MINI-SENTINEL METHODS

EVALUATING STRATEGIES FOR DATA SHARING AND ANALYSES IN DISTRIBUTED DATA SETTINGS

Prepared by: Jeremy A. Rassen, ScD (1), John Moran (1), Darren Toh, ScD (2), Mary K. Kowal (1), Karin Johnson, PhD (3), Azadeh Shoabi, MS, MHS (4), Tarek A. Hammad, MD, PhD, MSc, MS (5), Marsha A. Raebel, PharmD (6), John H. Holmes, PhD, (7), Kevin Haynes, PharmD, MSCE (7), Jessica Myers, PhD (1), Sebastian Schneeweiss, MD, ScD (1) and the Members of the Mini-Sentinel Strategies for Data Sharing and Analysis Workgroup

Author Affiliations:

(1) Division of Pharmacoepidemiology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School
(2) Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute
(3) Group Health Research Initiative
(4) Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration
(5) Division of Epidemiology, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration
(6) Kaiser Permanente Colorado Institute for Health Research
(7) Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania

November 30, 2012

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

ACKNOWLEDGEMENTS

The authors would like to acknowledge the participation of all the members of the Workgroup for their valuable contributions to the knowledge and discussion that formed the basis of this Report. We would also like to acknowledge the leadership and contributions of Richard Platt, MD, MSc, Principal Investigator of the Mini-Sentinel Project, and to acknowledge the contributions of Clara Kim, PhD, of the FDA.

Mini-Sentinel Methods

Evaluating Strategies For Data Sharing And Analyses In Distributed Data Settings

Table of Contents

I. INTRODUCTION

A. BACKGROUND

The Mini-Sentinel program is charged with conducting investigations that will facilitate the development and implementation of the Sentinel System while meeting the mandates of the 2007 FDA Amendments Act (FDAAA) legislation. Mini-Sentinel will identify the best scientific strategies and methods for delivering a reliable and efficient Sentinel System.

An examination of studies evaluating the strengths and weaknesses of a distributed database model against centralized data models led Mini-Sentinel to adopt the distributed data setting.[1] In the distributed data setting, data remain under the direct control of Data Partners. It is thought that maintaining data at Data Partners' sites will, among other benefits, encourage the ongoing collaboration of a wide range of major sources of health care data. In addition to minimizing the security concerns of Data Partners, the distributed data approach may better satisfy patient privacy guidelines mandated by the Health Insurance Portability and Accountability Act (HIPAA) and FDAAA, as well as the desire of health care organizations to maintain control of their proprietary information.

Having established the distributed database setting as the approach to be utilized in Mini-Sentinel, this project undertook the next necessary investigative step: *Evaluating Strategies for Data Sharing and Analysis in Distributed Data Settings.*

There are a number of issues involved when deciding the best methods to conduct medical product safety evaluations in Mini-Sentinel.

The overall objective of this project is to provide Mini-Sentinel with recommended analytic and data handling strategies that will most effectively balance the intersecting challenges and potential resolutions facing Mini-Sentinel public health activities, specifically in areas of: Epidemiological Fundamentals, Shared Data, Analytic Integrity and Flexibility, Operational Feasibility and Efficiency as well as the Protection of Personal Health Information (PHI) as mandated by the HIPAA Privacy Rule.

B. SUMMARY OF GOALS OF TASK ORDER

The following discussion summarizes the tasks required for the *Evaluation of Strategies for Data Sharing and Analysis in Distributed Data Setting* workgroup under a Mini-Sentinel Task Order.

The first step was to identify and describe the major strategies currently in use (or proposed) for both the analysis and handling of data. "Data handling" here refers to the methods employed in gathering and transmitting data within a distributed setting. The methods of data handling can vary between Data Partners. Specific analytic and data handling strategies are outlined in detail. In addition, data sharing and analysis experiences were drawn from conversations with researchers at the Vaccine Safety Datalink

[1] http://www.brookings.edu/~/media/Files/events/2010/0111_sentinel_workshop/03_Brown.pdf

(VSD), the Observational Medical Outcomes Partnership (OMOP) and the HMO Research Network (HMORN).

Once the major strategies for analysis and data handling were enumerated and described, key issues for each strategy were identified.

Each separate analysis and data handling strategy was evaluated in detail. In assessing the strengths and limitations of each approach, more than fifty criteria were reviewed, discussed, and rated for each of the methods being evaluated.

C. STATEMENT OF TASKS ACCOMPLISHED

The Working Group undertook the following tasks:

1. Enumerated and described the major available analytic strategies that have been used or proposed for use in similar distributed data systems (See Section IV.A)
2. Enumerated and described the major available data handling strategies that have been used or proposed for use in similar settings (See Section IV.A)
3. Enumerated the key issues that are important to consider when evaluating different strategies for data handling and analysis (See Section VI)
4. Evaluated each analytic and data handling strategy, and assessed the strengths and limitations of each approach with respect to the identified issues (See Section VI)
5. Created a report that included the deliberation process and recommendations of potential analytic and data handling strategies that most appropriately balance the relevant data and analytic issues in the Mini-Sentinel environment.

D. GUIDING PRINCIPLES

The primary goal of this project is to recommend potential analytic and data handling strategies for various types of activities envisioned in Mini-Sentinel that most effectively balance the intersecting issues of shared data, analytic integrity and flexibility, operational feasibility and the protection of personal health information (PHI) as mandated in the HIPAA Privacy Rule.

The evaluation was driven by the following guiding principles:

- The primary purpose of the recommended approach is the evaluation of hypothesized medical product safety issues across a distributed data network. The approach may not be optimal for data mining or other non-hypothesis-driven queries.
- The recommended approach should broadly protect the privacy of patients and the proprietary information of the participating institutions, as Mini-Sentinel prefers to accept a moderate amount of operational complexity in exchange for reducing the amount of identifiable information the partners are required to share.
- At the same time, the recommended approach should also meet the epidemiological needs – including adequate confounding control – required by a medical product safety surveillance system to produce valid results.

- To the extent that these goals may conflict, the recommended approach should strike a favorable balance and attempt to address any conflict pragmatically.
- Operational efficiency, including staffing levels, necessary expertise, and other concerns that affect the day-to-day operation of a distributed medical product safety surveillance network at the data partner sites, MS Operations Center, and other involved entities should form a large part of the evaluation and recommendation.

E. SUMMARY OF WORKGROUP ACTIVITIES

The Workgroup leaders for this investigative report are: **Jeremy A. Rassen, ScD** (Assistant Professor of Medicine at Harvard Medical School and Director of Computational Pharmacoepidemiology in the Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women's Hospital) and, **John H. Holmes, PhD, FACMI** (Associate Professor of Medical Informatics in Epidemiology in Biostatistics and Epidemiology at the Hospital of the University of Pennsylvania).

In consultation with the FDA, MSOC, the Methods Core, and other experts in the field, the Workgroup leaders were responsible for constituting the full membership of the Workgroup, guiding the group's activities and delivering a detailed report of the findings at the conclusion of the investigations and evaluations. The Workgroup members are listed in the Appendix (See Section IX).

The general responsibilities of the Workgroup members were to:

- Participate in bi-weekly phone conferences
- Contribute methods for investigation, as well as criteria by which to judge methods
- Contribute critical evaluations of methods under investigation
- Contribute report sections or other work products
- Review and discuss reports and other work products

1. Literature Review

An essential step for the group was to conduct a semi-formal literature review to accurately identify and describe the major methods, in use or proposed, for analysis and handling of health care data. The Workgroup sought to enhance its understanding of existing strategies by drawing upon the efforts of others engaged in the field through this review of published articles.

a. Literature Review Sources

- Workgroup members put forth known references in the scientific literature for consideration and inclusion.
- A list of references drawn from known papers was assembled and filtered.
- Limited, structured searches for pertinent references in PubMed were conducted

b. Literature Review Actions

The Workgroup distilled the methods found in the literature and enumerated the major strategies around which the investigation would revolve. The objective was to attain a greater understanding of the technical details of available methodologies, including identifying any limitations.

c. Literature Review Deliverables

Deliverable by-products of the Workgroup efforts in compiling and distilling the literature review include a detailed reference list, a PDF library (to the extent permitted by copyright law), and a list of the methods to be considered in this report.

2. Conversations with leading researchers at prominent distributed data networks

Direct discussion with researchers at several actively engaged distributed data research networks augmented the findings of the literature review. Conversations were held with staff members at VSD, HMORN, and OMOP. Their experiences in conducting analyses in distributed data settings served to underscore the strengths and limitations of particular analytic and data handling strategies. The methods used by each network and the lessons learned provided high value, field-level input providing additional depth and dimension to this report.

3. Extensive Evaluation of Five Key Methodologies

After completing the literature review, enumerating the key methodologies to be assessed, and identifying the key criteria against which each method would be appraised, the Workgroup began evaluation for each of the major strategies:

1. Covariate Sharing
2. Aggregated Data Sharing (with and without small cells removed)
3. Distributed Regression
4. Confounder Summary Score-Based Pooling
5. Meta-Analysis

With more than 50 criteria established for assessment against each method, the Workgroup made its analysis during and in-between the bi-weekly conference calls. Each method was assessed with key performance characteristics under the major headings of Epidemiological Fundamentals, Analytic Flexibility, Privacy, and Operational Efficiency. The Workgroup members discussed all enumerated issues of interest to determine ratings along each criterion. A detailed breakdown of the scoring by method, by criteria is found in Section VI.

The final step in Workgroup activity is the creation of the investigative report together with the incumbent methodological recommendation(s) for Mini-Sentinel.

II. SUMMARY OF LITERATURE REVIEW

Recognizing that developing and testing strategies for data sharing and analyses, in both distributed and centralized settings, has generated a body of relevant scientific literature, the Workgroup reviewed the pertinent published articles. The goal of this review process was to identify and describe the major strategies, in practice and proposed, for handling and analyzing shared data that would form the basis of the investigative report. The Workgroup identified the strengths and limitations of each major method as reviewed in the literature, and noted where modifications and improvements within those methods might be considered.

A. DESCRIPTION OF LITERATURE REVIEW METHODOLOGY

We gathered articles for review from three main sources. First, relying upon the expertise of the Workgroup, we solicited from each member any pertinent references already known to them in the existing body of literature. Individual members recommended specific articles for group consideration and inclusion in the review, reference list or PDF library. Second, we reviewed references for citations relevant to our evaluations. Third, we performed limited searches for appropriate article references in PubMed. The search terms included the following MeSH headings:

- Databases, Distributed
- Database, Distributed
- Distributed Database
- Distributed Databases
- Multicenter Studies
- Multicenter Studies as Topic
- Meta-Analysis
- Meta-Analysis as Topic

Given the broad nature of these MeSH terms and the great number of papers identified, we screened papers carefully and included only those thought to be most relevant in our literature review.

B. LITERATURE REVIEW FINDINGS

A full list of all papers evaluated appears in Section VII.

The literature review identified the major methods for data sharing and analyses in distributed data settings evaluated in this report. These methods became the focus of the Workgroup’s evaluation and are described in detail in Section IV.D. The findings revealed a clear grouping of major methodologies. Note that the strategies identified and evaluated are not mutually exclusive. That the approaches reviewed reflect areas of overlap as can be seen in the evaluation and scoring chart (see Section VI).

In some instances, methods that excel in their focus on sharing data may fail to provide adequate and mandated protection covering patient-level information. In other instances, a strategy may place its greatest emphasis on transmitting summary information. Other methods center on an epidemiological issue such as adjusting for confounders or may instead reflect a strong statistical emphasis.

III. SUMMARY OF DISCUSSIONS WITH DISTRIBUTED DATA NETWORK INVESTIGATORS

A. OVERVIEW

As part of this investigation, we conducted interviews with key staff members of the VSD, OMOP, and HMORN.

A key goal of engaging in discussions with scientists at VSD, HMORN and OMOP was to augment our literature review and Workgroup knowledge with field-level input relevant to our evaluations. All three data networks have grappled with many of the same methodological issues discussed in this report. Significantly, each organization found that the distributed data setting was far more effective, efficient, and secure than the centralized data pooling approaches of the past.

The discussions, detailed below, revealed common challenges like non-uniform data structure across sites, significant quality control variations among Data Partners, and inconsistent programming and data management skills. These findings, as well as other insights gained from these discussions, were used as inputs into the formal method evaluation process (see Section VI.A).

The dialogues revealed an overall need to develop models with sufficient flexibility to attract the participation of major Data Partners while still maintaining strong data definition consistency across sites to facilitate authoritative and expeditious analyses.

Most of the organizations we spoke with conveyed that if there had been more time spent developing the data model with sites in their network at the outset, issues of data quality, uniformity of file structure, and inconsistent variable definition could have been reduced.

The interviews also underscored that the distributed data setting encouraged greater participation from Data Partners who needed to know that their patient data and proprietary data management techniques would remain maximally secure within their own organizations.

A frequent theme in these discussions indicated that the adoption of more uniform data management standards across participating sites is a highly desirable goal directly affecting the issue of data quality. It is a challenge that remains in the forefront of current efforts in the distributed data setting.

In sum, the consultations contributed the element of practical experience. We would like to thank the scientists from each organization for sharing their historical observations, current challenges, and proposed enhancements in pursuit of attaining effective, secure, and efficient strategies for data sharing and analysis.

B. DISCUSSION WITH VACCINE SAFETY DATALINK INVESTIGATORS

1. Description of VSD approaches

In 1990 the CDC and FDA established the VSD and the Vaccine Adverse Event Reporting System (VAERS) to monitor vaccine safety. Addressing the gaps in scientific knowledge about the connections between adverse events and immunizations has always been at the core of the VSD’s pursuit for enhanced immunization safety.

The VSD project includes a large linked database that uses administrative data sources at each of ten health plans. Each participating site gathers data on vaccination (vaccine type, date of vaccination, concurrent vaccinations), medical outcomes (outpatient visits, inpatient visits, urgent care visits), birth data, and census data.

The VSD is under the direct administration of the CDC's National Center for Health Statistics (NCHS) and thus ultimately under the supervision of the U.S. Department of Health and Human Service.

VSD's goals in brief:[2]

- To conduct population-based research on immunization safety questions
- To evaluate immunization safety hypotheses that arise from medical literature, passive surveillance systems, adjustments to immunization schedules, and introduction of new vaccines
- To guide national immunization policy decisions
- To partner with healthcare providers, public health officials, and others to ensure the public has the best available information regarding the risks and benefits of immunization

The following health plans are currently part of the VSD, all of them are also members of the HMORN (described in the next section):

1. Group Health Cooperative, Seattle, Washington
2. Harvard Pilgrim Health Care, Boston, Massachusetts
3. HealthPartners Research Foundation, Minneapolis, Minnesota
4. Kaiser Permanente Northwest, Portland, Oregon
5. Kaiser Permanente Medical Care Program of Northern California, Oakland, California
6. Kaiser Permanente Colorado, Denver, Colorado
7. Kaiser Permanente of Georgia, Atlanta, GA
8. Kaiser Permanente of Hawaii, Honolulu, Hawaii
9. Marshfield Clinic Research Foundation, Marshfield, Wisconsin
10. Southern California Kaiser Permanente Health Care Program, Los Angeles, California

Currently, data for more than 18 million persons spanning 16 years are available for VSD research.

From 1990 until 2000 data was sent from participating VSD sites (six to begin with) and centrally pooled at the CDC for analysis. The 'central pooling' approach proved sub-optimal and in 2000 a confluence of forces led the VSD to adopt the distributed data model to facilitate these benefits:

- Data providers retaining permanent, physical control of their data

[2] http://www.cdc.gov/vaccinesafety/Activities/VSD.html

- Local content experts maintaining a close relationship with the data and its protection
- Eliminating the need to create, secure, maintain, and manage access to a complex, central data warehouse.

During the period of 2000-2004 participating sites created large data files on an annual basis. In order to respond more effectively to signals from VAERS (reports of rare, possible adverse events), VSD required data updates more frequently than those annual intervals. (For example, an annual data file would be due at the end of September and from it a file created with all patient data ending with the previous December. The ultimate result generated analyses that were over a year and a half old.)

In 2005 to address this analytic lag, the VSD launched its Rapid Cycle Analysis program (weekly dynamic data files). It monitors real-time data to compare rates of adverse events in recently vaccinated people with rates among unvaccinated people.

VSD Rapid Cycle Analysis process:

1. When a new vaccine is introduced, VSD begins an active surveillance program using the weekly dynamic data files.
2. Code is written centrally to manipulate individual data files and then distributed.
3. Participating sites then prepare and send aggregated data to the central facility. The data is sent in aggregated counts form, with cells indicating groupings of patients and counts indicating how many patients fall into that grouping (see Section IV.D.2).
4. The central facility performs a quality evaluation and the analysis.

The checking of data quality is crucial to the integrity of on-going analysis since the quality varies significantly across VSD sites. For example, a signal may be generated at one site because a coder has either entered incorrect code (or used code incorrectly) thus creating a non-existent outcome. With adverse outcomes so rare among vaccines, it takes very little input error to contaminate results. In addition there are many variations in the way files are structured for analysis and very few people in the country are engaged in carrying out that critical work (about 12 programmers in all).

While the weekly dynamic data files contain most of the annual constants, such as birthdate, sex, and vaccine exposures, they do not consistently include lab, pharmacy, mortality, or birth certificate data. If needed, this information must be obtained from other electronic sources and can add a great deal more programmer time to the already extensive amounts required for maintaining the dynamic weekly data files and creating the annual files.

The system’s rapid response capabilities are also challenged by time consuming issues involving protocol development, Institutional Review Board (IRB) approvals, and issues related to inconsistencies across participating sites.

(It should be noted that while VSD is an HMORN affiliated research network, it operates separately. Although VSD data is coming from the same sources and patients, a different data structure is being used; this is a common issue among the large research consortiums.)

2. Lessons learned from discussion with VSD investigators

In discussion with VSD staff, these key themes emerged as crucial to advancing the effectiveness of vaccine safety monitoring and adverse event rapid response.

1. Reducing variations in data quality across sites will reduce the occurrence of false signals.
2. Extensive programmer time can be reduced if the weekly dynamic data files consistently contained lab, pharmacy, mortality, and birth certificate data in addition to the annual constants.
3. Greater uniformity in the way files are structured for analysis across sites could further expedite investigation and response.
4. Less variation in the way each site chooses to carry out its analysis would advance efficiency and quality.

C. DISCUSSION WITH HMO RESEARCH NETWORK INVESTIGATORS

1. Description of HMORN approaches

The HMORN, formed in 1994, is a consortium of 19 plans or delivery systems with large, defined patient populations and formal, recognized research capabilities.[3] HMORN functions as a network of research centers that work within or in close partnership with integrated health systems across the United States and in Israel.

The research network's critical hierarchy consists of a governing board, an asset stewardship committee, a Virtual Data Warehouse (VDW) Operations Committee (VOC), an administrator's forum, a multi-center IRB coordination group, and a knowledge management group. HMORN sites carry out public domain health research on a wide range of diseases and cross-cutting topics such as cancer, vaccine safety, heart disease, pharmacoepidemiology, and more.[4]

Structurally, under the supervision of the VOC, there exists one working group for each data area (e.g., pharmacy, enrollment, and census). Each working group is volunteer-based and led by a principal investigator and analyst(s). The working group is responsible for data in their area, including data update and quality checks which must be conducted annually, at a minimum.

Current members of HMORN include:

1. Group Health Cooperative, Seattle, WA
2. Kaiser Permanente Northwest, Portland, OR
3. Kaiser Permanente Northern California, Oakland, CA
4. Palo Alto Medical Foundation, Palo Alto, CA

[3] http://www.geisinger.org/research/gchr/Collaborations/HMORN.html

[4] http://www.hmoresearchnetwork.org/about.htm

5. Kaiser Permanente Southern California, Pasadena, CA
6. Kaiser Permanente Hawaii, Honolulu, HI
7. Lovelace Health System, Albuquerque, NM
8. Kaiser Permanente Colorado, Denver Colorado
9. Scott and White Health Plan, Temple, TX
10. Health Partners, Minneapolis, MN
11. Essentia Health, Duluth, MN
12. Marshfield Clinic, Marshfield, WI
13. Henry Ford Health System, Detroit, MI
14. Kaiser Permanente Georgia, Atlanta, GA
15. Kaiser Permanente Mid-Atlantic, Rockville, MD
16. Geisinger Health System, Danville, PA
17. Harvard Pilgrim Healthcare, Boston, MA
18. Fallon Community Health Plan, Worcester, MA
19. Maccabi Healthcare Services, Tel Aviv, Israel

HMORN’s goals in brief:

- To be recognized as the Nation’s premier resource for population-based health and health care research
- To contribute to national and global dialogues on health research and policy by serving as a credible source of evidence-based information
- To promote and establish the Network as a preferred research partner of funding agencies
- To foster Network-led collaborative studies
- To share methodologies, best practices, and consultative expertise

HMORN member health systems provide comprehensive healthcare for approximately 15 million people.

They offer extensive clinical data resources to qualifying research projects and have the ability to coordinate these data resources in support of large, varied programs of multi-site, multi-purpose studies.[5] At the core of this ability is the VDW—a virtual data warehouse created by mapping data from local systems into a common format.

[5] http://www.hmoresearchnetwork.org/resources/toolkit/HMORN_Research-Process-and-Partnership-Primer.pdf

HMORN's VDW evolved in 2003 through an effort to organize data to further facilitate use in multi-site cancer research while simultaneously maintaining local control and security of patient data. In this and other respects it mirrors the distributed data settings under evaluation in this report.

HMORN describes VDW with these process characteristics:[6]

- A non-centralized database assures patient privacy is protected.
- Sites agree on the data to make available for research, and on definitions and formats to apply in order to standardize those data.
- Raw administrative, clinical and claims data are transformed to the agreed upon set of data standards at every HMORN site.
- Each institution's VDW data remain at their site until a study-specific need arises then the required study data are extracted.
- After IRB and HIPAA requirements are met, a program can be written at one member site to be run at other sites (with custom capabilities if required).

The HMORN process in brief is:

1. Write a work plan (statistical program), code description, and anticipated return of values
2. Have the lead site write the code and send it out to the local sites
3. Have the local sites run the code and review the results against the work plan
4. Send the results (in the form of an analytic work file) back if the data meets expectation, conforms to the IRB-approved specifications, and the format is as specified.
5. The central site analyzes the analytic file and prepares study results (see Section V.D.1).

The research protocol, including data sharing, is typically developed by the research team for each study. As all research data originate from health care or administrative data, each protocol must comply with relevant federal and state laws pertaining to research. In addition, each protocol must follow institutional policies. These policies may reflect institutional considerations such as intellectual property. As a result, HMORN studies have experienced cross-site variation in the data-sharing specifics that an IRB or other relevant institutional bodies will approve, for example, what is considered "minimum necessary" and how small cell counts must be handled. The HMORN has pursued several initiatives to streamline IRB review to the extent possible.

The more granularity, or perceived risk of identification, involved in a study, the more the IRBs/institutions will mandate protective guidelines such as limiting the number of people allowed to see the data, and requiring secure communication, restricting where and for how long the file(s) can be stored. When no IRB review is required, such as for public health activities or preparatory to research in most states, HMORN still pursues a 'minimum necessary' approach in order to responsibly manage

[6] http://www.hmoresearchnetwork.org/resources/resources_home.htm

member health information. *Note:* most public health reporting (e.g., mandatory reporting of communicable diseases) does not involve HMORN but is organized through the individual affiliated provider groups.

2. Lessons learned from discussion with HMORN staff

In the discussion with HMORN staff concerning what aspects of their structure they would alter if they could, three primary areas emerged.

1. More time developing the data model with site-by-site input would have established better variable definitions across the network. The present lack of definition consistency between sites remains problematic.
2. A greater emphasis on fundamental organizational structure would have generated more specific and consistent expectations from the participating sites, including a more uniformly coordinated data model and enhanced quality checking protocol.
3. The management of the network as a whole remains unfunded. Funding is undertaken on a study-by-study basis.

D. DISCUSSION WITH OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP INVESTIGATORS

1. Description of OMOP approaches

OMOP was formed in 2007 as an initiative funded by the Foundation for the National Institutes of Health (FNIH)[7] and with contributions from 16 for-profit and one non-profit (Pharmaceutical Research Manufacturers of America) donors. The partnership functions in collaboration with academic institutions, the pharmaceutical industry, the FDA and other federal agencies.

OMOP is self-described as "…a public-private partnership designed to help improve the monitoring of drugs for safety…"[8] The primary goal is to evaluate efficient and beneficial analytic methods applied to large healthcare databases to identify and assess the safety of drugs.

OMOP's operational structure consists of an Executive Director, a Senior Program Manager, nine Research Investigators, and a Statistics and Programming Team.

Some current and past research collaborators include:

- GE Healthcare
- Thomas Reuters
- Department of Veterans Affairs PBM Center for Medication Safety
- Indiana University and the Regenstrief Institute

[7] http://www.fnih.org/sites/all/files/documents/15%20Timeline.pdf
[8] http://omop.fnih.org/node/22

- Massachusetts General Hospital
- Hewlett Packard, HP Labs- Auburn University, Harrison School of Pharmacy
- Humana
- i3 Drug Safety (now OptumInsight)
- Partners HealthCare System
- Computer Sciences Corporation (CSC)
- Eli Lilly and Company
- Merck Research Laboratories
- Harvard Pilgrim Health Care Institute
- SDI Health
- University of Utah

At the outset, OMOP focused on evaluating and testing different data access alternatives. They set up and tested both centralized and distributed data approaches.

- A centralized environment was set up for utilization by those data holders open to sharing the actual data with other partners. This was originally done within a secure hosting environment engineered by Computer Sciences Corporation (CSC). It is now secured in an Amazon cloud computing environment that retains multiple disparate data sources in their own SAS data sets and schemas. In this approach OMOP acts as a central hub and creates a secure virtual instance for partners to upload data. The partners then take control and run analyses according to methods set by OMOP. In such instances OMOP is quite specific as to infrastructure: UNIX machines, Oracle and SAS datasets (sometimes translated from text files).
- In the distributed environment the partners established their own system infrastructures (Oracle, SAS or SQL Server). Each site maintained control whether to run analyses through an Amazon cloud or their own on-site server (as in the case of the EU's Adverse Drug Reaction project).

OMOP worked in collaboration with those participating partners who built their own servers. Each partner translated their data to the common data model and then ran analyses according to OMOP methods. Specifications for the data model were provided but not for their individual computing environments.

In the distributed setting it is possible to stand up virtual machines, run the analysis, and share the outcomes with the hub. If for example it was desired to carry out high dimensional propensity scoring (HD-PS) analysis on five separate databases, they could stand up five VM's, run the program, and report the results to a common server.

It is important to note that OMOP's use of the term 'centralized' is somewhat of a misnomer as both of the above settings are in fact instances of distributed analyses.

Specific Notes on the OMOP Environment

OMOP developed most of the programming at a central lab. However, in some circumstances code was written by members of the partnership after receiving the applicable standards and practices. In such a situation OMOP fulfilled the role of an independent tester. If the contributed code was proven to have no issues in testing and was in accord with the partnership's practices, OMOP would write the standardized documentation and release it publically.

Under both settings OMOP tested code with SAS and Oracle libraries. While SAS and Oracle presented no problems, processing SQL statements was problematic and those sites using an SQL Server had to perform code modifications before running analyses.

When using SAS, partners were told to apply ANSI SQL. While most adhered to that guideline, some programs still did not prove out in quality assurance checks. Non-SAS programs could not use SAS data sets and instead were engineered through Oracle or text files.

2. Lessons learned from discussion with OMOP staff

The fact that OMOP accepted a wide range of programming environments (SAS, Oracle, SQL) from partners had distinct benefits as well as some drawbacks.

1. The overriding benefit of OMOP's environment flexibility was that it made the partnership more attractive to collaborators. If participation had depended on all members first becoming proficient with SAS for example, there would have been far fewer data partners. Flexibility allowed the broadest access to potential data sources and enabled more collaborative analyses. (On the other hand, if Sentinel wanted to bring on a partner without SAS, it could be problematic.)
2. A drawback to the flexible environment approach was that OMOP had to provide substantial support for many distinctly different settings. For example, even within a specific environment there could be heterogeneity requiring OMOP assistance.

IV. EVALUATION OF ANALYSIS TECHNIQUES

A. OVERVIEW OF METHODS EVALUATED

The selection of major methods to be evaluated relied upon a literature review, discussions with workgroup members, and conversations with researchers at collaborative networks such as VSD, HMORN, OMOP, and MSOC. Five primary methods were identified for inclusion:

1. Analysis with full covariate sharing
2. Analysis with aggregated data sharing, with and without small cells removed/masked
3. Analysis with distributed regression methods
4. Analysis with confounder summary score-based methods, including propensity- and disease risk score-based methods
5. Meta-Analysis

B. TERMINOLOGY

In describing the analytic methods below, these frequently used terms are defined as follows:

- A **site,** sometimes called a "center" in the literature, refers to any participating Mini-Sentinel entity that will contribute data to the analysis. Sites are responsible for collecting, cleaning, organizing, and transmitting data.
- The **hub** (sometimes called a "master center" in the literature) refers to the MSOC or other location in which the data will be pooled for centralized analysis.

C. ACTIVITIES COMMON TO ALL METHODS

Before discussing each method in greater detail, it should be noted that the following study characteristics were identified as common throughout all the methods evaluated in this report.

1. A common data model is developed and defined for utilization across all the participating sites.
2. A common protocol is generated that defines the patient population to be studied.
3. The common protocol also defines and implements the definitions of all standard and study-relevant covariates. (It should be noted that this may entail more covariates than will be used in the study, if certain covariates are ultimately determined to be not meaningful to the analysis.)
4. After application of the patient selection criteria noted in the protocol, a standard "Table 1" is created.
5. After application of the analyses noted in the protocol, the transmission of diagnostic information is as follows.
6. For matched analysis, a "Table 1" (summary of the patient population, as characterized by all measured covariates and stratified by treatment status) is created for the unmatched and matched cohorts, with measures of absolute and standardized differences between treatment groups for each confounder.
7. For stratified analysis, a Table 1 is created for each stratum (e.g., PS decile, quintile) with measures of absolute and standardized differences between the treatment groups for each confounder.
8. For score-based approaches (propensity scoring or disease risk scoring) a plot is created of the distribution of the score in each exposure group before and after applying restrictions (such as matching, trimming, etc.).
9. SAS log files are generated and shared to detect gross errors in coding or application of the protocol.

D. DESCRIPTION OF METHODS EVALUATED

An overarching consideration with each of the combined database scenarios reviewed is to assess how well each method can meet the goals of analytic integrity and flexibility, as well as operational feasibility, while simultaneously maintaining privacy protection for patients and health plans.

1. Method One: Analysis with Full Covariate Sharing

Analysis with full covariate sharing is the most analytically straightforward of the methods we evaluated, as it entails the full sharing of individual patients' covariate information. In this approach each participating site creates an analytic dataset based upon the study protocol developed and provided by the hub. These datasets are then transmitted to the hub where analytic models of interest are run on the submitted data.

In a cohort-based analysis, the composition of the datasets involves each patient having one or more rows of data, the primary elements of which will include all available information pertaining to:

- Exposure status
- Occurrence of outcome events
- Date(s) of service when exposure was determined, at a pre-determined level of granularity (day, month, quarter, etc.)
- Details on medical encounters prior to and downstream of exposure, with dates measured as days from the index date

In this approach patient data will include all additional covariate information available. Though all of the covariates should be thought to be "minimally necessary" at the time of the protocol dissemination, to maintain the hub's analytic flexibility, it should be noted that they may ultimately not all be used for the analysis. The possibility exists that the transmission of full covariate information may include PHI. Reasonable techniques to guard against exposing PHI, such as age categorization, would be employed to lessen the risk, though transmission of some PHI may be expected nonetheless.

Patient identifiers may be included in this approach, but they need not be directly identifiable (such as with SSNs). Instead, the use of unique, random numbers could be effectively employed as identifiers to connect individuals across data sources.

While this method has significant advantages in terms of speed, analytic flexibility, and its minimal requirements for statistical expertise at each site, its main disadvantage is that it presents the possibility of privacy concerns.

2. Method Two: Aggregated Data Approach (without small cells removed)

This method employs the technique of aggregating or collapsing like patients' information into cells, after which adjustments are made for confounders based upon the counts of patients in each cell.

These aggregate cells will contain exposure, outcome status, follow-up time, and covariate information. One cell, for example, may represent patients who share the following combination of criteria.

CRITERION	YES/NO	COUNT
Exposed?	Yes	317
Outcome event?	No	
Followed for 180 days?	Yes	
Aged 75-85?	Yes	
Female?	No	
History of cardiovascular disease?	No	

The cell would indicate exactly how many patients met these criteria, e.g., 317. Subsequent cells would also be created for each possible combination of exposure, outcome, and covariates.

The total number of cells transmitted is determined by the number of observed combinations of exposure, outcome, follow-up time, and covariates. Zero cells would not have to be transmitted and no additional covariate information beyond the minimum required for the study might be sent to the hub.

While no specific patient-level information is sent, "small" cells (with varying patient thresholds of <6 and <11, for example) would be transmitted. The hub would then run the analysis on the aggregated data.

In this scenario, with numerous covariates and rare outcomes, it would be anticipated that cells with only one or few patients would be common. Under such circumstances this approach would provide approximately the same patient privacy protection as method one above.

The main advantage of this method is that little patient-level information is shared, especially when small cells have been removed or collapsed. Little statistical expertise is needed at each site. Its primary disadvantage includes the fact that certain statistical models (e.g., Cox proportional hazards) may be harder to achieve because detailed patient-level information is not available.

3. Method Two (Variant): Aggregated Data Approach (with small cells removed or masked)

In response to the privacy issues present when small cells are included, the Centers for Medicare and Medicaid Services (CMS), mandates[9] that cells contain a minimum of 11 patients. While improving the level of patient privacy, this approach diminishes the effectiveness of the analytic study. For instance, if cells with fewer than 11 patients were simply dropped or excluded from the analysis to meet the CMS

[9] http://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/downloads/cms-r-0235l.pdf

regulation, the statistical power and flexibility required to reveal rare safety outcomes might not be available.

Combining a series of cells until they met the minimum privacy protection threshold of >10 patients is another possibility that has been evaluated. This, however, would result in the merging of dissimilar patients into a single stratum thereby generating a substantial loss of detailed confounder information and significantly diminishing the validity of the analysis.

The small cells (those containing outcome events, for example) may be the most relevant to the analysis, so removing or collapsing those cells and losing covariate information may be quite detrimental.

4. Method Three: Distributed Regression Approach

Distributed regression encompasses a suite of methods that can be used to estimate regression models from distributed patient data. Unlike the other approaches described thus far, these estimates would be obtained without participating data sites providing individual patient data or stratum-specific data to the central analytic hub. Each site provides their summary statistics, which can then be combined at the hub to produce regression estimates for analysis.

Methods have been developed for estimating regression parameters from distributed data for all of the exponential family generalized linear models, including data with normal, Poisson, and binomial distributed outcomes. Such analytic flexibility while protecting patient-level information is strength of the distributed regression approach.

The primary advantage of this approach is to provide the analytic hub with estimates of regression parameters that are equivalent to estimates that could have been obtained if the hub had full patient-level data access, but without the need for explicit data sharing. Among the disadvantages of this method, there is a large potential statistical burden on each participating site, there may be little analytic flexibility, and the choice of models will be dictated by available software.

5. Method Four: Score-Based Approaches

The utilization of propensity scores (PS) and disease risk scores (DRS) in conducting pooled data analysis may offer a balance between the full utilization of covariate information and the maintenance of patient privacy that other methods cannot provide. Since a PS or DRS is a means of summarizing many covariates into a single opaque number, the method allows for the protection of individual patient-level information while retaining the content needed to adjust for confounding. Use of score-based approaches also provides enhanced statistical and analytic flexibility that is often limited in other approaches we have described.

In the score-based approach, after establishing the design components of a desired study (sites and data sources, cohort criteria, exposure and outcome definition, covariates available by site, patient subgroups etc.) the participating sites segment patient-level information into three categories.

- **Shared covariates:** Non-confidential or limited dataset covariates including age in decades, sex, date of exposure (in day-, month-, or quarter-level granularity as appropriate) , drug exposure status, event and censoring dates
- **Private covariates, universally available:** Covariates available at all sites, where the covariates may be considered information that cannot be shared or that an institution does

not wish to share. Examples include medical history variables created from diagnosis codes, prescription drug use, and recorded procedures.

- **Private covariates, available on a center-specific basis:** Covariates available at certain sites, where the covariates may be considered information that cannot be shared or that an institution does not wish to share. Examples include blood pressure, weight, lab values, and other fields that certain sites (such as those with electronic health record [EHR] data) may have available.

Example dataset for a single site (in this case, Site 6); other sites' data would look similar:

Site ID	Patient ID	Exposed?	Outcome Event?	Sex	Age 60-75?	Univ. PS
6	1	No	No	F	Yes	0.52
6	2	Yes	No	F	No	0.68
6	3	Yes	Yes	F	No	0.74
6	4	No	No	M	Yes	0.23
…	…	…	…	…	…	…

With these covariates defined, each center may estimate and record several scores. The first score would be based on the shared and private universal information (ScoreUniv). Optionally, sites could estimate a high dimensional propensity or disease risk score .

If the centers vary with respect to the amount of data available – for example, if one center has EHRs but the others do not – a second score might be estimated. This score, noted ScoreLocal, is based on the private center-specific information as well as the shared and private universal information.

Once the participating sites have implemented their basic study design, separated their measured covariates into the three patient information categories and estimated their scores, they can create center-aggregated files (see the example above) for transmission to the analytic hub. These site generated files would typically contain the study patient identifier, center identifier, exposure, outcome, shared covariates, and estimated scores. If desired, the scores can be used for the basis of creating balanced cohorts through matching.

Score-based approaches offer the advantage of transmitting epidemiological information in an opaque, privacy-oriented set of numbers, but may add complexity to analyses that require clear access to the underlying data. Errors in coding of variables and other "real-world" issues may be harder to detect in score-based methods than in approaches that share covariate data in a transparent form.

6. Method Five: Meta-Analysis

Meta-analysis is arguably the most straightforward approach in terms of structure of all the methods evaluated in this report. However, while the fundamental premise of meta-analysis is that each site conducts its own separate study, results may be ambiguous or undetectably flawed.

The hub defines a protocol which is then distributed to the individual collaborating sites. Each participating site would then either create the study in their data per the protocol, or execute a centrally-supplied macro file to do the study execution. Only the results of the statistical analysis are shared: the sites transmit their findings (point estimate, variance, and common diagnostic data) to the hub. With this method no PHI is forwarded to the hub as part of a site's transmitted data.

Upon receipt of each site's statistical results, the hub applies the received point estimates and variances to compute a singular, study-wide result for point estimate and confidence level. The hub also has the option to perform standard heterogeneity tests to ascertain whether all or only some of the sites qualify to be included in the summary estimates.

From a statistical standpoint, this approach should generate the mathematical equivalent of Method One.

The advantage of this method is the protection of patient privacy and the minimal data sharing that the sites must do. However, meta-analysis also has certain inherent disadvantages, including difficulties in detecting errors, the inability to perform post-hoc analyses without requiring participating sites to repeat analyses, and the lack of flexibility for subgroup analyses and the like.

Lastly, if each site is to execute its own study without centrally-supplied programming code – an approach that we ultimately do not recommend – risks such as inconsistent statistical capability and varying quality control mechanisms across sites (as noted in our discussions with investigators from other distributed studies) are incurred.

V. SUMMARY OF HIPAA PRIVACY RULE COMPLIANCE

The Workgroup engaged legal expertise to evaluate the five key methodologies in this report to determine whether they would comply with the HIPAA Privacy Rule. Kirsten B. Rosati of Coppersmith Schermer & Brockelman PLC drafted a memorandum entitled "Privacy Evaluation of Proposed Mini-Sentinel Statistical Methods", which appears in full as an appendix to this report.

The memorandum describes a setting in which the surveillance activity is carried out under a public health authority, and another setting in which the activity is carried out under a more general research setting. *As explained below and in greater detail in the full memorandum, each of the proposed statistical methods complies with the HIPAA Privacy Rule when applied in the context of a Public Health Authority. In addition, the methods each also comply with the HIPAA Privacy Rule when used in a research setting, given that certain criteria are met. Those criteria are described in detail in the memorandum.*

Section I of the memorandum discusses the HIPAA rules applicable to Mini-Sentinel activities. Section II applies that discussion to each statistical method. Section III summarizes compliance.

A. GENERAL CONCLUSIONS OF THE MEMORANDUM

As stated in the memorandum:

> Each of the proposed statistical methods will comply with the HIPAA Privacy Rule. Under some of these methods, the information disclosed to the MSOC [Mini-Sentinel Operations Center] will be de-identified under the HIPAA Privacy Rule. Moreover, even if the information is not de-identified under some of the statistical methods, the use or disclosure to the MSOC or its subcontractors still meets the requirements of the HIPAA Privacy Rule because the Operations Center and its subcontractors are functioning as public health authorities on behalf of the FDA[10,11]; use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization.

Note that the full memorandum addresses the issue of HIPAA as applied to “research”, distinct from “public health activity”.

B. KEY TERMS WITHIN THE MEMORANDUM

- **Covered Entities**: These include health plans, health care clearinghouses, and health care providers that engage in electronic “standard transactions” with health plans, such as electronic billing.
- **Individually Identifiable Health Information**: Health information, including demographic information collected from an individual that identifies an individual or where there is a reasonable basis to believe the information can be used to identify the individual. (See de-identification in Section V below.)
- **Protected health information:** A subset of individually identifiable health information that excludes certain health information held by employers and educational institutions.
- **Public Health Authority:** An agency or authority of the United States, a State, a territory, a political subdivision of a State or territory, or an Indian tribe, or a person or entity acting under a grant of authority from or contract with such public agency, including the employees or agents of such public agency or its contractors or persons or entities to whom it has granted authority, that is responsible for public health matters as part of its official mandate.
- **De-Identification**: HIPAA permits two ways to de-identify individually identifiable health information. First, a covered entity may follow the “safe harbor” method of de-identification and remove or code all of the HIPAA identifiers in the information. (A list of identifiers can be found in the complete memorandum.) The second method of de-identification is to have a

[10] McGraw, D., Rosati, K. and Evans, B. (2012), A policy framework for public health uses of electronic health data. Pharmacoepidem. Drug Safe., 21: 18–22. doi: 10.1002/pds.2319

[11] http://mini-sentinel.org/work_products/About_Us/HIPAA_and_CommonRuleCompliance_in_the_Mini-SentinelPilot.pdf

qualified statistical expert determine that the risk is very small that the information could be used alone, or in combination with other available information, to identify the patient.

C. MINI-SENTINEL ACTIVITIES AND HIPAA RULES

1. HIPAA Applies to Covered Entities and Business Associates

HIPAA applies to covered entities (as defined above) and will also apply to their business associates. Thus, to the extent that any of the data sources are business associates of HIPAA covered entities, they also will have to comply with the HIPAA Privacy Rule requirements discussed in the memorandum.

2. HIPAA Applies to Individually Identifiable Health Information (PHI)

The HIPAA Privacy Rule protects individually identifiable health information as summarized above and further detailed in the complete memorandum.

3. HIPAA Applies to Internal Use, As Well as External Disclosure of PHI

The HIPAA Privacy Rule permits covered entities to disclose PHI for a variety of public health purposes.

4. Data “Curation” is a Health Care Operation Permitted under HIPAA

The internal use of PHI to curate data for the purpose of producing datasets to send to the M-S Operations Center is itself a health care operation under HIPAA that is permitted without individual authorization.

5. Verification of Identity and Authority to Request Protected Health Information

To disclose PHI, data sources must confirm the recipient’s identity and that the recipient has the legal authority to request the PHI. A covered entity is entitled to rely on written confirmation on FDA letterhead that the M-S Operations Center and its subcontractors are acting on behalf of the FDA, and that they have the legal authority to request PHI for the Mini-Sentinel project.

6. Compliance with the Minimum Necessary Standard

HIPAA covered entities must observe the minimum necessary standard in using or disclosing PHI for any purpose other than treatment. This means that a covered entity must make reasonable efforts to limit the information to the minimum amount of information that is necessary to accomplish the intended purpose of the use or disclosure.

7. The Accounting Requirement

The HIPAA Privacy Rule currently requires covered entities to provide an “accounting” of disclosures of PHI to individuals at their request, with various exceptions, including disclosures that are made for treatment, payment and health care operations.

D. THE FIVE KEY METHODOLOGIES AND HIPAA COMPLIANCE

In the full memorandum each of the five key methods are reviewed in detail against the HIPAA legal standards under the assumption of M-S operating under public health authority.

In brief, all five methods were found to be compliant with HIPAA when applied in a public health authority setting. All five methods were found to be compliant with HIPAA when applied in a more general research setting, provided that certain criteria were met.

This summary serves only to provide highlights of that evaluation. Evaluation of the methods in a research setting is covered in the memorandum.

1. Method One: Analysis with Full Covariate Sharing [Public Health Authority Setting]

This method complies with the HIPAA Privacy Rule. The Operations Center and its subcontractors are functioning as public health authorities on behalf of the FDA; use of PHI by or disclosure of PHI to the Operations Center or its subcontractors is permitted as a public health activity without individual authorization. However, the transmission of full covariate information likely will include some HIPAA identifiers. The method does not however violate the HIPAA Privacy Rule to disclose PHI as long as the disclosure is for work contracted through the FDA, a public health authority conducting public health activities.

2. Method Two: Aggregated Data Approach [Public Health Authority Setting]

This method complies with the HIPAA Privacy Rule. The Operations Center and its subcontractors are functioning as public health authorities on behalf of the FDA; use of PHI by or disclosure of PHI to the Operations Center or its subcontractors is permitted as a public health activity without individual authorization. In addition and where applicable, internal data curation is a health care operation permitted under HIPAA without individual authorization.

It is not necessary to fully de-identify the PHI to comply with the HIPAA Privacy Rule. Rather, disclosure of aggregate information under Method Two, even with small cells, complies with HIPAA as long as the disclosure to the Operations Center or its subcontractors is for work contracted through the FDA, [or] it is for public health activities.

Finally, the disclosure of aggregated information complies with the HIPAA minimum necessary standard.

3. Method Three: Distributed Regression Approach [Public Health Authority Setting]

This method complies with the HIPAA Privacy Rule. First, it appears that the information disclosed to the Operations Center will be de-identified under the HIPAA Privacy Rule. Moreover, even if the information is not fully de-identified, the disclosure still meets the requirements of the HIPAA Privacy Rule because the Operations Center and its subcontractors are functioning as public health authorities on behalf of the FDA. Use by or disclosure of PHI to the Operations Center or its subcontractors is permitted as a public health activity without individual authorization.

Even if it is possible to recreate an individual identifier, in rare cases, this method would comply with the HIPAA Privacy Rule (as noted with regard to Method One and Method Two) because disclosures of PHI for Mini-Sentinel purposes are permitted as public health activities. Method Three would also comply with the minimum necessary standard.

4. Method Four: Score-Based Approaches [Public Health Authority Setting]

This method complies with the HIPAA Privacy Rule. Most of the information disclosed to the Operations Center will be de-identified under the HIPAA Privacy Rule. Moreover, even if the information is not de-identified, the use or disclosure still meets the requirements of the HIPAA Privacy Rule because the Operations Center and its subcontractors are functioning as public health authorities on behalf of the FDA. Thus, use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization.

5. Method Five: Meta-Analysis [Public Health Authority Setting]

This method also complies with the HIPAA Privacy Rule. The information disclosed to the Operations Center will be de-identified under the HIPAA Privacy Rule. Moreover, the internal use of the PHI by the covered entities to create the de-identified information for the Operations Center complies with the HIPAA Privacy Rule for two reasons: (1) if the data sources are subcontractors to the Operations Center, they are functioning as public health authorities on behalf of the FDA, and use of PHI by the subcontractors is permitted as a public health activity without individual authorization; (2) the internal data curation is a health care operation permitted under HIPAA without individual authorization.

VI. METHODS EVALUATION TABLE

In the following section is a descriptive and quantitative table that the Workgroup used to evaluate the five distributed data methods.

The table was an integral part of the project’s design from inception. The purpose of the table is to bring together, in one place and in concise terms, an overview of the key characteristics deemed essential when considering each of the five analytic methods reviewed by the Workgroup. Furthermore, the table was purposely designed to:

1. Generate a series of Workgroup discussions that would lead to a refined identification of precisely what key characteristics should be enumerated for the methods evaluated.
2. Serve as a platform for individually ‘scoring’ each key issue in the context of each method and thereby further elicit detailed discussion among Workgroup members to reach consensus scores for each issue enumerated
3. Augment the Workgroup’s final descriptive recommendations by providing a means for rendering individual scores into cumulative total scores for each of the five methods

The key characteristics for consideration in the table are divided into four major categories:

1. Epidemiological Fundamentals
2. Analytic Flexibility
3. Privacy and Regulatory Compliance
4. Operational Efficiency

For each of these categories, a preliminary list of key issues provided the basis for discussion. The Workgroup reviewed every key characteristic allowing for an expanded understanding of each approach's strengths and limitations.

A. QUANTITATIVE ANALYSIS

The table served as the basis of a quantitative analysis of each distributed analysis method. The goal of the quantitative analysis was to obtain a numeric sense of the relative performance of each method through a consistent methodology. While small variations in score may more reflect the particulars of the scoring system rather than the exact performance of the methods, large gaps in score between one method and another are likely informative.

1. Methodology

For each criterion noted in the table, a score of 1 to 5 was assigned to each of the five methods. One was the lowest score and indicated that a method performed poorly with respect to the noted criterion; conversely, 5 was the highest score and indicated that a method performed well.

Each criterion was also assigned a relative weight in order to gauge the importance of each element as compared to the others and in the broader scheme of Mini-Sentinel. The scores were assigned as follows:

- 5 = Issue is of core significance
- 4 = Issue is significant
- 3 = Issue is a 'very nice to have'
- 2 = Issue is a 'nice to have'
- 1 = Issue is not important
- 0 = Issue is not feasible

The assignment of the weight was greatly informed by the discussions with the members of other distributed data network (see Section III). The areas identified as important or challenging in these discussions (including logistical challenges, staffing requirements at various sites, variations in installed software and computing facilities, and other noted issues) were rated appropriately per the Workgroup's assessment of the relevance to Mini-Sentinel and other investigators' prior experience.

A method's total score was calculated as the sum total of its criterion scores (1 to 5) multiplied by each criterion's relevance.

2. Public Health Authority versus Research

For those evaluation criteria that had to do with privacy and HIPAA compliance, we created two cases: an evaluation of each method under the assumption of Mini-Sentinel's operation as part of FDA's public health authority, as well as a use of Mini-Sentinel in a broader research setting. (The specifics of these two settings are described in Appendix B.)

B. EVALUATION TABLE

Characteristic	Covariate Sharing (CS)	Aggregated Data Keeping Small Cells (AGGKEEP)	Aggregated Data Removing Small Cells (AGGREMOVE)	Distributed Regression (DR)	Score-Based Pooling (SBP)	Meta-Analysis (META)	Relative Weight (PUBLIC HEALTH AUTHORITY)*	Relative Weight (RESEARCH)*	Notes on Methods	Notes on Relative Weight
Epidemiological Fundamentals										
Ability to obtain simple "Table 1" descriptive data (frequency, means, median, standard deviations, etc.) for the study population	5	5	5	5	5	5	5	5	A Table 1 is assumed to be generated separately and as such to be common across methods.	
Ability to control for confounding	5	5	3	5	5	5	5	5		
Ability to handle continuous covariates.	5	2	2	5	5	5	4	4	Continuous covariates must be categorized for the aggregated data approach.	
Ability to handle cohort designs	5	5	2	5	5	5	5	5	AGGREMOVE method may not be able to supply full follow-up time information, if stratified by days.	
Ability to handle self-controlled designs	5	4	4	4	5	5	5	5		
Ability to operate as well as possible with a rare exposure and rare outcome	4	4	2	4	3	4	3	3	This is a difficult case in which no method may perform optimally. Scores are hard to estimate with rare exposure or outcome. AGGREMOVE will require collapsing cells and losing information.	This is a less likely scenario and may require special handling no matter which method is used.
Ability to operate as well as possible with a rare exposure and common outcome	4	4	2	4	4	4	5	5	This is a rare case as few outcomes will be common. Disease risk scores will operate well in this case. AGGREMOVE will require collapsing cells and losing information.	
Ability to operate as well as possible with a common exposure and rare outcome	4	4	2	4	5	4	5	5	Propensity scores will operate well in this case. AGGREMOVE will require collapsing cells and losing information.	

Ability to operate as well as possible with a common exposure and common outcome	5	5	5	5	5	5	5	5	This is a rare case as few outcomes will be common.	
Ability to detect and handle treatment effect heterogeneity across pre-specified subgroups	5	5	5	5	5	5	5	5	No flexibility to account for non-pre-specified subgroups. For AGG and DR, subgroups must be handled at the site. For SBP, subgroup indicators must be shared.	
Ability to detect and handle treatment effect heterogeneity across sites	4	4	4	4	4	4	4	4		
Ability to detect and handle information content heterogeneity	3	3	3	3	5	4	5	5	SBP can use techniques such as hd-PS which extract maximal information content from each site.	
Ability to evaluate dose-response relationships	5	3	3	3	4	4	3	3	This can be viewed as a special case of subgroup analysis.	Dose information access is probably not as reliable in the M-S environment.
Ability to identify and handle (remove, do subgroup analysis, etc.) patients who may be overly influential to results.	5	2	2	2	5	2	5	5	DR may require re-running of models at the site to remove influential patients. Influential patients can be removed with AGG only if they are together in single cell(s).	Given the rare nature of many outcomes, some patients can be very influential.
Ability to evaluate whether treatments are used in similar circumstances across sites.	3	3	2	2	5	3	5	5	Important aspect of study. May require analysis beyond what is accomplished here. META would require additional diagnostic information.	
Ability to handle practice pattern, treatment usage, and patient heterogeneity across sites. Can sites be pooled?	5	3	2	3	4	2	5	5	Important aspect of study. May require analysis beyond what is accomplished here. META would require additional diagnostic information.	
Ability to determine if analytic/model assumptions hold and to perform diagnostic common measures	5	5	5	5	5	5	5	5	Diagnostics are common to all methodologies.	
Ability to determine if covariate adjustment was adequate	4	4	2	2	5	4	5	5	For SBP, this includes balance diagnostics.	

Ability to adjust for individual covariates specified at design time	5	5	3	5	5	5	5	5	AGGREMOVE's ability to adjust will depend on frequency of covariates.	
								0		
Analytic Flexibility								0		
Flexibility to adjust for individual covariates not foreseen at design time	4	1	1	1	2	1	2	2	For CS, additional covariates can be added to the extent that they are included with the dataset. For SBP, use of a technique like hd-PS will allow for additional covariates.	Assumes a certain level of 'agony' in doing a round-trip with sites. Can always accomplish the same goal by redesigning the study.
Flexibility to examine effects in non-pre-specified subgroups.	4	1	1	1	1	1	2	2		Assumes a certain level of 'agony' in doing a round-trip with sites. Can always accomplish the same goal by redesigning the study.
Hub's ability to alter analytic plan (ITT versus As-Treated, stratified versus matched, etc.)	5	5	5	1	5	1	5	5	For DR, new analytic code will need to be pushed out to sites.	
Hub's ability to create a summary confounder score	5	5	3	1	3	1	3	3	For SBP, a summary score is included	
Hub's ability to match patients on individual covariates	5	1	1	1	4	1	2	2		Generally important, but less so given the standard diagnostics that are available.
Hub's ability to determine which patients could plausibly have received either of the treatments under study, and to trim those who are not comparable	5	3	2	1	4	1	4	4	For AGG, entire cells must be trimmed.	
Hub's ability to perform post hoc sensitivity analyses using different methods on the same data	5	4	4	1	5	1	3	3		
Ability to support horizontal partitioning of data	5	3	3	1	4	1	5	5	M-S data environment will not change. SBP will require aggregating data from across databases to calculate the score, or will require two scores. AGG methods will require aggregation separately for each database. DR will need all data before	There may be key confounders or other items stored in other systems that may need to be linked.

									running regression model.	
Ability to support vertical partitioning of data	5	5	5	5	5	5	5	5	M-S data environment will not change.	Vertical partitioning is at the core of M-S.
Hub's ability to detect subtle errors in analysis at sites.	2	2	2	2	2	1	3	3	General: may be easy to detect obvious errors but harder to determine hidden errors.	This is generally important but many not be feasible.
Investigators' overall ability to understand and sense transparency in analyses and results (not a "black box")	5	4	4	2	4	2	4	4	A highly subjective criterion.	This ability flows from many other characteristics that are evaluated elsewhere on this table.
Transparency in results -- ability to see into the methods	5	4	3	4	4	1	4	4		Important to be able to 'see into' methods to validate results.
Privacy and regulatory compliance										
Compliance with HIPAA, HITECH, and other regulations that may apply	5	5	5	5	5	5	5	5		
Ability to distribute study data without compromising patient privacy in a PUBLIC HEALTH AUTHORITY setting	5	5	5	5	5	5	5	5	Per memo by KR, all methods are compliant under HIPAA law.	
Ability to distribute study data without compromising patient privacy requirements in a RESEARCH setting	2	2	4	4	4	5	0	5		M-S is expected to operate primarily in the public health authority setting
Ease of de-identifying information and creating a limited data set in a RESEARCH setting	2	2	5	5	5	5	0	2		
Need for DUA in a RESEARCH setting	2	2	2	2	2	5	0	2	Those methods that require a DUA are scored low; a DUA would be required for	

									use of a limited data set	
Ability to support anonymous linkage of patients, as required by participating sites in a PUBLIC HEALTH AUTHORITY setting	5	1	1	1	5	2	2	2	META patients cannot be linked.	
Ability to support anonymous linkage of patients, as required by participating sites in a RESEARCH SETTING	5	1	1	1	5	2	0	4		
Ability to transmit full patient covariate information (or the epidemiological equivalent, such as a propensity score) without sharing institutions' potentially proprietary information	1	2	4	4	4	5	5	5		
Operational Efficiency										
Ease of transferring information (as required) for analysis or summarization	3	4	4	4	4	5	5	5	For CS, data can be sent as-is but may require a more secure connection. For AGG, data must be a summarized. For SBP, a score model must be run. AGG data may require more explanation.	
Ease of handling updates in analysis plan (including need for re-distribution of SAS code)	4	4	4	2	4	4	4	4	For DR, new software may be required.	Many analyses may require "round trips". The system should be able to support those without too much overhead.
Ease of re-use of data for parallel questions, as required	5	3	3	3	4	3	2	2	For AGG and SBP, not all data may be in the dataset.	Reuse is not an important case, as most times, a new analysis will be kicked off. Cases such as sequential monitoring will be handled separately.
Computing time required at individual sites	5	4	4	2	4	3	2	2	Computing complexity varies for each approach. DR requires back-and-forth between sites and hub, depending on complexity of model.	

Computing time required at hub	4	5	5	2	4	5	2	2	DR requires back-and-forth between sites and hub, depending on complexity of model. META would require very little time at the hub.	
Predicted need for staff time in each participating site	4	4	4	4	4	3	3	3	META may require more staff time.	
Predicted level of statistical and programming expertise needed within each participating site	4	4	4	3	4	2	4	4	META may require more statistical training.	
Predicted ability to execute analysis in a timely fashion	4	4	4	4	4	4	4	4		If timeliness needs to be increased, additional staff or other remedies could be taken.
Predicted "scalability": whether additional studies will require less effort than initial studies	4	3	3	3	4	3	3	3	Depends on nature of follow-on studies. Methods with full data may be more straightforward than those that have summary data.	
RAW SCORE	**205**	**168**	**154**	**150**	**203**	**163**				
WEIGHTED SCORE (PUBLIC HEALTH AUTHORITY)	**780**	**678**	**593**	**596**	**780**	**625**				
WEIGHTED SCORE (RESEARCH)	**818**	**700**	**631**	**634**	**834**	**678**				

VII. CONCLUSIONS AND RECOMMENDATIONS

Over the course of this year-long project, the Workgroup performed a literature search, had conversations with experienced researchers in the field, consulted a leading lawyer in the area of patient privacy, and engaged in an in-depth analysis of five key methods for performing analyses in a distributed data network.

The recommendations of the Workgroup distinguish two separate scenarios: use of the Mini-Sentinel system in the context of the public health authority, and use more broadly under a medical research scenario. The basis of the recommendations – and the basis of today's Mini-Sentinel system – assumes the public health authority. However, we did consider the research case as well as Mini-Sentinel may broaden in scope in the future.

A. RECOMMENDATION IN BRIEF

In the context of the public health authority, this Workgroup primarily recommends the Score-Based Approach and also recommends the Covariate Sharing Approach. In the context of general research, this Workgroup uniquely recommends the Score-Based Approach. Overall, the Score-Based Approach serves today's Mini-Sentinel needs, will accommodate future growth in Mini-Sentinel's scope or use cases, and accomplishes the goals of the Mini-Sentinel System while sharing a minimal amount of data. Nonetheless, the ultimate choice between the two methods will be made on a case-by-case basis.

B. DISCUSSION OF RECOMMENDATION

Based on the group's scoring methodology and the discussions on the bi-weekly phone calls, two methods rose to the top as candidates for the Workgroup's recommendation for methodology. Each of these methods was in the top tier of score both when the surveillance activity was assumed to be conducted under public health authority and when a general approach to medical research was assumed.

The Covariate Sharing method tied for highest score among all the methods when operating under public health authority. The transparency of the method; light burden on participating sites with respect to packaging, analyzing, and transmitting data; and flexibility in the analysis to be undertaken were all key factors for Covariate Sharing's high score. However, there are several downsides to the method – mainly in the area of patient privacy and institution's ability to protect proprietary data – and these played a smaller factor in the quantitative scoring since many of requirements of HIPAA did not apply when Mini-Sentinel operated as a public health entity. Furthermore, a more qualitative rating system may additionally disfavor Covariate Sharing as an approach for its substantially larger need for raw data sharing. Finally, Covariate Sharing may place an administrative burden on partners that are Covered Entities, as without a data use agreement disclosure accounting may be required.

The Score-Based Approach method received the highest score among all the methods when operating under a general research environment and tied for highest score when operating under public health authority. The method is somewhat less transparent than the Covariate Sharing approach; variables that would have been shared "in the clear" must be summarized into a score that can be shared without risk of identifying patients. Despite this, the Workgroup determined that this method would support the analytic goals of Mini-Sentinel, though with some acceptable compromises with respect to analytic flexibility and features such as ad-hoc subgroup analyses. Importantly to many data partners, the Score-Based Approach will also protect partners' proprietary information whether or not de-identification of patients or limited data sets are required by law, and may have a lighter administrative burden; this may be an important concern for certain partners. The "cost" of the approach to partners is some increase in analytic capacity at the partners' sites in order to create the scores, though the Workgroup feels that this can be largely automated in SAS macros or other such facilities.

The Aggregated Data approaches were determined to not be feasible; when small cells were included, there were simply no great advantages as compared to Covariate Sharing but there were a fair number of disadvantages. When small cells were omitted or collapsed, the Aggregated Data approach was determined to not support the analytic goals of Mini-Sentinel.

The Distributed Regression approach had certain desirable properties, but was determined to not be robust enough for the variety of analyses envisioned by Mini-Sentinel nor flexible enough to support the analytic needs of the system. It remains an interesting option for certain applications.

Meta-Analysis has the great advantages of simplicity in analysis and no need to share data, but masks the complexity needed to create the meta-analytic estimates. The Workgroup determined that Meta-Analysis was too inflexible and too opaque for the robust operation of a distributed data network. The potential requirement for analytically-trained staff at each partner site was also a strong disadvantage.

When electing an approach to use broadly in Mini-Sentinel, it is important to emphasize that Mini-Sentinel places a high priority on minimizing the amount of potentially identifiable data that data partners need to share. This is the case even though the program is legally permitted to obtain identifiable information. The reasons for doing this are both to respect patients' confidentiality to the greatest degree possible, and also to lower the barriers to data partners' participation. Importantly, the data partners often have separate business reasons for wanting to minimize the amount of information they share.

C. RECOMMENDED COURSE OF ACTION

Given the desire by the Mini-Sentinel program to trade a moderate amount of operational complexity in exchange for reducing the amount of identifiable information the partners are required to share, the Workgroup chose the Score-Based Approach for its primary recommendation. In cases where parties accept greater information sharing, the Workgroup also endorses the Covariate Sharing approach. Both approaches will fulfill Mini-Sentinel's analytic needs.

Operationally, the Workgroup recommends that Mini-Sentinel build tools that ultimately support both methodologies.

One reasonable approach would be the following:

- Mini-Sentinel should build a general framework for requesting data items from Data Partners.
- For a given data item, the framework should support returning the item "in the clear" and/or as part of one or more scores. For example, a "history of MI" data item could be returned both as an individual covariate (perhaps for use as a subgroup identifier) as well as part of a propensity score.

- If Mini-Sentinel wishes to use the Score-Based Approach for a given study, then the data items requested should include minimal information (exposure status, outcome status, subgroup indicators, and the like) plus covariates summarized in scores.
- If Mini-Sentinel wishes to use the Covariate Sharing Approach for a given study, then the data items should all be requested "in the clear".
- Software for the analytic hub should be able to classify the returned data elements for proper analytic handling. Some analytic techniques would be standard adjustment for individual covariates, matching on scores, stratification on scores, and subgrouping by covariates.

VIII. REFERENCES BY TOPIC

A. GENERAL METHODS

1. AOCTG. Chemotherapy in advanced ovarian cancer: an overview of randomised clinical trials. Advanced Ovarian Cancer Trialists Group. BMJ 1991;303:884-93.
2. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol 1999;28:1-9.
3. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010;48:S114-20.
4. Brown J, Lane, K., Moore, K., Platt, R. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative: Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care; 2009.
5. Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: a new paradigm. Health Aff (Millwood) 2009;28:454-66.
6. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. Epidemiology 1993;4:295-302.
7. Friedenreich CM. Commentary: Improving pooled analyses in epidemiology. International Journal of Epidemiology 2002;31:86-7.
8. Greene SM, Geiger AM. A review finds that multicenter studies face substantial challenges but strategies exist to achieve Institutional Review Board approval. J Clin Epidemiol 2006;59:784-90.
9. Greene SM, Geiger AM, Harris EL, et al. Impact of IRB requirements on a multicenter survey of prophylactic mastectomy outcomes. Ann Epidemiol 2006;16:275-8.
10. Kahn M. A Pragmatic Framework for Single-Site and Multi-Site Data Quality Assessment in Electronic Health Record-Based Clinical Research.
11. Kulldorff M RLD, Margarette Kolczak, Edwin Lewis, Tracy Lieu, and Richard Platt. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. Sequential Analysis 2011;30:58–78.
12. Paul SR, Donner A. A comparison of tests of homogeneity of odds ratios in K 2 x 2 tables. Stat Med 1989;8:1455-68.
13. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;158:915-20.
14. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323-37.
15. Strom B. The future of pharmacoepidemiology. In: Strom B, ed. Pharmacoepidemiology. Chichester, UK: John Wiley & Sons; 2005.
16. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res 2011.

B. SHARING IDENTIFIABLE DATA

1. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Collaborative Group on Hormonal Factors in Breast Cancer. Lancet 1996;347:1713-27.
2. Smith-Warner SA, Spiegelman D, Ritz J, et al. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. Am J Epidemiol 2006;163:1053-64.

C. SHARING WITH AGGREGATED INFORMATION

1. Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. BMC Public Health 2006;6:235.
2. Wolfson M, Wallace SE, Masca N, et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010;39:1372-82.

D. SHARING WITH PROPENSITY SCORES

1. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149-56.
2. Herrinton LJ, Curtis JR, Chen L, et al. Study design for a comprehensive assessment of biologic safety using multiple healthcare data systems. Pharmacoepidemiol Drug Saf 2011;20:1199-209.
3. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. Pharmacoepidemiol Drug Saf 2010;19:848-57.
4. Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. Med Care 2010;48:S83-9.

E. DISTRIBUTED REGRESSION

1. Fienberg S, Karr, A., Nardi, Y., Slavkovic, A. “Secure” Log-Linear and Logistic Regression Analysis of Distributed Databases. In: 56th Session of the ISI; 2007; 2007.
2. Fienberg SE, William J. Fulp, Aleksandra B. Slavkovic and Tracey A. Wrobel. “Secure” Log-Linear and Logistic Regression Analysis of Distributed Databases. Lecture Notes in Computer Science 2006;4302:277-90.
3. Fienberg SE, Nardi, Y. and Slavković, A. B. Valid statistical analysis for Logistic regression with multiple sources. In: Lesk PKaM, ed. Proc Workshop on Interdisciplinary Studies in Information Privacy and Security -- ISIPS 2008. New-York: Springer-Verlag; 2009.
4. Ghosh J, Jerome P. Reitera, Alan F. Karr. Secure computation with horizontally partitioned data using adaptive regression splines. Computational Statistics and Data Analysis 2006.
5. Karr AF, Xiaodong Lin, Ashish P Sanil, Jerome P Reiter. Secure Regression on Distributed Databases. Journal of Computational and Graphical Statistics 2005;14:263-79.

6. Karr AF, William J. Fulp, Francisco Vera, S. Stanley Young, Xiaodong Lin, Jerome P. Reiter. Secure, Privacy-Preserving Analysis of Distributed Databases. TECHNOMETRICS 2007;49.
7. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des 2005;19:739-47.
8. Li L, Fang Zhang, Allyson M. Abrams, Ken Kleinman. A SAS Macro for Secure Logistic Regression on Multi-Site Electronic Health Databases. Boston, MA: Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute; 2010.
9. Sanil A, Karr, A., Lin, X., Reiter, J. Privacy Preserving Regression Modelling Via Distributed Computation. In: Tenth ACM SIGKDD Internat Conf on Knowledge Discovery and Data Mining. Seattle, Washington, USA; 2004:677-82.
10. Slavkovic AB, Yuval Nardi, Stephen Fienberg, Alan Karr, and Matthew Tibbits. “Secure” Logistic Regression with Distributed Databases. In: **PADM-ICDM**. Omaha, NE; 2007.
11. Zhang F, LingLing Li, Allyson M. Abrams., Ken Kleinman. A SAS Macro for Secure Linear Regression on Multi-Site Electronic Health Databases. Boston: Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute; 2010.

F. META-ANALYSIS

1. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Early Breast Cancer Trialists' Collaborative Group. Lancet 1992;339:1-15.
2. Beral V. "The practice of meta-analysis": discussion. Meta-analysis of observational studies: a case study of work in progress. J Clin Epidemiol 1995;48:165-6.
3. Berlin JA. Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. Am J Epidemiol 1995;142:383-7.
4. Chalmers TC. Problems induced by meta-analyses. Stat Med 1991;10:971-9; discussion 9-80.
5. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140:290-6.
6. Olkin I. Re: "A critical look at some popular meta-analytic methods". Am J Epidemiol 1994;140:297-9; discussion 300-1.
7. Shapiro S. Meta-analysis/Shmeta-analysis. Am J Epidemiol 1994;140:771-8.
8. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;309:1351-5.
9. Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet 1991;338:1127-30.

G. HMO RESEARCH NETWORK

1. Andrade SE, Raebel MA, Brown J, et al. Outpatient use of cardiovascular drugs during pregnancy. Pharmacoepidemiol Drug Saf 2008;17:240-7.
2. Andrade SE, Raebel MA, Brown J, et al. Use of antidepressant medications during pregnancy: a

multisite study. Am J Obstet Gynecol 2008;198:194 e1-5.

3. Brown J, John Holmes, Judy Maro, Beth Syat, Kimberly Lane, Ross Lazarus, Richard Platt. Design Specifications for Network Prototype and Cooperative To Conduct Population-Based Studies and Safety Surveillance, Effective Health Care Research Report No. 13. (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.). Rockville, MD: Agency for Healthcare Research and Quality; 2009 July 2009.
4. Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiol Drug Saf 2007;16:1275-84.
5. Chan K. HMO Research Network. In: Strom B, ed. Pharmacopidemiology. Chichester, UK: John Wiley & Sons; 2005.
6. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf 2001;10:373-7.
7. Raebel MA, Carroll NM, Andrade SE, et al. Monitoring of drugs with a narrow therapeutic range in ambulatory care. Am J Manag Care 2006;12:268-74.
8. Raebel MA, Lyons EE, Andrade SE, et al. Laboratory monitoring of drugs at initiation of therapy in ambulatory care. J Gen Intern Med 2005;20:1120-6.
9. Raebel MA, McClure DL, Chan KA, et al. Laboratory evaluation of potassium and creatinine among ambulatory patients prescribed spironolactone: are we monitoring for hyperkalemia? Ann Pharmacother 2007;41:193-200.
10. Raebel MA, McClure DL, Simon SR, et al. Laboratory monitoring of potassium and creatinine in ambulatory patients receiving angiotensin converting enzyme inhibitors and angiotensin receptor blockers. Pharmacoepidemiol Drug Saf 2007;16:55-64.
11. Raebel MA, McClure DL, Simon SR, et al. Frequency of serum creatinine monitoring during allopurinol therapy in ambulatory patients. Ann Pharmacother 2006;40:386-91.
12. Simon SR, Andrade SE, Ellis JL, et al. Baseline laboratory monitoring of cardiovascular medications in elderly health maintenance organization enrollees. J Am Geriatr Soc 2005;53:2165-9.

H. VACCINE SAFETY DATALINK

1. Vaccine Safety Datalink. (Accessed at http://www.cdc.gov/vaccinesafety/vsd/.)
2. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. Pediatrics 1997;99:765-73.
3. Go AS, Magid DJ, Wells B, et al. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. Circ Cardiovasc Qual Outcomes 2008;1:138-47.
4. Klein NP, Fireman B, Yih WK, et al. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. Pediatrics 2010;126:e1-8.

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

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

7. Velentgas P, Bohn RL, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. Pharmacoepidemiol Drug Saf 2008;17:1226-34.

I. SENTINEL

1. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010;48:S114-20.

2. 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:S45-51.

3. Herrinton LJ, Curtis JR, Chen L, et al. Study design for a comprehensive assessment of biologic safety using multiple healthcare data systems. Pharmacoepidemiol Drug Saf 2011;20:1199-209.

4. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med 2009;151:341-4.

5. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel Network--improving the evidence of medical-product safety. N Engl J Med 2009;361:645-7.

J. OMOP

1. Madigan D, Ryan P. What can we really learn from observational studies?: the need for empirical assessment of methodology for active drug safety surveillance and comparative effectiveness research. Epidemiology 2011;22:629-31.

2. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2011.

3. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010;153:600-6.

K. OTHER DISTRIBUTED NETWORKS

1. 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:S45-51.

2. Fireman B, Lee J, Lewis N, Bembom O, van der Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. Am J Epidemiol 2009;170:650-6.

3. Gliklich RE DN, eds. Registries for Evaluating Patient Outcomes: A User's Guide (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHSA29020050035I TO1.). Rockville, MD: Agency for Healthcare Research and Quality; 2007.

4. Magid DJ, Gurwitz JH, Rumsfeld JS, Go AS. Creating a research data network for cardiovascular disease: the CVRN. Expert Rev Cardiovasc Ther 2008;6:1043-5.
5. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med 2009;151:341-4.
6. McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A self-scaling, distributed information architecture for public health, research, and clinical care. J Am Med Inform Assoc 2007;14:527-33.
7. Moore KM, Duddy A, Braun MM, Platt R, Brown JS. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. Pharmacoepidemiol Drug Saf 2008;17:1137-41.
8. OSCAR - Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment. 2011. (Accessed at http://omop.fnih.org/oscar.)
9. Partnership OMO. Generalized Review of OSCAR Unified Checking. 2011.
10. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther 2011;90:883-7.
11. Wagner AK, Chan KA, Dashevsky I, et al. FDA drug prescribing warnings: is the black box half empty or half full? Pharmacoepidemiol Drug Saf 2006;15:369-86.

L. MISCELLANEOUS

1. 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:S45-51.
2. Fireman B, Lee J, Lewis N, Bembom O, van der Laan M, Baxter R. Influenza vaccination and mortality: differentiating vaccine effects from bias. Am J Epidemiol 2009;170:650-6.
3. Gliklich RE DN, eds. Registries for Evaluating Patient Outcomes: A User's Guide (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHSA29020050035I TO1.). Rockville, MD: Agency for Healthcare Research and Quality; 2007.
4. Magid DJ, Gurwitz JH, Rumsfeld JS, Go AS. Creating a research data network for cardiovascular disease: the CVRN. Expert Rev Cardiovasc Ther 2008;6:1043-5.
5. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med 2009;151:341-4.
6. McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A self-scaling, distributed information architecture for public health, research, and clinical care. J Am Med Inform Assoc 2007;14:527-33.
7. Moore KM, Duddy A, Braun MM, Platt R, Brown JS. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. Pharmacoepidemiol Drug Saf 2008;17:1137-41.
8. OSCAR - Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment. 2011. (Accessed at http://omop.fnih.org/oscar.)

9. Partnership OMO. Generalized Review of OSCAR Unified Checking. 2011.

10. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther 2011;90:883-7.

11. Wagner AK, Chan KA, Dashevsky I, et al. FDA drug prescribing warnings: is the black box half empty or half full? Pharmacoepidemiol Drug Saf 2006;15:369-86.

IX. APPENDIX

A. APPENDIX A: WORKGROUP MEMBERS

Brigham and Women's Hospital – Division of Pharmacoepidemiology

- Mary Kowal - Workgroup Project Manager
- Jessica Myers - Member
- Jeremy Rassen - Workgroup Leader
- Sebastian Schneeweiss - Member

Columbia University Department of Statistics

- David Madigan - Member

Food and Drug Administration

- Carlos Bell - Member
- Lauren Choi - Member
- Jane Gilbert - Member
- Laura Governale - Member
- Tarek Hammad - Member
- Monika Houstoun - Member
- Clara Kim - Member
- Carolyn McCloskey - Member
- Afrouz Nayernama - Member
- Antonio Paredes - Member
- Quocbao Pham - Member
- Robert Pratt - Member
- Marsha Reichman - Member
- Mary Beth Ritchey - Member
- Melissa Robb - Member
- Mitra Rocca - Member
- Azadeh Shoaibi - FDA Lead
- Jingwen Tan - Member
- Ram Tiwari - Member

HMORN: Group Health Research Institute

- Karin Johnson - Member

HMORN: Harvard Pilgrim Health Care Institute

- Meghan Baker - Member
- Jeff Brown - Member
- Roberta Constantine - FYI
- Kara Coughlin - MSOC Research Assistant
- Lingling Li - Member
- John Moran - Member
- Darren Toh - MSOC Scientific Lead
- Madhavi Vajani - MSOC Project Manager

Kaiser Permanente Colorado

- Marsha Raebel – Member

Observational Medical Outcomes Partnership (OMOP)

- Patrick Ryan - Member

University of Pennsylvania School of Medicine

- John Holmes - Workgroup Leader
- Kevin Haynes – Member

B. APPENDIX B: PRIVACY EVALUATION OF PROPOSED MINI-SENTINEL STATISTICAL METHODS

Coppersmith Schermer & Brockelman PLC

DATE: September 28, 2012

TO: Members of the Mini-Sentinel Data Sharing and Analysis in Distributed Data Setting Workgroup

FROM: Kristen Rosati, Coppersmith Schermer & Brockelman PLC

RE: Privacy Evaluation of Proposed Mini-Sentinel Statistical Methods

You have asked me to analyze five different statistical methods that might be employed by the Mini-Sentinel project, to determine whether those methods comply with the Health Insurance Portability and Accountability Act ("HIPAA") Privacy Rule.[1] As I explain below, each of the proposed methods will comply with HIPAA.

Section I of this memorandum discusses the HIPAA rules applicable to Mini-Sentinel activities. Section II then applies that discussion to each statistical method. Section III provides my recommendations for compliance.

I. HIPAA and Mini-Sentinel Activities

A. Applicability of the HIPAA Privacy Rule

1. HIPAA Applies to Covered Entities and Their Business Associates

The HIPAA Privacy Rule applies to HIPAA "covered entities." Covered entities include health plans, health care clearinghouses, and health care providers that engage in electronic "standard transactions" with health plans, such as electronic billing.[2]

After the effective date of final amendments to the HIPAA Privacy Rule, which are expected late summer this year, many requirements of the HIPAA Privacy Rule also will apply to "business associates" of HIPAA covered entities.[3] These regulations will implement section 13404 of the Health Information Technology for Economic and Clinical Health Act (HITECH) Act,[4] which requires HIPAA business associates to comply with many requirements of the Privacy Rule. Thus, to the extent that any of the data sources are business associates of HIPAA

[1] 45 C.F.R. Part 160 and Part 164, Subpart E.

[2] 45 C.F.R. § 160.102 (applicability); 45 C.F.R. § 160.103 (definition of covered entity).

[3] *See* 75 Fed. Reg. 40868 (July 14, 2010) (proposed amendments to the HIPAA Privacy Rule). *See 45* C.F.R. § 160.103 (definition of business associate).

[4] *Codified at* 42 U.S.C. § 17934.

covered entities, they also will have to comply with the HIPAA Privacy Rule requirements discussed in this memorandum, as of the effective date of final amendments to the Rule.

Some of the data sources participating in the Mini-Sentinel project may not be HIPAA covered entities or business associates, and thus will not be subject to the rules discussed in this memorandum.

2. HIPAA Applies to Individually Identifiable Health Information (Protected Health Information)

The HIPAA Privacy Rule protects "individually identifiable health information." Individually identifiable health information is "health information, including demographic information collected from an individual" that identifies an individual or where "there is a reasonable basis to believe the information can be used to identify the individual."[5] "Protected health information" (PHI) under the HIPAA Privacy Rule is a subset of individually identifiable health information that excludes certain health information held by employers and educational institutions."[6]

HIPAA permits two ways to "de-identify" information so that it is no longer protected by the Privacy Rule.[7] First, a covered entity may follow the "safe harbor" method of de-identification and remove or code all of the HIPAA "identifiers" in the information. These identifiers include all of the following data about individuals and their family members, household members, or employers:

- Name;
- Street address, city, county, precinct, or zip code (unless only the first three digits of the zip code are used and the area has more than 20,000 residents);
- All elements of dates (except year) directly related to an individual;
- Age over 89 (unless aggregated into a single category of age 90 and older);
- Telephone numbers;
- Fax numbers;
- Email addresses;
- Social security numbers;
- Medical record numbers;
- Health plan beneficiary numbers;

[5] 45 C.F.R. § 160.103 (defining "individually identifiable information" as "information that is a subset of health information, including demographic information collected from an individual, and: (1) Is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (i) That identifies the individual; or(ii) With respect to which there is a reasonable basis to believe the information can be used to identify the individual".

[6] 45 C.F.R. § 160.103 (defining "protected health information").

[7] 45 C.F.R. § 164.514(a)-(b).

- Account numbers;
- Certificate/license numbers;
- Vehicle identifiers, serial numbers, and license plate numbers;
- Device identifiers and serial numbers;
- Web Universal Resource Locators (URLs) and Internet Protocol (IP) addresses;
- Biometric identifiers, such as fingerprints;
- Full-face photographs and any comparable images; or
- Any other unique identifying number, characteristic, or code.

If a covered entity has actual knowledge that, even with these identifiers removed the remaining information could be used alone or in combination with other information to identify the individual, then the information still must be treated as PHI.[8]

You have asked me to explain in more detail how dates are protected. The Privacy Rule treats as identifiers "all elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death." An OCR/NIH fact sheet explains that time periods of less than one year are considered to be elements of dates under § 164.514(b)(1) of the Privacy Rule:

> Q: May information de-identified under the Privacy Rule's "safe-harbor" method contain a data element that identifies a time period of less than a year (e.g., the fourth quarter of a specific year)?
>
> A: No. The Privacy Rule's "safe-harbor" method for de-identifying health information requires removal of, among other elements, all elements of dates directly related to an individual, except for year. Thus, a data element such as the fourth quarter of a specified year must be removed if a covered entity intends to de-identify data using the "safe-harbor" method. However, fewer identifiers may need to be removed under the Privacy Rule's alternative method for deidentification, where a qualified statistician, applying generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, determines that the risk of re-identification is very small. Thus, it may be possible for certain elements of dates to be considered de-identified where this second method allows it. See section 164.514(b)(1) of the Privacy Rule.
>
> As an alternative to de-identified data, the Privacy Rule would permit a covered entity to use or disclose information about dates in the form of a limited data set.

See "Health Services Research and the HIPAA Privacy Rule," NIH Pub. No. 05-5308, available at http://privacyruleandresearch.nih.gov/pdf/HealthServicesResearchHIPAAPrivacyRule.pdf.

[8] 45 C.F.R. § 164.514(b)(2).

So, any time interval that reveals a specific day, week, month, quarter or other time period within a particular year would be treated as identifier. For example, a data cut that includes only information from patients that received a medical product of interest in June 2010 would be treated the same as disclosing information that the date of medical product administration for those patients was June 2010. On the other hand, descriptions of time that are specified as an offset from a date coded only as a year would not be PHI. For example, if a dataset contained the information that a patient was exposed in 2010 and that she experienced an adverse event 127 days after her exposure, the notation of "127 days" would not be PHI because that would not reveal a specific month or other specific time period within that year. As another example, "Month 1" in a dataset is not a HIPAA identifier if it does not reveal which specific month it is.

If identifiers are coded before access, review, collection or release for the research, the code may not be derived from any information about the individual. For example, the code may not be derived from the individual's social security number, medical record number or name (such as initials), and may not be capable of being translated to identify the individual.

The second method of de-identification is to have a qualified statistical expert determine that the risk is very small that the information could be used alone, or in combination with other available information, to identify the patient.[9] The statistical expert must be a person with knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information non-individually identifiable, and must document the methods and results of the analysis that justifies the conclusion of very small risk.[10] For this analysis, whether or not the "identifiers" are in the information is not relevant. For example, a statistical expert could conclude that there is a very small risk of identification if certain dates of services are present in the information.

3. HIPAA Applies to Internal Use, As Well as External Disclosure, of Protected Health Information

The HIPAA Privacy Rule applies to the internal use of PHI by covered entities, as well as the external disclosure of PHI to third parties.[11] Use or disclosure of PHI is permitted if the requirements of at least one of the provisions in the HIPAA Privacy Rule are met.[12]

[9] 45 C.F.R. § 164.514(b) ("(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination."

[10] *Id.*

[11] *See* 45 C.F.R. § 164.502 "(a) Standard. A covered entity may not use or disclose protected health information, except as permitted or required by this subpart or by subpart C of part 160 of this subchapter."). *See also* 45 C.F.R. § 160.103 (defining "use" and "disclosure").

As applied to Mini-Sentinel activities, that means that covered entities' *internal* use of their PHI for Mini-Sentinel purposes must comply with one of the rules discussed below. In other words, it does not obviate the need to comply with HIPAA simply by avoiding disclosure of PHI to the Operations Center or other Mini-Sentinel participants. This issue will be addressed in the analysis below.

B. Disclosures of Protected Health Information for Public Health Activities

The provision of PHI to the Food and Drug Administration ("FDA"), the Mini-Sentinel Operations Center, and other participants subcontracted to the Operations Center, is to support a public health activity and is thus permitted under the HIPAA Privacy Rule without patient permission (called an "authorization" under HIPAA). The HIPAA Privacy Rule permits covered entities to disclose PHI for a variety of public health purposes, including to:

> [A] public health authority that is authorized by law to collect or receive such information for the purpose of preventing or controlling disease, injury, or disability, including, but not limited to, the reporting of disease, injury, vital events such as birth or death, and the conduct of public health surveillance, public health investigations, and public health interventions; or, at the direction of a public health authority, to an official of a foreign government agency that is acting in collaboration with a public health authority.[13]

The FDA is a "public health authority" under HIPAA, which is defined as:

> an agency or authority of the United States, a State, a territory, a political subdivision of a State or territory, or an Indian tribe, or a person or entity acting under a grant of authority from or contract with such public agency, including the employees or agents of such public agency or its contractors or persons or entities to whom it has granted authority, that is responsible for public health matters as part of its official mandate.[14]

[12] Where multiple rules exist that may permit use or disclosure of PHI, it is not necessary to meet the terms of all of those rules. See 45 C.F.R. § 164.502(a)(1) "Permitted uses and disclosures. A covered entity is permitted to use or disclose protected health information as follows:
(i) To the individual;
(ii) For treatment, payment, or health care operations, as permitted by and in compliance with §164.506;
(iii) Incident to a use or disclosure otherwise permitted or required by this subpart, provided that the covered entity has complied with the applicable requirements of §164.502(b), §164.514(d), and §164.530(c) with respect to such otherwise permitted or required use or disclosure;
(iv) Pursuant to and in compliance with a valid authorization under §164.508;
(v) Pursuant to an agreement under, or as otherwise permitted by, §164.510; and
(vi) As permitted by and in compliance with this section, §164.512, or §164.514(e), (f), or (g)."

[13] 45 C.F.R. §164.512(b)(1)(i).

[14] 45 C.F.R. §164.501 (emphasis added).

The release of PHI to the FDA for purposes of medical product safety surveillance is for the "conduct of public health surveillance" purposes, as contemplated by the Rule.[15]

Moreover, the Mini-Sentinel Operations Center and its subcontractors are also functioning as "public health authorities," because they are acting under contract with or under a grant of authority from the FDA. The Mini-Sentinel Operations Center is performing its functions under contract with the FDA. Moreover, even though the Operations Center subcontractors do not have a direct contract with the FDA, FDA has issued a letter to the Mini-Sentinel Operations Center explaining that both the Mini-Sentinel Operations Center and its subcontractors are acting under a grant of authority from the FDA.[16] Data sources thus may release PHI to the Mini-Sentinel Operations Center and its subcontractors as "public health authorities" for the purpose of the Mini-Sentinel pilot medical product safety surveillance queries. The internal use of PHI by the Operations Center's subcontractors for public health activities similarly would be permitted under the HIPAA Privacy Rule.[17]

Where a use or disclosure of PHI is to a public health authority, the HIPAA Privacy Rule *does not* require the covered entity to have an IRB or Privacy Board determine whether the covered entity may make the use or disclosure.

C. Disclosures of Protected Health Information for Research

The Department of Health and Human Services (HHS) Office for Human Research Protections (OHRP) has concluded that activities related to the Sentinel Initiative are <u>not</u> research that requires review by an Institutional Review Board (IRB).[18] Thus, the use and

[15] "[T]he Privacy Rule specifically permits covered entities (such as pharmacists, physicians or hospitals) to report adverse events and other information related to the quality, effectiveness and safety of FDA-regulated products both to the manufacturers and directly to FDA." *See* http://www.fda.gov/medwAtch/hipaa.htm (citing HHS Office for Civil Rights Guidance Explaining Significant Aspects of the Privacy Rule at page 28).

[16] See July 19, 2010 Letter from Dr. Rachel Behrman, FDA, to Dr. Richard Platt, Harvard Medical School and Harvard Pilgrim Health Care.)

[17] 45 C.F.R. § 164.512(b)(2) ("Permitted uses. If the covered entity also is a public health authority, the covered entity is permitted to use protected health information in all cases in which it is permitted to disclose such information for public health activities under paragraph (b)(1) of this section.").

[18] January 19, 2010 Letter from Dr. Jerry Menikoff, Director of the OHRP, to Dr. Rachel Behrman, Acting Associate Director of Medical Policy, Center for Drug Evaluation and Research at the FDA, explaining that OHRP "has determined that the regulations this office administers (46 CFR part 46) do not apply to the activities that are included in the [FDA] Sentinel Initiative.") Dr. Behrman then wrote a letter to Dr. Richard Platt at Harvard Pilgrim Health Care (the Mini-Sentinel's prime contractor managing the Operations Center), providing Dr. Menikoff's letter and concluding that the OHPR's "assessment applies to the work being conducted by [Harvard Pilgrim Health Care] and its subcontractors under contract number HHSF2232009100061, as the purpose of this contract is to carry out Sentinel Initiative activities that are included in the [FDA] Sentinel Initiative." (See April 2, 2010 Letter from Dr. Rachel Behrman, FDA, to Dr. Richard Platt, Harvard Medical School and Harvard Pilgrim Health Care.)

disclosure of information for Mini-Sentinel purposes is not "research." This means that data sources providing information for Mini-Sentinel purposes are not required by federal regulation to obtain approval of their IRBs for participation in Mini-Sentinel, and are not required to obtain a determination from their IRBs that these activities are "exempt."

For reference purposes, however, the remainder of this section sets forth the HIPAA regulations application to research. Under the HIPAA Privacy Rule, covered entities may use PHI internally for research or disclose PHI externally to third parties for research only if the requirements of at least one of nine rules below are met:[19]

1. The research involves only de-identified data;

2. The research uses or discloses a "Limited Data Set" and the covered entity has a "Data Use Agreement" in place with the recipient of the PHI;

3. The research subject or the subject's authorized representative has signed a written HIPAA authorization;

4. An IRB has waived the requirement for authorization;

5. The activities are just to prepare for research and required representations are obtained from the researchers;

6. The use or disclosure is for patient recruitment purposes, within the limits described below;

7. The research involves only the information of decedents and required representations are obtained from the researchers;

8. The disclosure of the PHI is required by law; or

9. The research is "grandfathered" under the HIPAA rules.

1. The Research Involves Only De-identified Information

The HIPAA Privacy Rule protects only "individually identifiable health information." Information that has been appropriately de-identified thus may be used or disclosed without restrictions under the Privacy Rule. HIPAA permits two ways to "de-identify" information,[20] both of which are discussed in Section A above.

[19] 45 C.F.R. § 164.512(i) (general rules for use and disclosure of patient information for research). Other HIPAA rules are cited as applicable.
[20] 45 C.F.R. § 164.514(a)-(b).

2. The Research Uses or Discloses a "Limited Data Set"

This is the HIPAA compliance method most commonly used in health services research. A "Limited Data Set" is partially de-identified patient information. A Limited Data Set excludes all of the "identifiers" listed in Section A above, except that a Limited Data Set may include: (1) geographic designations above the street level or PO Box; (2) dates directly related to a patient; or (3) any other unique identifying number, characteristic, or code that is not expressly listed as an "identifier."[21] A covered entity may disclose a Limited Data Set for research, public health or "health care operations" purposes if the recipient signs a "Data Use Agreement" in which the recipient agrees to protect the confidentiality of the information.[22]

A Data Use Agreement that complies with the HIPAA Privacy Rule must include the following items:

> (A) Establish the permitted uses and disclosures of such information by the limited data set recipient. The data use agreement may not authorize the limited data set recipient to use or further disclose the information in a manner that would violate the requirements of this subpart, if done by the covered entity;
>
> (B) Establish who is permitted to use or receive the limited data set; and
>
> (C) Provide that the limited data set recipient will:
> (1) Not use or further disclose the information other than as permitted by the data use agreement or as otherwise required by law;
> (2) Use appropriate safeguards to prevent use or disclosure of the information other than as provided for by the data use agreement;
> (3) Report to the covered entity any use or disclosure of the information not provided for by its data use agreement of which it becomes aware;
> (4) Ensure that any agents, including a subcontractor, to whom it provides the limited data set agrees to the same restrictions and conditions that apply to the limited data set recipient with respect to such information; and
> (5) Not identify the information or contact the individuals.[23]

It is not necessary to have a separate Data Use Agreement for each individual disclosure. Rather, the Privacy Rule permits a blanket Data Use Agreement, as long as the agreement meets the requirements in the list above.[24] For example, each data source may have one Data Use

[21] 45 C.F.R. § 164.514(c).

[22] Id.

[23] 45 C.F.R. § 164.514(e)(4).

[24] The Preamble to the Privacy Rule indicates substantial flexibility in how Data Use Agreements may be structured. *See* 67 Fed. Reg. at 53236, 53237 (Aug. 14, 2002).

Agreement in place with the Operations Center to support all disclosures of PHI for the Mini-Sentinel project.

3. The Subject or the Subject's Authorized Representative Has Signed a Written HIPAA Authorization

This is the most common HIPAA compliance method in clinical trial, where there are face-to-face interactions with individuals. A HIPAA-compliant authorization form must include a number of items:[25]

- A specific and meaningful description of the PHI to be used or disclosed in the research (such as the subject's medical records or other more limited portions of the record, such as laboratory results);
- The name or specific identification of the persons or class of persons authorized to make the disclosure (such as the subject's physicians and treating hospitals);
- The name or specific identification of the persons or class of persons who will have access to the PHI (such as the research site, principal investigator, IRB, sponsor, other third parties involved in the research, data safety monitoring board, FDA, and HHS);
- A description of the specific research protocol or study;
- An expiration date or event (such as the end of the study), or a statement that the authorization has no expiration;
- A statement of the subject's right to revoke the authorization in writing and a description of how to do so;
- A statement that the subject may not revoke the authorization as to information already disclosed for the research where the information is necessary to maintain the integrity of the study data, or a description of other exceptions where the subject may not revoke the authorization;
- A statement that the entity disclosing the PHI may not condition treatment, payment, enrollment or eligibility for benefits on the subject signing the authorization. If the individual will not be allowed to participate in the clinical trial without signing the authorization, the authorization must include a statement to that effect;
- A statement that the information disclosed for the research may be subject to redisclosure by the recipient and no longer be protected by the federal privacy rule;
- If the subject will not be given access to medical records during the study, a statement that the subject agrees to the denial of access when consenting to participate in the study, and that the right of access to the records will be reinstated upon completion of the study
- The subject's signature and the date of signature; and
- If the authorization is executed by a personal representative of the subject (the subject's health care decision maker), a description of that person's authority to act for the subject.

[25] 45 C.F.R. § 164.508.

4. An IRB Waives the Requirement for Authorization

If it is not feasible to get research subjects' authorization, researchers may ask an IRB to waive authorization. To have the IRB grant this request, the researcher must demonstrate three things:

1. The use or disclosure of the subjects' identifiable information involves no more than minimal risk to their privacy, based on: (a) an adequate plan to protect information identifying the subjects from improper use and disclosure; (b) an adequate plan to destroy information identifying the subjects at the earliest opportunity consistent with conduct of the research (unless there is a health or research justification for retention or if retention is required by law); and (c) adequate written assurances that the information identifying the subjects will not be reused or disclosed to any other person or entity, except as required by law, for authorized oversight of the study, or for other research permitted by the rules;

2. The research could not practicably be conducted without the waiver or alteration of authorization; and

3. The research could not practicably be conducted without access to and use of information identifying the subjects.

If the researchers can get HIPAA authorizations from the research subjects for some purposes but not others, the researchers can ask the IRB for partial waiver or alteration of the authorization. For example, researchers can ask the IRB to waive authorization for the initial review of records to determine which patients may be appropriate subjects, or may ask the IRB to approve verbal authorization if the contact with the subjects will be by phone.

5. The Activities Are to Prepare for Research

If researchers merely want to access, review or collect PHI to prepare for research, researchers may obtain that information if they provide the covered entity with the following representations in writing:

1. The PHI is sought solely to prepare for research;

2. The PHI is necessary to prepare for research; and

3. No information identifying individuals will be removed from the premises in the course of the review.

Activities to prepare for research include activities such as preparing a research protocol or developing a research hypothesis, identifying prospective research participants, or screening patient records to identify whether there are a sufficient number of patients at a facility to function as a site for a clinical trial.[26] Contacting patients to solicit participation in a clinical trial is not an activity to prepare for research,[27] and is covered in Section 6 below.

If researchers will need to remove the information from the covered entity's premises to review it to prepare for research, the researchers must ask the IRB to waive authorization instead, or another HIPAA option must be satisfied. In its guidance document, entitled "Health Services Research and the HIPAA Privacy Rule," OCR provided more details on when remote access to a server containing PHI is removing the PHI from the premises.[28]

6. The Use or Disclosure Is for Study Recruitment

HIPAA permits the use or disclosure of PHI for patient recruitment.[29] First, a health care provider may contact the provider's own patients to determine if the patients are interested in participating in a clinical trial. If the provider or the provider's employees contact the providers' own patients, that use of PHI is for either "treatment" (if the clinical trial involves treatment) or "health care operations" purposes, both of which are permitted without patient authorization under HIPAA.[30] The health care provider also may use a non-employed third party (including the researcher) to contact patients for recruitment purposes, but would first have to obtain a business associate agreement with the third party.[31] Finally, the researcher can request an IRB to partially waive authorization under Section 4, so that authorization is not required for the initial contact, but will be sought for enrollment in the study. Contacting patients for recruitment is not a "preparatory to research" activity under Section 5 above.[32]

7. The Research Involves Only the PHI of Decedents

Where the research involves only the information of deceased individuals, researchers may access this information if they provide the covered entity with the following representations in writing:

[26] *See* Clinical Research and the HIPAA Privacy Rule (NIH), at p, 11, available at http://privacyruleandresearch.nih.gov/pdf/clin_research.pdf.

[27] *Id.*

[28] http://privacyruleandresearch.nih.gov/pdf/HealthServicesResearchHIPAAPrivacyRule.pdf.

[29] HHS, *Clinical Research and the HIPAA Privacy Rule*, p. 4 (NIH 6/22/04), at 11, available at http://privacyruleandresearch.nih.gov/clin_research.rtf.

[30] 45 C.F.R. § 164.501 and § 164.506.

[31] 45 C.F.R. § 164.502(e) and § 164.504(e).

[32] *See* OHRP, *Clinical Research and the HIPAA Privacy Rule*, p. 4 (NIH 6/22/04), at http://privacyruleandresearch.nih.gov/clin_research.rtf ("Under the "preparatory to research" provision, covered entities may use or disclose PHI to researchers to aid in study recruitment. The covered entity may allow a researcher, either within or outside the covered entity, to identify, but not contact, potential study participants under the "preparatory to research" provision.").

1. The use or disclosure of information is sought solely for the research on the information of decedents;

2. The information is necessary for the research; and

3. The researcher will provide documentation of the death of the research participants upon request.

8. The Disclosure of the PHI Is Required by Law

HIPAA permits the disclosure of PHI if that disclosure is required by another law.[33] HIPAA covered entities thus may disclose PHI to the Food and Drug Administration as required by the FDA regulations, the Office for Human Research Protections as required by the Common Rule, and other government agencies if required by statute or regulations.

9. The Research is "Grandfathered" under the HIPAA Privacy Rule

Research is "grandfathered" under HIPAA if the participant signed an informed consent before April 14, 2003 (and the informed consent has not been modified since that date) or if the IRB waived informed consent before April 14, 2003.[34] This does not apply to any subjects enrolled in a study after April 14, 2003 or to subjects who signed a new informed consent document after this date. If research is grandfathered under HIPAA, researchers may continue to use the subject information they have and also may continue to collect information from the subject.

D. Data "Curation" is a Health Care Operation Permitted under HIPAA

In some of the statistical methods described below, a data source will curate its PHI in order to produce data sets to send to the Operations Center. This internal use of PHI to curate the data for this purpose is itself a "health care operation" under HIPAA that is permitted without individual authorization.[35] In other words, covered entities may utilize their own PHI to create data sets to disclose to others (as long as the disclosure of the data set produced is permitted by another rule).

[33] 45 C.F.R. § 164.512(a).

[34] 45 C.F.R. § 164.532(c).

[35] 45 C.F.R. § 164.501(defining "health care operations" as including "creating de-identified health information or a limited data set" … "[c]onsistent with the applicable requirements of §164.514."

E. Verification of Identity and Authority to Request Protected Health Information

To disclose PHI, data sources must confirm the recipient's identity and that the recipient has the legal authority to request the PHI.[36] A covered entity is entitled to rely on written confirmation on FDA letterhead that the Mini-Sentinel Operations Center and its subcontractors are acting on behalf of the FDA, and that they have the legal authority to request PHI for the Mini-Sentinel project.[37] FDA has issued a letter to the Mini-Sentinel Operations Center explaining that both the Mini-Sentinel Operations Center and its subcontractors are acting under a grant of authority from the FDA, pursuant to the legal authority provided by the FDAAA.[38]

F. Compliance with the Minimum Necessary Standard

HIPAA covered entities must observe the "minimum necessary standard" in using or disclosing PHI for any purpose other than treatment. This simply means that a covered entity must make reasonable efforts to limit the information to the minimum amount of information that is necessary to accomplish the intended purpose of the use or disclosure.[39] A covered entity may not disclose the entire medical record unless there is a specific justification for doing so.[40]

Under the HIPAA Privacy Rule, a covered entity may rely on a public health authority's determination that the data requested are the minimum necessary data that the agency needs to fulfill the purpose of its request.[41] When FDA (or the Operations Center or its subcontractors,

[36] 45 C.F.R. § 164.514(h)(1)(i).

[37] 45 C.F.R. § 164.514(h)(2)(ii)(C) (allowing a covered entity, when making disclosure to a person acting on behalf of a public official, to rely on "a written statement on appropriate governmental letterhead that the person is acting under the government's authority or other evidence or documentation of the agency, such as a contract for services … that establishes that the person is acting on behalf of the public official"; 45 C.F.R. § 164.514(h)(2)(iii)(A) (permitting a covered entity to rely on the written statement of a public agency regarding the legal authority under which it is requesting PHI, or an oral statement if a written statement is impracticable).

[38] See July 19, 2010 Letter from Dr. Rachel Behrman, FDA, to Dr. Richard Platt, Harvard Medical School and Harvard Pilgrim Health Care.

[39] 45 C.F.R. § 164.502(b)(1).

[40] 45 C.F.R. § 164.514(d)(5).

[41] See 45 C.F.R. § 164.514(d)(3)(iii) ("A covered entity may rely, if such reliance is reasonable under the circumstances, on a requested disclosure as the minimum necessary for the stated purpose when: (A) Making disclosures to public officials that are permitted under § 164.512, if the public official represents that the information requested is the minimum necessary for the stated purpose." While §13405(b) of the Health Information Technology for Economic and Clinical Health Act (the HITECH Act), codified at 42 U.S.C. § 17935, contains a provision that requires covered entities to determine what is the minimum amount of PHI for a disclosure, the proposed amendments to the HIPAA Privacy Rule to implement the HITECH Act do not modify a covered entity's ability to rely on minimum necessary representations by public officials. (See Notice of Proposed Rule Making, "Modifications to the HIPAA Privacy, Security, and Enforcement Rules under the [HITECH] Act," at 75 Fed. Reg. 40868 (July 14, 2010).

acting on behalf of FDA) sends a query to a covered entity, Mini-Sentinel policies require the request to be limited to what is required to evaluate the particular medical product safety issue. Covered entities thus may rely on these public health authority requests as being limited to the minimum amount of PHI necessary for the Mini-Sentinel activities.

Moreover, §13405(b) of the Health Information Technology for Economic and Clinical Health Act (the HITECH Act)[42] contains a provision that specifies that a covered entity must limit PHI to a "Limited Data Set" if practicable, or if that is not practicable, to the minimum necessary to accomplish the intended use, disclosure or request. While this statutory provision is not yet reflected in regulation, it does clarify that information restricted to a Limited Data Set will comply with the minimum necessary standard.

G. The Accounting Requirement

The HIPAA Privacy Rule currently requires covered entities to provide an "accounting" of disclosures of PHI to individuals at their request, with various exceptions, including disclosures that are made for treatment, payment and health care operations.[43] Section 13101 of the HITECH Act requires HHS to issue standards for accounting requirements. The OCR issued proposed regulations that will change the accounting requirement;[44] these regulations have not yet been issued in final form.

As explained in Section C above, the HIPAA Privacy Rule permits a covered entity to release a "Limited Data Set" for research, public health and health care operations purposes, as long as the covered entity first obtains a Data Use Agreement with the recipient of the Limited Data Set.[45] The Rule does not require a covered entity to include a disclosure of a Limited Data Set in an accounting, as long as a Data Use Agreement is in place.[46]

If the disclosure is to a "public health authority," that disclosure does not need to be limited to a Limited Data Set; rather, covered entities may release fully-identifiable PHI to public health authorities or may release a Limited Data Set without a Data Use Agreement in place.[47] However, a covered entity must include in an accounting, a disclosure to a public health authority if the covered entity discloses full PHI to the public health authority, or if it discloses a Limited Data Set to a public health authority without a Data Use Agreement in place. I thus recommend obtaining a Data Use Agreement for disclosures of Limited Data Sets to public health authorities, if possible.

[42] *Codified at* 42 U.S.C. § 17935.
[43] 45 C.F.R. § 164.528.
[44] *See* 76 Fed. Reg. 31426 (May 31, 2011).
[45] 45 C.F.R. § 164.514(d).
[46] *See* 45 C.F.R. § 164.528(a)(1) (viii) (exempting disclosures from the accounting requirement if they are "part of a limited data set in accordance with §164.514(e)").
[47] 45 C.F.R. § 164.512(b).

II. Application to Proposed Statistical Methods

As you have described to me, the goal for choosing a statistical method is to meet the goals of robust analytic integrity and flexibility, while simultaneously maintaining strong privacy protection for patients and data sources. This section explains each method and then applies the HIPAA legal standards discussed above. If I have not explained any of these statistical methods correctly, please let me know and I will re-evaluate my conclusions.

A. Method One: Analysis with Full Covariate Sharing

Conclusion: This method complies with the HIPAA Privacy Rule for two reasons. One, the Operations Center and its subcontractors are functioning as "public health authorities" on behalf of the FDA; use of PHI by or disclosure of PHI to the Operations Center or its subcontractors is permitted as a public health activity without individual authorization. Two, the information included in this method constitutes a "Limited Data Set" under the HIPAA Privacy Rule; use by or disclosure to the Operations Center or its subcontractors is permitted for research purposes if the covered entity has a Data Use Agreement in place with the recipient. Moreover, internal data curation to produce the Limited Data Set is a health care operation permitted under HIPAA without individual authorization.

Discussion: In this approach, each participating site creates an analytic dataset based on a protocol developed and provided by the Operations Center. The participating site then transmits its dataset to the Operations Center, which runs analytic models of interest on the submitted data.

These datasets will include individual-level data, including information related to the following data elements:

- Exposure status: whether the individual has been prescribed the target medical product being evaluated;
- Date that the target medical product was prescribed to the individual;
- Occurrence of outcome events: whether the individual has experienced an adverse health outcome, such as a heart attack, stroke, and the like;
- Details on the individual's medical encounters before and downstream of the exposure being evaluated. These details will include medical information related to the existence of co-morbidities, medical procedures, pre-existing medical conditions, and other conditions that may affect the analysis of whether the target medical product caused or potentially caused the adverse event observed in that individual.

Under this method, no information that *directly* identifies an individual will be transmitted to the Operations Center, such as name, address, social security number, medical record number or health plan membership number. However, the transmission of full covariate information likely will include HIPAA "identifiers," such as dates directly related to individuals (i.e., medical product exposure dates, dates of adverse events, dates related to previous medical diagnoses, etc.) and the geographic location of the individual.

When any of the HIPAA identifiers are included, that information technically is PHI under HIPAA, unless the information is certified as "de-identified" by a statistician. (See discussion in Section I(A), above.) My understanding is that techniques will be used to lessen the amount of PHI transmitted, such as such age categorization versus actual birth dates, but some PHI will be transmitted. I do not know whether there are plans to obtain a statistical certification of de-identification.

As described in Section I above, it does not violate the HIPAA Privacy Rule to disclose PHI under this statistical method to the Operations Center for Mini-Sentinel activities. There are two reasons why this method complies with the HIPAA Privacy Rule. First, as long as the disclosure is for work contracted through the FDA, it is for "public health activities."

Second, if the disclosure is for "research" purposes, disclosure of identifiers limited to dates related to individuals and geographic designations is a "Limited Data Set" that is permitted as long as the recipient has signed a Data Use Agreement. I thus recommend that each data source sign a Data Use Agreement with the Operations Center and any subcontractors that will receive the Limited Data Set. This will ensure compliance with the HIPAA research rules described in Section I (C) above, and also will obviate the need for data sources to list such disclosures in their "accountings" to individuals as described in Section I(G) above. The data curation of the data sources to produce a Limited Data Set constitutes a "health care operation" permitted under the Privacy Rule.

Finally, because this information is restricted to data elements permitted in a Limited Data Set, it complies with the HIPAA minimum necessary standard.

B. Method Two: Aggregated Data Approach

Conclusion: This method complies with the HIPAA Privacy Rule for two reasons. One, the Operations Center and its subcontractors are functioning as "public health authorities" on behalf of the FDA; use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization. Two, the information included in this method constitutes a "Limited Data Set" under the HIPAA Privacy Rule; disclosure to the Operations Center or its subcontractors is permitted for research purposes if the disclosing covered entity has a Data Use Agreement in place with the recipient. Moreover, internal data curation is a health care operation permitted under HIPAA without individual authorization.

Discussion: This method employs the technique of aggregating or collapsing like individuals' information into cells, after which adjustments are made for confounders based upon the counts of patients in each cell. These aggregate cells will contain exposure, outcome status, follow-up time, and covariate information, but will not be individual-level data (unless the cell includes only one individual). You provided an example of a cell that represents individuals who share the following combination of criteria:

Exposed?	Yes
Outcome event?	No
Followed for 180 days?	Yes
Aged 75-85?	Yes
Female?	No
History of cardiovascular disease?	No

The cell would indicate the number of patients that met these criteria. The total number of cells transmitted is determined by the number of observed combinations of exposure, outcome, follow-up time, and other covariates. No additional covariate information beyond the minimum required for the study would be sent to the Operations Center. The Operations Center would then run analytic models of interest on the aggregated data.

While no specific individual-level information is sent, "small" cells (with varying thresholds of <6 and <11 individuals) would be transmitted. With numerous covariates and rare outcomes, it would be anticipated that cells with only one individual would be common. The inclusion of small cells with one individual's information may constitute PHI, if that cell represents a HIPAA identifier. For example, if the data included in a cell of one is a date related to individuals (such as the month or day a medical product is prescribed to that individual), that small cell information would be treated as PHI and be subject to the same analysis as in Method One.

Removal of those small cells would create "de-identified" information because the lack of individual-level HIPAA identifiers means the information is no longer PHI under HIPAA. While de-identified information may be used without restriction under HIPAA, you have explained that this approach diminishes the effectiveness of the analytic study. As you have explained, if cells with fewer than 11 patients were simply dropped from the study, the statistical power and flexibility required to reveal rare safety outcomes would no longer be available. Similarly, combining a series of cells until they met the minimum threshold of >10 individuals would result in the merging of dissimilar patients into a single stratum, thereby generating a substantial loss of confounder information and significantly diminishing the efficacy of the analysis.

As described in Section I, it is not necessary to fully de-identify the PHI to comply with the HIPAA Privacy Rule. Rather, disclosure of aggregate information under Method Two, even with small cells, complies with HIPAA for two reasons. First, as long as the disclosure to the Operations Center or its subcontractors is for work contracted through the FDA, it is for "public health activities."

Second, if the disclosure of HIPAA identifiers is limited to dates related to individuals and geographic designations, it is a "Limited Data Set" that may be disclosed that is permitted as long as the recipient has signed a Data Use Agreement. I thus recommend that each data source sign a Data Use Agreement with the Operations Center and any subcontractors that will receive the Limited Data Set. This will ensure compliance with the HIPAA research rules described in Section I (C) above, and also will obviate the need for data sources to list such disclosures in their "accountings" to individuals as described in Section I(G) above. Data curation by the data sources to produce a Limited Data Set constitutes a "health care operation" permitted under the Privacy Rule.

Finally, the disclosure of aggregated information complies with the HIPAA minimum necessary standard. Even if small cells are included, they will be limited to the HIPAA identifiers that are included in a Limited Data Set; disclosure of a Limited Data Set complies with the minimum necessary standard.

C. Method Three: Distributed Regression Approach

Conclusion: This method complies with the HIPAA Privacy Rule. First, it appears that the information disclosed to the Operations Center will be "de-identified" under the HIPAA Privacy Rule. Moreover, even if the information is not fully de-identified, the disclosure still meets the requirements of the HIPAA Privacy Rule for two reasons. One, the Operations Center and its subcontractors are functioning as "public health authorities" on behalf of the FDA; use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization. Two, any information included in this method would fall within a "Limited Data Set" under the HIPAA Privacy Rule; use by or disclosure to the Operations Center or its subcontractors is permitted for research purposes if the disclosing covered entity has a Data Use Agreement in place with the recipient. Moreover, internal data "curation" is a health care operation permitted under HIPAA without individual authorization.

Discussion: Distributed regression encompasses a suite of methods that can be used to estimate regression models from distributed individual-level data. Unlike the other approaches, these estimates would be obtained without participating data sites providing any individual-level data to the Operations Center. Instead, each site would provide its summary statistics, which then can be combined at the Operations Center to produce regression estimates for analysis.

As you have explained, this approach is intended provide the analytic hub with estimates of regression parameters that are equivalent to estimates that would otherwise be obtained if the hub had full access to the data from the participating sites. This method places a greater statistical burden on participating sites, some of which may not have the necessary statistical or analytic staff to provide the hub with the required estimates. In addition, variances in generating summary statistics at the sites could reduce analytic flexibility that may be needed in later stage analyses at the hub.

You have explained that in most cases it would be nearly impossible to recreate individual patient-level information from site-provided summary statistics. If no identifiers for an individual are produced, the summary statistics would be de-identified under the standards discussed in Section A above.

Moreover, even if it is possible to recreate an individual identifier in rare cases, this method would comply with the HIPAA Privacy Rule. As noted with regard to Method One and Method Two, this would comply with HIPAA for two reasons: (1) disclosures of PHI for Mini-Sentinel purposes are permitted as "public health activities"; and (2) disclosure of Limited Data Set information are permitted as long as the recipient signs a HIPAA-compliant Data Use Agreement. Method Three would also comply with the minimum necessary standard.

D. Method Four: Score-Based Approaches

Conclusion: This method complies with the HIPAA Privacy Rule. First, it appears that most of the information disclosed to the Operations Center will be "de-identified" under the HIPAA Privacy Rule. Moreover, even if the information is not de-identified, the use or disclosure still meets the requirements of the HIPAA Privacy Rule for two reasons. One, the Operations Center and its subcontractors are functioning as "public health authorities" on behalf of the FDA; use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization. Two, any information included in this method would fall within a "Limited Data Set" under the HIPAA Privacy Rule; use by or disclosure to the Operations Center or its subcontractors is permitted for research purposes if the disclosing covered entity has a Data Use Agreement in place with the recipient. Moreover, internal data "curation" is a health care operation permitted under HIPAA without individual authorization.

Discussion: You explained the score-based approach in the following way:

> "The utilization of propensity scores (PS) and disease risk scores (DRS) in conducting pooled data analysis offers an optimal balance between the full utilization of covariate information and the maintenance of patient privacy that other methods cannot provide. Since a PS or DRS is a means of summarizing many covariates into a single opaque number it allows for the protection of individual patient-level information while retaining the content needed to adjust for confounding. Use of score-based approaches also provide enhanced statistical and analytic flexibility that is often limited in other approaches we have described.
>
> In the score-based approach, after establishing the design components of a desired study (sites and data sources, cohort criteria, exposure and outcome definition, covariates available by site, patient subgroups etc.) the participating sites segment patient-level information into three categories:

- **Shareable:** Non-confidential covariates including age in decades, sex, index dates, medical product exposure status, event and censoring dates
- **Private, Universal:** Covariates available at all sites, where the covariates are considered non-disclosable PHI under HIPAA standards
- **Private, Center-Specific:** Covariates that only certain sites can provide based on the granularity of their patient data or their access to EHRs and lab values. As above, the covariates would be considered non-disclosable PHI under HIPAA standards.

Example dataset for Center 6:

Center ID	Patient ID	Exposed?	Outcome Event?	Sex	Age 60-75?	Prop. Score.
6	1	No	No	F	Yes	0.52
6	2	Yes	No	F	No	0.68
6	3	Yes	Yes	F	No	0.74
6	4	No	No	M	Yes	0.23
…	…	…	…	…	…	…

With these covariates defined, each center may estimate and record several scores. The first score would be based on the shareable and private universal information (ScoreUniv). Optionally, sites could estimate a high dimensional propensity or disease risk score (hd-PS) [citation omitted].

If the centers vary with respect to the amount of data available – for example, if one center has EMRs but the others do not – a second score might be estimated.

This score, noted ScoreLocal, is based on the private center-specific information as well as the shareable and private universal information.

Once the participating sites have implemented their basic study design, separated their measured covariates into the three patient information categories and estimated their scores, they can create center-aggregated files (See table "x") [citation omitted] for transmission to the analytic hub. These site generated files would typically contain the study patient identifier, center identifier, exposure, outcome, shareable covariates, and estimated scores."

Under this method, my understanding is that the vast majority of information transmitted would be summaries of information that would not include individual-level data, and would thus be "de-identified" under the standards in Section I(A). If individual-level data would occasionally be available, the next step would be to evaluate whether that individual-level data contains any PHI. Descriptions of time that are specified as an offset from a date coded only as a year would <u>not</u> be PHI. A specific day, week, month, quarter or other time period within a particular year (e.g. the first quarter of 2012) is PHI. Of course, however, even where a specific time period within a particular year is included in the data, the information could be treated as de-identified if a qualified statistical expert determines that the risk is very small that the information could be used alone, or in combination with other available information, to identify the patient in that data.

The presence of PHI would trigger the need for HIPAA compliance. As with the other methods, however, this would comply with HIPAA for two reasons: (1) disclosures of PHI for Mini-Sentinel purposes are permitted as "public health activities"; (2) disclosure of Limited Data Set information (including dates related to individuals) are permitted as long as the recipient signs a HIPAA-compliant Data Use Agreement. Once again, if the information does not exceed a Limited Data Set, it complies with the minimum necessary standard, as well.

E. Method Five: Meta-Analysis

Conclusion: This method complies with the HIPAA Privacy Rule. The information disclosed to the Operations Center will be "de-identified" under the HIPAA Privacy Rule. Moreover, the internal use of the PHI by the covered entities to create the de-identified information for the Operations Center complies with the HIPAA Privacy Rule for two reasons: (1) if the data sources are subcontractors to the Operations Center, they are functioning as "public health authorities" on behalf of the FDA, and use of PHI by the subcontractors is permitted as a public health activity without individual authorization; (2) the internal data "curation" is a health care operation permitted under HIPAA without individual authorization. If the data sources are themselves performing "research," the results of which are disclosed to the Operations Center, the data sources would be required to comply with one of the HIPAA research rules described in Section I(C).

Discussion: Under this method, the Operations Center will define a protocol that is distributed to the individual sites. Each site will conduct an analysis of its own data by creating the cohort, defining covariates, measuring exposure, and identifying outcome events. Each site will carry out its own statistical analysis to determine the treatment effects of interest and confidence intervals. The sites will transmit their findings to the Operations Center (point estimate, variance, and common diagnostic data). The sites will not include any individual-level data to the Operations Center. Upon receipt of each site's statistical results, the Operations Center will apply the received point estimates and variances to compute a singular, study-wide result for point estimate and confidence level. The Operations Center also has the option to perform standard heterogeneity tests to ascertain whether all or only some of the sites qualify to be included in the summary estimates. You have expressed the concern that, because each site conducts its own separate study under this meta-analysis study, the results may be ambiguous or undetectably flawed.

Because no individually identifiable health information will be disclosed to the Operations Center, that disclosure does not trigger the application of the HIPAA Privacy Rule.

Moreover, the internal use of the PHI by the covered entities to create the de-identified information for the Operations Center complies with the HIPAA Privacy Rule for two reasons: (1) if the data sources are subcontractors to the Operations Center, they are functioning as "public health authorities" on behalf of the FDA, and use of PHI by the subcontractors is permitted as a public health activity without individual authorization; (2) the internal data "curation" is a health care operation permitted under HIPAA without individual authorization. If the data sources are themselves performing "research," the results of which are disclosed to the Operations Center, the data sources would be required to comply with one of the HIPAA research rules described in Section C.

III. Conclusion

As explained above, each of the proposed statistical methods will comply with the HIPAA Privacy Rule. Under some of these methods, the information disclosed to the Operations Center will be "de-identified" under the HIPAA Privacy Rule. Moreover, even if the information is not de-identified under some of the statistical methods, the use or disclosure to the Operations Center or its subcontractors still meets the requirements of the HIPAA Privacy Rule for two reasons. One, the Operations Center and its subcontractors are functioning as "public health authorities" on behalf of the FDA; use by or disclosure of PHI to the Operations Center or its subcontractors therefore is permitted as a public health activity without individual authorization. Two, information included in each method would fall within a "Limited Data Set" under the HIPAA Privacy Rule; use by or disclosure to the Operations Center or its subcontractors is permitted for research purposes if the disclosing covered entity has a Data Use Agreement in place with the recipient. Moreover, internal data "curation" is a health care operation permitted under HIPAA without individual authorization.

If you have any questions, please do not hesitate to call.

KBR