MINI-SENTINEL PRISM

IDENTIFYING COMPLEMENTARY DATA SOURCES REPORT

Prepared by: Corrie Paeglow, DrPH, MPH, on behalf of the Mini-Sentinel Activity 6 workgroup

Author Affiliations: HealthCore, Inc, Alexandria, VA

January 17, 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.

Mini-Sentinel PRISM

Identifying Complementary Data Sources Report

I. BACKGROUND

The Mini-Sentinel project is an FDA sponsored initiative designed to develop an active surveillance system for healthcare safety issues, including vaccine and drug safety. Within Mini- Sentinel, the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) program is currently establishing mechanisms for conducting vaccine safety research using healthcare claims data. The data for these projects are provided by Mini-Sentinel Data Partners and consist of variables developed from healthcare claims data.

Administrative healthcare claims data are a valuable resource for conducting vaccine safety research. However, healthcare claims do not typically capture all data that may be necessary for this type of research, including potentially important confounders, such as race/ethnicity and socioeconomic status. In addition, when using claims for safety surveillance research, it is usually desirable to validate the outcome of interest. Often, the data necessary for such validation activities are not available in healthcare claims and must be sought through medical records or another source.

One of the goals of the PRISM project is to identify and evaluate complementary data sources for the healthcare claims data provided by the Data Partners. The goal of PRISM Activity 6, the subject of this report, is to identify such data sources and investigate the feasibility of and process for matching these data sources to administrative healthcare claims data. The project was undertaken in two phases:

- Enumeration of available data sources with the theoretic possibility of linking to claims and ultimately improve the data capture of PRISM activities;

- Selection of data sources of highest priority based on their potential contribution to PRISM, and detailed information gathering about these data sources

The first goal of this project was to develop a comprehensive list of potential complementary data sources. This was achieved through three tasks: (a) development of a preliminary relevant list of data sources, (b) engagement of the workgroup for expansion of the list, and (c) engagement of Mini-Sentinel and other reviewers for feedback and incorporation of their comments into a comprehensive list.

II. METHODS

A. PRELIMINARY LIST

The workgroup required that a data source capture vaccine exposure, potential confounding/effect modifying characteristics or outcomes for lives represented by the Mini- Sentinel Data Partners to ensure that all the included data sources are relevant. Using only these criteria the project team began generating a list of complementary data sources, leveraging the results of a previous report developed for the FDA "Evaluation of Potential Data Sources for the FDA Sentinel Initiative – Final Report" by Booz Allen Hamilton. This report, which included lists of categories of data sources and specific data sources, served as a starting point for the workgroup's search.

The FDA also shared a list of priority outcomes for vaccine safety research, generated by vaccine safety experts at the FDA. The workgroup specifically searched for data sources that collected data related to these priority outcomes. In addition, members of the workgroup from HealthCore brainstormed to compile a list of potential data sources and categories of data sources.

A preliminary list of complementary data sources was researched via literature reviews and internet searches. Data source information on the following criteria was collected:

- Population captured
- Capture of vaccines, covariates of interest, or endpoints of interest
- Value provided by using this data source
- Active/historical status
- Website address

This information was essential to determine whether a particular data source could enhance claims data for vaccine safety research. All information was entered in a spreadsheet that served as the working document for the project.

The preliminary list identified 106 data sources in ten categories.

B. WORKGROUP ENGAGEMENT

The workgroup consisted of members from four Mini-Sentinel Data Partners (Aetna, HealthCore, Humana and Kaiser Permanente), the Harvard Mini-Sentinel Operations Center and the FDA, who met eight times between February and July 2011. During these meetings, the workgroup reviewed a spreadsheet containing the preliminary list of complementary data sources and expanded the number of data sources and categories in which these data sources fell. At the end of this process, a total of 129 data sources were identified. These data sources fell into sixteen broad categories:

1. Consumer Data
2. Demographic Data
3. Disease Registry
4. Non-vaccine Exposure Registry
5. Health Information Exchange
6. Hospital Data
7. Laboratory Data
8. Member Self-reported Data
9. Pharmacy Benefits Management Data
10. Pharmacy E-prescribing Network
11. Practice-based Research Network
12. Retail Clinic
13. Retail Pharmacy
14. Specialty Pharmacy
15. Vaccine Exposure Registry
16. Vital Statistics Data

C. MINI-SENTINEL & EXTERNAL REVIEWERS ENGAGEMENT

The preliminary list of complementary data sources spreadsheet, with 16 categories, was circulated to Mini-Sentinel Core Leads, Mini-Sentinel members and non-Mini-Sentinel reviewers. Reviewers were invited to add additional data sources, provide comments on data sources already on the list, and provide information on any experience they had linking with any of the data sources.

In total, there were 248 complementary data sources identified. **Table 1** shows the number of data sources that fell into each category of data sources, along with how many were assigned a priority category for PRISM, as discussed in further detail in the "Prioritization of Data Sources" section. In addition, the attached spreadsheet contains the full list of complementary data sources and the information collected about the sources. Please note the information was, in many cases, taken from the data source's website and has not been independently verified.

Table 1. Number of Data Sources in each Category

	Number of Data Sources in Category	Priority for PRISM				
		High	Medium-High	Medium	Medium-Low	Low
Consumer Data	1					1
Demographic Data	1			1		
Disease Registry	59	20		25		14
Non-vaccine Exposure Registry	26					26
Health Information Exchange	37		37			
Hospital Data	8					8
Laboratory Data	2	2				
Member Self-reported Data	4				4	
Pharmacy Benefits Management Data	9				9	
Pharmacy E-prescribing Network	1				1	
Practice-based Research Network	50	50				
Retail Clinic	4	4				
Retail Pharmacy	14					14
Specialty Pharmacy	3					3
Vaccine Exposure Registry	25	25				
Vital Statistics Data	4	3				1
Column Total	**248**	**104**	**37**	**26**	**14**	**67**

III. PRIORITIZATION OF DATA SOURCES

After enumerating the data sources, the workgroup turned its attention to prioritizing the data sources. The workgroup developed specific qualitative criteria through a process of discussion described. This process began with a discussion of each group of data sources, and individual data sources within a particular group. Through this discussion each source was assigned two priorities: a priority for PRISM, based on the data source's potential value to vaccine safety research, and a priority for further investigation under this activity. We differentiated between these priorities because some data sources, such as vaccine registries, are highly valuable for vaccine safety research but are already being used as part of PRISM, and do not need to be investigated further. As the workgroup reviewed the data sources, a set of criteria for prioritizing the data sources emerged from the discussion. The final step of the prioritization process was formalizing the criteria that emerged from the conversation into a framework for evaluating data sources.

In parallel with these discussions, a literature review was undertaken to identify examples of a particular data source being matched to healthcare claims data or used in vaccine safety research. The results of this literature review are presented in a separate spreadsheet submitted with this report.

A. SUMMARY OF WORKGROUP DISCUSSIONS

The primary factor considered in the decision about a given data source's priority for PRISM was whether the data source enhanced claims data the Data Partners have or merely duplicated it. For example, pharmacy benefits management (PBM) programs have rich data about drug exposures that might be relevant covariates in vaccine safety studies. Most of the data, however, are already captured in claims data and matching to PBM data is unlikely to significantly enhance claims data. Because of this, PBM data were given a ranking of low priority for PRISM. This same logic accounts for the low priority ranking of several other data sources, including specialty pharmacy data and hospital data sources such as the Premiere Perspectives Database and Pediatric Health Information System (PHIS).

Another important consideration was the richness of the data available in the data source. Some data sources, such as vaccine registries, only capture data about one exposure, outcome or covariate. Electronic Medical Records (EMR) from a Practice-based Research Network, in contrast, are likely to have rich data on vaccine exposures, outcomes of interest and potential covariates. All the information is stored in one location, making it an ideal data source for capturing a variety of information. Health Information Exchanges also provide access to EMR data but were not designed to support research, which suggests that the data may be harder to access than data collected by Practice-based Research Networks. Accordingly, the workgroup assigned practice-based research networks a high priority for PRISM and Health Information Exchanges a medium-high priority.

We also considered whether the data source captures data that is not available from another data source. Retail clinics, for example, are a potential source of information about very short- term adverse events, such as fainting and anaphylaxis; the FDA has noted that many reports to the Vaccine Adverse Events Reporting System (VAERS) come from retail clinics. These short-term, transient events are not likely to be captured in claims data or any other data source we enumerated but represent important vaccine safety concerns. Thus, this data source was given a high priority. Other data that are unlikely to be captured in another source include laboratory data and vital statistics data captured in the National Death Index and NCHS Fetal Death Data.

In order to effectively match to claims data, the complementary data source must include personal identifiers that allow direct linkage to an individual included in the claims data. Any data source that does not include such identifiers was given a low priority for PRISM. This included several large hospital discharge databases maintained by state departments of health. In the same vein, a low priority was assigned to US census data. Census data can be used to impute characteristics such as race/ethnicity, socio-economic status; however as this is imputed and not directly linked data it was assigned a low priority.

Another important consideration is the extent of the capture of the data source. Member self- reported data, such as data collected by health plans through health screening and disease management programs, is not collected for the majority of members enrolled in the health plan, and therefore given a low priority. Population-based data sources will capture all the outcomes or exposures in a particular population, thus they will capture all health plan members who are also in the catchment population of the data source. These sources do not raise issues of bias when matched to claims data and used to ascertain exposures or outcomes. Population-based data sources that were identified included state cancer registries and birth defects registries.

Just having extensive capture, however, did not guarantee a data source a high priority ranking; it must also capture data that are of particular interest for PRISM. Birth defects have been identified by the FDA as a priority outcome for vaccine safety research, thus population-based birth defects registries were given a high priority for PRISM. SEER cancer registries are also population-based, however cancer is not a priority outcome of interest for vaccine studies and these sources were assigned a medium priority. Some data sources collected data that may be more useful for drug safety studies than vaccine safety studies, and these data sources were given a low priority, although they may be of more interest for other biologic or drug safety studies.

248 data sources were assigned a priority for PRISM. The prioritization distribution was as follows: 104 high priority, 37 medium-high priority, 26 medium priority, 14 medium-low priority and 67 low priority. Table 1, above, shows the number of data sources in each category and the prioritization assignment distribution.

Priority for Further Investigation. The workgroup prioritized the 141 data sources assigned a high or medium-high priority for PRISM for further investigation. Data sources with a high priority for PRISM were assigned a high priority for further investigation unless Mini-Sentinel had previous experience matching with it, as is the case with vaccine registries. Health Information Exchanges were the only data sources assigned a medium-high PRISM priority, and considered high priority for further investigation as they capture data on outcomes, exposures and potential covariates of interest. Data sources that were medium priority for PRISM – population-based disease registries without priority outcomes of interest – were low priority for further investigation. Finally, data sources that were assigned a medium-low PRISM priority were assigned a low priority for further investigation. At the end of this process 112 data sources in 5 categories were determined to be high priority for further investigation and 29 data sources were determined to be low priority for further investigation. **Table 2** (below) presents the distribution of priority for further investigation.

Table 2. Priority for Further Investigation of Data Sources that are High Priority for PRISM

	Priority for Further Investigation		
	High	**Medium**	**Low**
Disease Registry	20		
Health Information Exchange	37		
Laboratory Data			2
Practice-based Research Network	50		
Retail Clinic	4		
Specialty Pharmacy			
Vaccine Exposure Registry			25
Vital Statistics Data	1		2
Column Total	**112**	**0**	**29**

B. FRAMEWORK FOR EVALUATING DATA SOURCES

New data sources are always being developed or coming to the attention of researchers. To aid in future evaluation of the suitability of new data sources for enhancing claims data for vaccine safety research the workgroup established a series of questions that can be used for this purpose.

- Does the data source enhance, or merely duplicate, the existing claims data?
  Sources that simply duplicate claims data should be assigned lower priority.

- How complete is the capture?
  Sources that have incomplete capture should be assigned lower priority.

- Does the data source capture an exposure or outcome that is of explicit interest to PRISM?
  Sources that did not capture exposures/outcomes of explicit interest should be assigned lower priority.

- Is the data linkable to an individual life in the claims data?
  Data that is not directly linkable to an individual should be assigned lower priority.

- Does the data source provide information on just one aspect of interest, or several?
  Data sources with multiple outcomes of interest should be assigned a higher priority.

- Are the data available elsewhere?
  Data that are not available elsewhere should be assigned higher priority.

- How feasible will it be to access the data?
  Although not a primary concern, data that can be accessed easily should receive slightly higher

priority than those that are more difficult to access. It may be hard to determine how feasible it is to access the data until the data source has been contacted.

IV. GENERAL STRATEGY FOR FURTHER INVESTIGATING DATA SOURCES

The final portion of Phase 1 of the project involved developing a strategy for investigating the selected data sources. One hundred and twelve data sources that fell into 5 categories were determined to be high priority for further investigation. We decided to investigate only 4 categories, excluding lupus registries as these registries though population-based are quite small. Also, while lupus is a priority outcome of interest for vaccine safety research, it is not currently an outcome of interest for any PRISM projects. Because these data sources fell into a small number of categories and we anticipate that data sources within each category are likely to have a large number of characteristics in common, we developed a strategy for investigating each of these categories rather than individual data sources.

We began our investigation of these data sources by sending out a questionnaire that collected general data about the data source. The questionnaire consisted of two sets of questions, a general set that was sent to all data sources, and a set of questions specific to that category of data source. The general set of questions asked specifics about personal identifiers included in the data, such as, if the data was stored electronically and whether the data source had ever been matched to administrative healthcare claims data. The specific set of questions asked about the coding systems used for data elements we anticipated to capture in the data. For example, the birth defects registry questionnaire asked how different types of birth defects are coded within the registry.

The questionnaire also asked for contact information for programmatic, technical, and legal contacts for the data source. When we received the completed questionnaire, this information was used to schedule a conference call between HealthCore and these individuals. In the call more detailed questions were asked that assessed the feasibility of and process for matching the data source's data to administrative healthcare claims data.

The questionnaire was widely distributed to maximize the number of responses. We utilized the internet and the workgroup's pre-existing network to retrieve contact information on people receiving the questionnaire. The questionnaire was distributed with a cover letter from the FDA explaining the project and encouraging potential respondents to participate. Given the wide reach of the questionnaire, we developed a strategy for prioritizing how many, and which, data sources to schedule follow-up calls. This strategy is outlined below by category of data source.

Birth Defects Registries. State birth defects registries collect data on every child diagnosed with a birth defect in the state. Three registries from different geographic sizes (large state, small state, and a metropolitan area) were contacted.

Health Information Exchanges (HIE). HIE are local data exchanges between service providers. Two HIEs were investigated, one that is large and well-established and another that is smaller and new-to-the-market.

Practice-Based Research Networks. Practice-Based Research Networks are data sources created by researchers that combine data from different sources. One research data network, the Practice Partner

Research Network, was investigated. The Practice Partner Research Network is large, well-established, and there is interest in matching to it for another PRISM project.

Retail Clinics. Retail clinics are healthcare providers located in non-healthcare settings. One retail clinic chain was contacted.

V. RESULTS

A. BIRTH CERTIFICATE DATA

1. New York State Birth Certificate Data

Personally Identifying Variables Available. All major personal identifiers, including name, date of birth, gender, address and social security number are available. These variables, with the exception of date of birth, are also available for the mother. However, no health insurance information is reported for the mother or the child.

Clinical Variables Available. (Note: additional detail is provided only for variables of special interest for PRISM-related work.) This state uses the revised birth certificate from developed by CDC NCHS in 2003. This form collects information on a variety of clinical variables, including:

1. Pregnancy-related complications
2. Congenital birth defects
    a. Anencephaly
    b. Meningomyelocele/Spina bifida
    c. Cyanotic congenital heart disease
    d. Congenital diaphragmatic hernia
    e. Omphalocele
    f. Gastroschisis
    g. Limb reduction defect (excluding congenital amputation and dwarfing syndromes)
    h. Cleft Lip with or without Cleft Palate
    i. Cleft Palate alone
    j. Down Syndrome
    k. Hypospadias
3. Characteristics of labor and delivery
4. Risk factors during pregnancy
5. Details of prenatal care
6. Birth Weight
7. Gestational age
    a. Estimated by obstetrician
    b. Date of LMP, which can be used to estimate gestational age
8. Parental demographics

In previous versions of the birth certificate much of the clinical information reported came from the mother's self-report. In the 2003 revision this information is collected from the mother's medical record.

Percent Completeness. The data will be complete for all births that take place within the state. For those births that take place in a health care facility, the majority of births, the data should be complete. It is not clear, however, how complete the information is for children who are not born in a health care facility or taken to a health care facility shortly after birth.

Method of Data Collection. All births within the state are required to be registered, and typically the facility where the child is born files the report with the Office of Vital Statistics.

Previous Linkage to Payer Data. To our knowledge, the data have not been linked to payer data. This is likely due to the legal restrictions on the release of the data described below.

Legal & Regulatory Requirements. This state has specific legislation that prohibits the release of birth certificate data. This is true even for FDA sponsored research for public health activities, thus we would not be able to use this data for any PRISM studies.

Technical Requirements. As we are not able to link to this data the technical requirements for doing so were not discussed.

Budget Requirements. As we are not able to link to this data the budget requirements for doing so were not discussed.

Overlap with Plan Membership. All children born in the state should be included, with the exception of one large metropolitan area that has its own vital statistics system. Thus, the degree of overlap with plans that have members in that state, but not the metropolitan area that have had a child should be close to 100%. Although it is possible for a child to be born out of state most states exchange birth certificate data with their neighbors when the mother resides in a different state, although there will be more of a lag in this data.

Completeness of Variables of Interest for PRISM. The data should be complete for all children born in a health care facility.

2. Commonwealth of Virginia Birth Certificate Data

Virginia does not currently use the revised 2003 birth certificate form. Thus, information on maternal risk factors and other non-delivery related maternal medical information are gathered through self-report rather than medical record review.

Personally Identifying Variables Available. All major personal identifiers, including name, date of birth, gender, address and social security number are available. These variables, with the exception of date of birth, are also available for the mother. No health insurance information is reported for the mother or the child.

Clinical Variables Available. (Note: additional detail is provided only for variables of special interest for PRISM-related work.)

1. Pregnancy-related complications
2. Congenital birth defects
    a. Anencephaly

   b. Hydrocephalus
   c. Microcephalus
   d. Other nervous system abnormalities
   e. Heart malformations
   f. Other circulatory system abnormalities
   g. Rectal atresia/stenosis
   h. Tracheoespohageal fistula
   i. Omphalocele
   j. Other gastrointestinal abnormalities
   k. Malformed genitals
   l. Renal agenesis
   m. Other urogenital abnormalities
   n. Cleft lip/palate
   o. Polydactyly/syndatyly/adactyly
   p. Club foot
   q. Diaphragmatic hernia
   r. Other musculoskeletal abnormalities
   s. Down's syndrome
   t. Other chromosomal abnormalities
   u. Other
3. Characteristics of labor and delivery
4. Risk factors during pregnancy
5. Details of prenatal care
6. Birth Weight
7. Gestational age
   a. Estimated by obstetrician
   b. Date of LMP, which can be used to estimate gestational age
8. Parental demographics

Percent Completeness. The data should be complete for all births that take place within the state. For those births that take place in a health care facility, the majority of births, the data should be complete. It is not clear, however, how complete the information is for children who are not born in a health care facility or taken to a health care facility shortly after birth.

Method of Data Collection. All births within the state are required to be registered, and typically the facility where the child is born files the report with the Office of Vital Statistics.

Previous Linkage to Payer Data. This data has been linked to Medicaid data, but we are unaware of any instances of it being matched to private payer data. An example of a link with Medicaid data is found in:

> Anum EA, Retchin SM, Garland SL, Straus JF III. Medicaid and preterm births in Virginia: an analysis of recent outcomes. *Journal of Women's Health*. 2010; 19, 11, 1969, 1975.

Legal & Regulatory Requirements. Matching to Virginia vital records data requires approval from the Virginia Department of Health Institutional Review Group and the State Registrar of Vital Records.

Technical Requirements. Files can be transmitted in a variety of formats, including Microsoft formats, ASCII, and SPSS.

Budget Requirements. The Office of Vital Statistics charges a flat fee of $60/hour for time spent on the linkage.

Overlap with Plan Membership. All children born in the state should be included, so the degree of overlap with plans that have members in that state who have had a child should be close to 100%. Although it is possible for a child to be born out of state most states exchange birth certificate data with their neighbors when the mother resides in a different state, although there will be more of a lag in this data.

Completeness of Variables of Interest for PRISM. The data should be complete for all children born in a health care facility.

3. New York City Birth Certificate Data

Personally Identifying Variables Available. All major personal identifiers, including name, date of birth, gender, address and social security number are available. These variables, with the exception of date of birth, are also available for the mother. However, no health insurance information is reported for the mother or the child.

Clinical Variable Available. (Note: additional detail is provided only for variables of special interest for PRISM-related work.)

1. Pregnancy-related complications
2. Congenital birth defects
    a. Anencephaly
    b. Meningomyelocele/Spina Bifida
    c. Cyanotic Congenital Heart Disease
    d. Congenital Diaphragmatic Hernia
    e. Omphalocele
    f. Gastroschisis
    g. Limb Reduction Defect
    h. Cleft Lip With or Without Cleft Palate
    i. Cleft Palate Alone
    j. Down Syndrome Karyotype
    k. Other Chromosomal Disorder
    l. Hypospadias
3. Method of diagnosis of congenital defect
4. Pregnancy-related complications
5. Congenital birth defects
6. Characteristics of labor and delivery
7. Risk factors during pregnancy
8. Details of prenatal care
9. Birth Weight
10. Gestational age

a. Estimated by obstetrician
b. Date of LMP, which can be used to estimate gestational age

11. Parental demographics

Percent Completeness. The data will be complete for all births that take place within the given state. For those births that take place in a health care facility, the majority of births, the data should be complete. It is not clear, however, how complete the information is for children who are not born in a health care facility or taken to a health care facility shortly after birth.

Method of Data Collection. All births within the state are required to be registered, and typically the facility where the child is born files the report with the Office of Vital Statistics.

Previous Linkage to Payer Data. The Office of Vital Statistics reported that they did not have any experience matching this data to payer data, and we found no examples of such a linkage in the literature.

Legal & Regulatory Requirements. A formal Data Use Agreement that set out the conditions for access, disclosure, retention, storage and disposition of the data would have to be executed. Once the parameters of the data match are specified, the Office of Vital Statistics would consult with the DOHMH Privacy Officer to determine if there are any unforeseen legal or regulatory barriers to provision of the requested data.

Technical Requirements. The list of subjects to be matched and results of the match must be transferred to/from OVS by a method approved by the Department of Mental Health and Hygiene Division of Informatics and Information Technology. Currently these methods include using encrypted e-mail (e.g., MacAfee End Point for data files less than 10 Mbytes) or an encrypted web-based delivery service (e.g., BISCOM for data files equal to or greater than 10 Mbytes).

Budget Requirements. The cost is dependent on the size of the request; for a small project their might not be any charge. There is a charge of $15 if a copy of a birth certificate is required; however this is unlikely to be necessary for PRISM research.

Overlap with Plan Membership. All children born in the metropolitan area should be included, so the degree of overlap with plans that have members in the metropolitan area who have had a child should be close to 100%.

Completeness of Variables of Interest for PRISM. The data should be complete for all children born in a health care facility.

4. Summary/Other Important Details

States vary in what data is collected on the birth certificate. However, most states use forms developed and recommended by The National Center for Health Statistics, which underwent revision in 2003. The revised form collects more precise data about the prenatal care initiation date and collects in-depth information about pregnancy- and labor-related complications. The new form requires the abstraction of data from the mother's medical record, while the old form relied exclusively on mother's self-report. While this form has been widely implemented nationwide, some states are still using the older birth

certificate form. This state, however, is currently transitioning to the new form and anticipates having the transition complete in several years.

While we investigated birth defects registry data as an entirely separate date source birth certificates also contain information about a limited number of birth defects as listed above. Birth defects registry data, as described below, is more difficult to match with than birth certificate data, thus it may be easier to use birth certificate data to investigate birth defects than it is to use registry data.

States vary with regard to their legislation pertaining to the release of birth certificate data. At least one state we contacted has laws that specifically preclude the release of birth certificate data. Thus, before pursing this option it will be necessary to confirm with each of the states of interest whether the laws of that state will allow birth certificate data to be disclosed. Such a review, however, is beyond the scope of this report.

Several barriers to sharing this data were mentioned by the respondents. Concerns were raised about disclosing information to groups that are affiliated with insurance carriers. Thus, before beginning any project that involves matching to these data sources it will likely be necessary to provide assurances about the firewalls between Mini-Sentinel Data Partners and their parent companies. The other major barrier that was mentioned included the availability of staff and resources to conduct the match. Offices of Vital Statistics are state or city run groups and as such often have limited staff. This staff may also be responsible for developing periodic reports using birth certificate data, and during the time when the report is under development the staff may not have time to work on additional projects.

B. BIRTH DEFECT REGISTRIES

1. Michigan Birth Defects Registry

Personally Identifying Variables Available. All major personal identifiers, including name, date of birth, gender, address and social security number are available. These variables, with the exception of date of birth, are also available for the mother. However, no health insurance information is reported for the mother or the child unless the child has Medicaid coverage, in which case his/her Medicaid ID number would be available. This information, however, is likely to be of little utility for PRISM projects.

Clinical Variables Available. The birth defects reporting form collects a limited amount of clinical detail. Birth defects are captured with ICD-9 diagnosis codes. There may also be some information about prenatal exposures captured using ICD-9 diagnosis codes.

Percent Completeness. Information is collected on over 400 birth defects, but while reporting is mandated the system is passive and relies on hospitals and clinics to provide reports. Thus, the data is most likely not entirely complete although the degree of completeness is likely to be high given the legal mandate to report. In addition, if the reporting forms are completed incorrectly they may have missing information. While the degree of missingness has not been ascertained, it is quite a bit higher than what is seen in birth certificate data.

Method of Data Collection. Data are collected through reports from hospitals and other health care facilities. There is a standardized reporting form that is available on the registry's web-site that includes instructions to ensure a high quality of reported data. There is also a manual that is available on the

web-site that gives detailed instructions for how to complete the forms accurately. The form includes both a written name of the defect and the ICD-9 code associated with the defect.

Previous Linkage to Payer Data. The registry reported that its data is routinely linked to Medicaid data, but we were unable to find any instance of this data being linked to private payer data.

Legal & Regulatory Requirements. Executing this kind of match would require several layers of regulatory approval. The Data Partners would need to sign a certificate of confidentiality and IRB approval would be needed. As the registry has never previously linked with private insurance data it might take time to first determine what the legal and regulatory requirements would be and then meet them.

Technical Requirements. No specific technical requirements were noted, nor were any technical barriers noted by the registry.

Budget Requirements. The general charge for record linkage work for research purposes is $850 for the link and $.75 per confirmed link.

Overlap with Plan Membership. All children diagnosed with a birth defect in the state should be included, so the degree of overlap for children diagnosed with a birth defect in the state should be 100%

Completeness of Variables of Interest for PRISM. A large number of birth defects are captured in this registry, so it could likely be used for a variety of outcomes of interest. However, beyond information about birth defects little other data is included in the registry. Data about prenatal exposures is collected using ICD-9 diagnosis codes, which suggests that maternal exposure to vaccines is not capture.

2. Arizona Birth Defects Monitoring System

Personally Identifying Variables Available. For vaccine safety research we would most likely want to match a mother who had received a vaccine during pregnancy to her child who may have developed a birth defect. This particular registry does not include any maternal identifying information which would make this difficult. It does include the child's birth certificate number which can be used to link back to the child's birth certificate and identify his/her mother. This would only work, however, if the child is born in the state.

Clinical Variable Available. Only data about birth defects diagnosed in the first year of life to a woman who resides in Arizona at the time of birth are collected. These defects are coded using a six digit classification developed by the CDC. No data about prenatal exposures are collected.

Percent Completeness. The completeness of the data will vary according to the means of ascertainment. For conditions for which there is active surveillance the data is likely to be very complete. Many conditions, however, are only subject to passive surveillance and in these cases the data are likely to be much less complete.

Method of Data Collection. Data for nine birth defects is captured through active surveillance, where a member of the registry staff goes to hospitals where children are born to seek out information about the condition the child has and collect data through chart abstraction. Other birth defects are tracked

through passive surveillance, with doctors and other health care providers providing reports. These reports are not mandated.

Previous Linkage to Payer Data. To our knowledge there have been no previous linkages of this data to payer data, nor did the respondent mention any such linkage.

Legal & Regulatory Requirements. Appropriate IRB approvals would need to be in place, as well as approval from the Office of Vital Statistics.

Technical Requirements. No specific technical requirements were noted.

Budget Requirements. While there would be some costs incurred for the linkage the registry preferred to not give an exact figure, instead stating that it would depend on the size and scope of the project at hand. The registry has undergone serious budget cutbacks recently, so for a successful linkage to take place Mini-Sentinel would need to be willing to cover all costs associated with that linkage.

3. Summary/Other Important Details

As both registries collect passive reports they do not have control over the quality of data they receive, and some fields may be missing. The incomplete information may include fields such as mother's first or last name, information that will be necessary for matching pregnant women who appear in the claims data with their children.

Given the lack of availability of data about the mother's first name in birth defects registry data both states recommended taking a two step matching approach, first matching the mother to her child using the birth certificate data and then matching the child to any reports s/he many have in the birth defects registry system. While this process would be somewhat slower than matching to only one data source, it may be expedited by the fact that both birth certificate data and birth defects data are housed in the Department of Vital Statistics. Additionally, the revised birth certificate form includes information about some birth defects. Thus, depending on what birth defect is of interest to a particular study it may not be necessary to match to the birth defects registry data; the necessary data may be available in the birth certificate data alone.

C. HEALTH INFORMATION EXCHANGES

1. HIE Number 1

Personally Identifying Variables. All major personal identifiers, including name, gender, address and social security number are available. In addition, health insurance information is available which will significantly enhance the ability to match to payer data.

Clinical Variables. Diagnosis, procedure, pharmacy, laboratory and vaccination data are captured in the source. Diagnosis and procedure codes are typically captured using ICD-9 diagnosis codes and CPT codes. There is wider variation in how medications, laboratory results and vaccinations are reported in the data.

Percent Completeness. Individual patients are not participants in the HIE, but rather providers and institutions are participants. Thus, a patient that sees multiple providers may not have complete data in

the HIE. There is no way to determine if this is the case, making it impossible to determine the percent completeness of this data for any given individual.

Method of Data Collection. Data is provided by participating providers and institutions through EMR systems that are at the provider's offices or providing institutions. These data collection systems allow information to be collected in both pre-specified drop down boxes and free text fields. When data are entered in drop-down boxes they automatically populate in HIE databases. In general, hospitals use the dropdown fields more than individual provider offices do.

Previous Linkage to Payer Data. This data has not been linked to payer claims previously. However, it has been regularly used in public health research activities, including regular reporting to the State Department of Health and CDC Public Health Information Messaging System (PHIN-MS) for surveillance activities.

Legal & Regulatory Requirements. Data can only be released from the HIE if the provider who sent the data into the HIE agrees to have it released. Thus, the completeness of the data will be impacted by the willingness of providers to release the data. However, the system was designed to facilitate separating data from providers and institutions that consent to have their information shared and those who do not.

Technical Requirements. Data can be transferred in any format the recipient would like. Initial data transfer will require some data infrastructure build-out, but after this is created it can be used for future data transfers. The organization has a number of contractors they regularly work with for such build-out which makes the process relatively efficient.

Budget Requirements. As the data source has never been matched to payer data they were unsure what the exact costs would be. However, the first data transfer would cost more than any subsequent transfers, as some infrastructure build-out costs would be incurred.

Overlap with Plan Membership. It is unknown to what extent the patients represented in the HIE overlap with plan membership.

Completeness of Variables of Interest for PRISM. In theory data on vaccinations, diagnoses, procedures and medications is available. Additional clinical information includes: allergies, radiological studies, notes from provider encounters and a list of providers seen is available for each patient. Demographic variables that are available include gender, date of birth, race and marital status. However, as noted above, this is limited by the degree to which a patient's providers participate in the HIE.

2. HIE Number 2

The second HIE that responded to our request for information indicated that they were not willing to participate in an FDA sponsored vaccine safety study. Because of this lack of interest we did not pursue a follow-up phone call with the group, and this information is solely from the information gathering tool they completed.

Personally Identifying Variables Available. All major personal identifiers, including name, gender, address and social security number are available. However, health insurance information is not available.

Clinical Variables Available. Data on diagnoses, procedures, medication use, laboratory results and vaccination are included in the HIE. Each of these are stored as original coded by the facility. While it is likely that facilities code using systems that are generally well-recognized and used we were not able to confirm that this was the case.

Percent Completeness. Individual patients are not participants in the HIE, but rather providers and institutions are participants. This particular HIE only works with hospital systems, so data would be available primarily for inpatient stays but may also be available from providers who are in the hospital system.

Method of Data Collection. Data is captured through EMR systems that are already being used by the providers.

Previous Linkage to Payer Data. No such linkage has been conducted.

Legal & Regulatory Requirements. The state in which this HIE is located has extensive legislation and regulation surrounding the disclosure of PHI. The HIE did not elaborate on what this legislation included, but this may be because they have never linked to an external data source and thus have not had to fully explore this issue.

Technical Requirements. No specific technical requirements were noted, nor were any technical barriers noted by the registry.

Budget Requirements. There would be a flat fee for conducting the match and a flat fee for each record that is matched.

Overlap with Plan Membership. It is unknown to what extent the patients represented in the HIE overlap with plan membership.

Completeness of Variables of Interest for PRISM. This registry captures data primarily from hospitals. As most vaccines are administered in the outpatient setting they may not be well-captured in this system. In addition, most codes from the primary care setting will not be captured. Given this limitation it's possible that only the most severe adverse reactions – those that require hospitalization – could be identified using this data.

3. Summary/Other Important Details

Representatives from two other HIEs responded to our request for information. One indicated that their files are stored in text format and because of this are difficult to use for research purposes and did not provide any further information. The other HIE requested a phone call prior to completing the information gathering tool and during that call indicated that they would complete the tool shortly thereafter. The completed questionnaire was not received, and when a follow-up e-mail was sent they replied that they would not be able to complete the questionnaire.

One theme that emerged in several of our conversations was that HIEs are currently in cycle of rapid development and improvement, and thus may be altered significantly by the time Mini- Sentinel is interested in accessing data contained in HIEs. One often-cited example was the continuity of care document. This document, a file containing a summary of a patient's medical history, is designed to be

highly portable and provide any provider the basic information he/she would need to care for a patient. This includes, but is not limited to, age, gender, allergies, immunizations and lab results. Currently not all HIEs have the capacity to create these documents; however, the US Healthcare Technology Standards Panel has adopted the continuity of care document (CCD) as a standard, making it likely that their use will expand in the near future. Continuity of care documents are designed to be accessible in a number of formats, which may make them easier to use for research purposes. Overall, HIEs may have the potential to be a source of data for vaccine safety research but they vary so widely that each HIE would need to be contacted separately to determine suitability.

D. PRACTICE-BASED RESEARCH NETWORK

1. Practice-Based Research Network 1

Personally Identifying Variables Available. The only personally identifying variables held by the network itself are date of birth and race. All other personally identifying information is held at the member practices and not transmitted to the network itself.

Clinical Variables Available. All practices that participate in the network must use one of three EMR systems Practice Partner, Lytec, or MediSoft's EHR. Data are extracted from these EMR systems periodically and this information is used to create a quarterly summary report that lists each patient seen by the practice and includes information about diagnoses, vital signs, medications used, lab test results and immunizations they have received.

Percent Completeness. The majority of practices (~95%) that participate in the network are family practice and internal medicine practices. Thus, any visits to a specialty provider are unlikely to be captured.

Method of Data Collection. Data are collected through each practice's EMR systems.

Previous Linkage to Payer Data. To our knowledge there have been no previous linkages to payer data, nor did the respondent report any such linkages.

Legal & Regulatory Requirements. IRB approval would be required for any research conducted using this data

Technical Requirements. To access data from this network that is personally identifiable, and thus can be matched to payer data, it would be necessary to approach each individual practice. It is possible that the technical requirements may vary by practice; however the requirement that each member use one of three EMR systems would likely facilitate the process

Budget Requirements. Accessing data would require interacting with each member practice from which data was desired, thus the network as a whole was not able to provide an estimate of the budget requirements.

Overlap with Plan Membership. Member practices are located around the United States, and it is difficult to tell to what extent members of the plans represented by the Data Partners will be found in any given EMR system.

Completeness of Variables of Interest for PRISM. The network does not hold all of the variables of interest for PRISM; however they are all likely to be held by the member practices.

2. Summary/Other Important Details

Several other practice-based research networks responded to our request for information. Each of them indicated that their networks are loosely structured with no data held by the network coordinating center, and also indicated that they did not think it would be possible to collaborate on a vaccine safety study.

E. RETAIL CLINICS

1. Retail Clinic 1

Although several retail clinic chains were invited to participate in this study, only one completed and returned the information gathering tool. A representative from this clinic also participated in a follow-up phone call to further clarify the information provided.

Personally Identifying Variables Available. Most major personal identifiers, including name, date of birth, gender and address are available. Social security number is also available in some cases, although the degree of completeness of this field is unknown. Health insurance information, such as plan name and subscriber number, is available for approximately 60% of people who use their insurance to cover services received at the clinic. Although it is not clear what percent of patients with health insurance submit a claim for services received at a retail clinic the clinic's representative felt that most insured individuals would indeed submit a claim.

Clinical Variables Available. Vaccinations are captured using CPT, ICD-9 diagnosis and HCPCS codes. Adverse events that are associated with vaccination would be captured in the EMR system in free text fields. Every visit to this retail clinic is followed-up the next day with a phone call, thus adverse events that are not immediate may be captured through this call. It is not clear, however, how often the clinic is able to reach the patient the day after the visit, thus this we cannot assess the completeness of the information.

Percent Completeness. Every person who receives a vaccination at a retail clinic should have a record of this vaccination in the clinic's data. Adverse events that occur in the clinic (i.e. immediately post-vaccination) would be captured in the EMR system, and some adverse events that occur in the day following vaccination would also be captured through the follow-up phone call mechanism discussed above.

Method of Data Collection. Data are collected as part of keeping a correct medical record and for billing purposes. This data is collected at each retail clinic site. All data from each individual clinic is aggregated at one central location and one EMR systems is used in all clinics.

Previous Linkage to Payer Data. There have been no previous linkages of this data to payer data. The data have, however, been used in several research activities, including an analysis of the use of retail clinics and prevention of emergency room visits and a study of the appropriateness of antibiotic prescription for bronchitis.

Legal & Regulatory Requirements. Appropriate Business Associate and Data Use Agreements would need to be in place for the health plans to access this data. IRB approval would also be needed, and an external review board – typically Quorum Institutional Review Board – is used for these purposes.

Technical Requirements. No major technical barriers were noted. The clinic's representative did state that typically before releasing data they clean it somewhat to make it easier for the recipient to use.

Budget Requirements. The cost of matching to this data would vary depending on the scope of the project and the number of individuals to be matched. As such, it would need to be discussed on a project-by-project basis.

Overlap with Plan Membership. This is unknown.

Completeness of Variables of Interest for PRISM. Every vaccination given in a retail clinic will be recorded in the EMR system. It is not clear, however, what percent of adverse events would be captured. Adverse events that occur immediately post-vaccination are more likely to be captured than those that occur after the patient has left the clinic.

2. Summary/Other Important Details

To identify an appropriate contact within a retail clinic the Workgroup collaborated with their internal contracting departments. This proved to be an important source of information about the data collected by retail clinics. For example, the Workgroup assumed that in most cases a visit to a retail clinic for a vaccination would not results in a claim being filed with the recipient's insurance company. The contracting group, however, felt that the majority of vaccinations would have an associated claim filed with the insurance company. This issue will need to be explored in greater depth before pursuing a data match; if the majority of vaccines do have an associated health insurance claim matching to data held by retail health clinics may not be useful for identifying individuals who were vaccinated.

VI. CONCLUSION

Table 3 summarizes the key contributions the data sources discussed above could have as well as the key challenges of linking these data sources to claims data.

Table 3. Potential Contributions and Feasibility of Complementary Data Source Linkage with PRISM Claims Data

	Key Contribution	Immediate Challenges to Linking with Claims Data**
Birth Certificate Registries		
Birth Certificate Registry 1	Information on gestational weight and age and some birth defects	The records cannot be released outside the Office of Vital Statistics per state law.
Birth Certificate Registry 2	Information on gestational weight and age and some birth defects	None foreseen, although it may take time to ensure the appropriate regulatory paperwork is in place prior to the match.

	Key Contribution	Immediate Challenges to Linking with Claims Data**
Birth Certificate Registry 3	Information on gestational weight and age and some birth defects	None foreseen, although it may take time to ensure the appropriate regulatory paperwork is in place prior to the match.
Birth Defects Registries		
Registry 1	Data on the occurrence of a wide variety of birth defects	The registry would want assurances that the data would never be transmitted to the subject's insurance company, and other regulatory approvals must be in place.
Registry 2	Data on a small number of birth defects, some of which are detected by active surveillance and others which are detected through passive surveillance	Data does not contain maternal identifiers so in order to link children with mothers you must first link to birth certificate data.
Health Information Exchanges		
HIE 1	Rich clinical detail, including lab values, available.	Not clear to what extent members of the HIE overlap with plan membership. If the degree of overlap is not large this may be an inefficient way of accessing records.
HIE 2	Rich clinical detail, including lab values, available.	The group is unwilling to participate in a vaccine safety study because of significant human resource constraints.
Practice-based Research Networks		
PBRN 1	Working with the PBRN may make member practices more receptive to requests to share medical records.	Data are not held by the network itself, thus it is not clear if working with the PBRN is any more efficient than the Data Partner's current process for accessing data held by providers. The member practices are geographically diverse, while insurance membership tends to be concentrated in certain states.
Other PBRN	Working with the PBRN may make member practices more receptive to requests to share medical records.	Data are not held by the network itself, thus it is not clear is working with the PBRN is any more efficient than the Data Partner's current process for accessing data held by providers.
Retails Clinics		
Retail Clinic 1	If most people with insurance who use the clinic have a claim filed, likely to be of little value. If not, may capture vaccinations.	It was difficult to contact these organizations, which suggests that they may not be interested in collaboration on research projects. In addition, if claims are already captured in payer data there is little value in matching to this data.

The data sources that show the most promise for future PRISM projects seem to be birth certificate data, where they can be released legally, and birth defects registry data. These sources contain appropriate personal identifiers for such a match, even if it needs to be a two part match to the birth certificate data and then the birth defects registry data. They also capture outcomes of interest for PRISM studies, and are generally representative of the population, although birth defects registry data will be somewhat less complete than birth certificate data.

Some of the remaining data sources may be good sources of data for enhancing claims, depending on the degree to which they captures members of the health plans represented by the data partners. For example, if HIE 1 contains data on a large number of lives represented by a data partner matching to this data could provide rich clinical detail that may be not available in another source. HIE 1 has the IT infrastructure in place to enable such a match; however it is not clear if most other HIEs have the capacity to do this. Each HIE would need to be investigated on its own to determine this.

Two data sources, practice-based research networks and retail clinics, are unlikely to prove worthwhile data sources. The PBRNs who responded to our request for information did not hold enough personal identifiers to be matched to claims data, thus we would need to approach each member practice to conduct the linkage. This is not likely to prove any more efficient than the data partners' current process for accessing medical records. Retail clinics present two challenges. First, it is not clear to what extent people with insurance who use a retail clinic would have a claim filed. If the majority of people do have a claim filed on their behalf matching to retail clinic data is unlikely to significantly enhance the claims data already held by the data partners. In addition, it was difficult to find a representative of a retail clinic who was willing to provide information for this project. This suggests that they may not be interested in collaborating with the FDA on any such effort.

VII. APPENDICES

A. APPENDIX A: WORKGROUP MEMBERS AND OTHER CONTRIBUTORS

1. Workgroup Members

Core Members

Robert L. Davis, MD, MPH	Kaiser Permanente
Gregory W. Daniel, PhD, MS, MPH	HealthCore
Vinit Nair, BS Pharm, MS, RPh	Humana
Cheryl N McMahill-Walraven, MSW, PhD	Aetna
Corrie Paeglow, DrPH, MPH	HealthCore

Other Contributors

Joseph Singer, MD	HealthCore
Chris Hetrick, BS	HealthCore
Leslie Ann Hazel-Fernandez, PhD, MPH	Humana
Michael Nguyen, MD	FDA - CBER
Tracy Lieu, MD, MPH	Harvard Pilgrim Health Care Institute
Carolyn Balsbaugh, MPH	Harvard Pilgrim Health Care Institute

2. Individuals who provided feedback on the data sources spreadsheet

Mini-Sentinel Core Leads

Lesley Curtis, PhD
Associate Professor in Medicine
Duke University School of Medicine

Sean Hennessey, PharmD, PhD
Assistant Professor of Pharmacology
University of Pennsylvania

Other Mini-Sentinel Affiliated Personnel

Nancy Dreyer, MPH, PhD
Chief of Scientific Affairs and Senior Vice President
Outcome Sciences, Inc.

Non-Mini-Sentinel Personnel

Sarah Clark, MPH
Associate Director, Child Health Evaluation and Research (CHEAR) Unit
University of Michigan

Arnold Potosky, PhD
Professor of Oncology
Director of Health Services Research
Georgetown University Medical Center