

## **Improving the Efficiency of Outcome Validation in the Sentinel System**

Duke-Margolis Center for Health Policy  
1201 Pennsylvania Ave, NW Suite 500 • Washington, DC 20004  
May 17, 2018

### **Discussion Guide**

#### **Introduction**

The rapid identification of adverse events related to medical products is of paramount concern to the U.S. Food and Drug Administration (FDA), patients, providers, and manufacturers of the product. The relative speed of identification depends on effective data systems and methods of analysis to quickly generate high quality evidence to inform a regulatory decision. To develop this capacity, FDA launched the Sentinel System as an “active surveillance” component of FDA’s postmarket surveillance systems. The Sentinel System, a distributed data network that links data sources from Sentinel Data Partner organizations through a common data model (CDM), augments surveillance data of regulated products reported voluntarily by manufacturers, clinicians, and patients.

The Active Postmarket Risk Identification and Analysis (ARIA) system is the component of the Sentinel System used by FDA for the surveillance of pharmaceutical products. Before using ARIA to conduct a safety assessment, FDA must first determine whether the data and methods under ARIA are sufficient for answering key regulatory questions of interest within a satisfactory level of precision. In 2016, FDA officially launched the Sentinel System, and the ARIA component is routinely being used as part of the agency’s portfolio of post-market surveillance activities. Recently, the FDA achieved a major milestone by including ARIA sufficiency as part of a pre-market decision to approve a new drug application.<sup>1</sup>

While Sentinel has achieved its initial core function to quickly identify and assess safety issues using electronic health care data, key challenges remain to fully realizing the value of this data infrastructure. In some instances the data and methods available in the Sentinel System may be insufficient to inform regulatory decisions. Agency analyses investigating reasons for this insufficiency have identified outcome validation as a major contributing factor.

The gold standard for outcome validation involves auditing medical records to verify the performance characteristics (i.e., positive predictive values (PPV), sensitivity, and specificity) of claims based-algorithms to ensure the validity of study results. To date, validating algorithms typically requires manual chart review to abstract key data elements from source records, which is often a time consuming, expensive, and technically challenging process given available administrative data may not accurately identify the outcome of interest. This process also limits the ability to calculate sensitivity and specificity leaving the PPV as the only practical performance metric for evaluating algorithms.

---

<sup>1</sup> U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Sinuva (mometasone furoate) NDA 209310 approval letter, December 18, 2017. Retrieved April 15, 2018, from: [https://www.accessdata.fda.gov/drugsatfda\\_docs/applletter/2017/209310Orig1s000ltr.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/applletter/2017/209310Orig1s000ltr.pdf)

The Duke-Margolis Center for Health Policy, in cooperative agreement with the FDA, is convening a private workshop to consider potential opportunities to address ARIA sufficiency gaps that are impeding outcome validation. Workshop discussion will explore how innovative technologies including artificial intelligence (AI) tools such as natural language processing (NLP), machine learning techniques, and electronic phenotyping could support more streamlined and automated validation processes. Input will be solicited from diverse stakeholders on the most promising approaches with these technologies, and will inform subsequent discussions between the FDA, Sentinel Data Partners, and Sentinel Operating Center to identify and develop high priority pilots for future implementation.

### **Key Challenges with Sufficiency Analysis in the Active Risk Identification and Analysis (ARIA) System**

The FDA conducts safety assessments in ARIA in concordance with requirements at Section 505(o)(3)(B) of the Federal Food, Drug, and Cosmetic Act. This provision requires that analytic tools in addition to the Sentinel CDM are sufficient for answering key questions of interest. The agency interprets ‘sufficient’ to mean the availability of adequate data (e.g. the drug or biologic of interest, comparators, confounders and covariates, health outcomes of interest) and appropriate tools to provide a satisfactory level of precision for answering a regulatory question. When ARIA is sufficient the system may be used to conduct assessments on:

- known serious risks related to the use of the drug (i.e. signal evaluation)
- signals of serious risk related to the use of the drug (i.e. signal refinement)
- unexpected serious risks when available data indicates potential for serious risk (i.e. signal detection)

Sentinel has informed a variety of regulatory actions including label changes, responding to Citizens Petitions, and becoming part of an Advisory Committee deliberation. Additionally, ARIA is also being used to generate evidence that may address a safety concern and help FDA to determine that no regulatory action is needed.

While there have been successful uses of Sentinel to inform regulatory decisions, there are instances where the infrastructure is deemed insufficient. Preliminary agency analyses have identified the lack of validated algorithms for using the CDM to identify outcomes of interest and study cohorts as a key barrier for ARIA sufficiency. Outcome validation is a critical step in safety assessment in which algorithms, typically based on International Classification of Disease (ICD) codes, are validated to ensure that health outcomes of interest are present in the study’s parameters (e.g. start date and length of follow up). The presence of these outcomes helps to establish temporal relationships between the drug exposure and outcome, and ultimately, whether any causal inferences may exist. The process to validate algorithms requires a manual chart review process to abstract relevant data from source medical records.

Although considered the gold standard, there remain many challenges for efficiently implementing chart review in the Sentinel System. This requires identifying specific records corresponding to the patients driving the risk estimate and working with several vendors at each Data Partner site to find, redact, abstract and adjudicate the medical record. The quantity of supplemental information needed from the medical records can vary from a single laboratory value (e.g., platelets) or multiple streams of information (e.g., vital signs, pathology, radiology data) depending on the outcome. While the Sentinel

System strives to collect standardized outcomes, varied approaches for data capture at different sites can lead to variable data quality and potentially misclassification of charts. Also, most medical charts exist in paper format, which further complicates current efforts to digitize and automate efforts. Collectively, these gaps contribute to ARIA insufficiency and prevent use of the Sentinel System by FDA.

### **Streamlining Outcome Validation: A Foundation for Greater ARIA Sufficiency**

Advancements with computing and artificial intelligence provide new opportunities to streamline the outcome validation process to address key challenges with manual chart reviews. There are potentially two phases of validation that could be improved through technologies and analytical approaches: 1) extracting and prepping standardized data for analyses; and 2) applying advanced analytics and sophisticated algorithms for the rapid identification of cohorts and outcomes.

Technologies that improve data for analysis such as optical character recognition (OCR), innovative abstraction software tools, and natural language processing (NLP) hold great promise for mining data from large and diverse databases. Given many health care settings still operate in a paper-based environment, OCR technology could be used to make images or PDFs of charts searchable or even machine readable and thus easier to identify important data elements. Novel software abstraction tools are being developed to complement OCR technology in a way that strengthens and standardizes data quality for more efficient abstraction. Natural language processing, a domain of artificial intelligence (AI) that enables the extraction of unstructured data into a standardized format based on pre-defined rules, could underpin a comprehensive data abstraction infrastructure that integrates rich clinical text found in electronic health records with administrative and claims data.

Once these data are extracted and standardized for analysis, analytical approaches can then be applied to identify study cohorts and outcomes. One approach includes electronic phenotyping, which could support more efficient uses of Sentinel data. Electronic (also known as computable) phenotypes are specifications for identifying patients or populations with a given characteristic or condition. Phenotyping could support safety assessments in Sentinel by providing standardized definitions based on a curated list of core data elements that would be scalable and re-usable for rapid querying across multiple settings and information systems.<sup>2</sup> Examples of distributed research networks developing and applying electronic phenotypes in research settings include the electronic Medical Records and Genomics (eMERGE) network and NIH Collaboratory among other efforts.<sup>3</sup>

Machine learning techniques, another domain of AI, provide potentially powerful tools that could also support streamlined outcome validation. In context of the Sentinel System, machine learning refers to the application of sophisticated prediction algorithms to varied data sources capable of identifying patients and populations with specific conditions not readily identifiable with existing claims data. If these data are in a machine readable format, then machine learning techniques, potentially in combination with NLP, could support “supervised” and in some cases even “unsupervised” statistical learning in which phenotypes are actively identified without pre-defined rules and annotated data. Such

---

<sup>2</sup> Richesson, R et al. A Framework to Support the Sharing and Re-Use of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. Electronic Data Methods (EDM) Forum. eGEMs (Generating Evidence & Methods) to improve patient outcomes. July 2016.

<sup>3</sup> Kirby, C, Speltz, P, Rasmussen, L et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016; 23:1046-1052.

approaches could be used to quickly identify important study cohorts and health outcomes while still maintaining high performance metrics of the algorithms.

### **Key Implementation Challenges for Developing and Applying Artificial Intelligence Tools in the Sentinel System**

A streamlined outcome validation process leveraging AI capabilities to support rapid identification of cohorts and outcomes is expected to improve FDA decision making. This will require developing the necessary data infrastructure suitable for NLP, machine learning, and other technologies for use in a distributed production environment. Effectively using these tools will require building a corpus of medical records that could support automated processes to extract key data elements while still maintaining compliance with all privacy and security regulations. There is growing experience with building pre-annotated records in health care contexts, however, considerable resources are usually needed for initial implementation.<sup>4</sup> Further, algorithm refinements that enable more sophisticated statistical approaches that can be applied to both structured and unstructured data may be needed to scale and reuse phenotypes across diverse health care settings.

In addition to these technical challenges, there are also unique considerations for how such capabilities could be implemented within Sentinel's distributed data environment. Depending on the approach, there will be different implications for governance, processes, and the types of tools required. Key considerations include:

- Impact on Sentinel's CDM including necessary expansions of the data model,
- Impact on Sentinel's existing legal and governance frameworks for how data is shared, including the confidentiality of patient data,
- Impact on the quality review and curation of data in the CDM, and
- Impact on the suite of tools for distributed querying.

As a first step in determining whether and how these tools could be incorporated into the Sentinel System, the initial focus will center on understanding the requirements and resources needed to improve data and methods sufficiency for streamlined outcome validation. The input received at the workshop will be used to support FDA's efforts to build a stronger foundation for ARIA sufficiency, and inform strategic planning of potential implementation approaches for more efficient outcome validation in collaboration with Sentinel Data Partners.

Funding for this conference was made possible in part by a cooperative agreement from the U.S. Food and Drug Administration Center for Drug Evaluation and Research. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services nor does mention of trade names, commercial practices, or organizations imply endorsements by the U.S. Government.

---

<sup>4</sup> Botsis, T, Buttolph, T, Nguyen, M et al. Vaccine adverse event text mining for extracting features from vaccine safety reports. J Am Med Inform Assoc. 2012; 19:1011-1018.

## Appendix A: Brief Overview of the Sentinel System

The Sentinel System comprises multiple medical product surveillance efforts – called components – within the Sentinel Initiative. The ARIA component assesses the safety of pharmaceuticals and works in coordination with these other components. The Blood Surveillance Continuous Active Network (BloodSCAN) supports the regulatory review of blood and blood products, the Postmarket Rapid Immunization Safety Monitoring (PRISM) system monitors vaccine safety, and the Surveillance of Tissues and Advanced Therapeutics (STAT) program focuses on surveillance and recipient safety evaluation of human cell-, gene-, tissue-based products, other advanced therapies, and antivenoms. The agency is also adding new capabilities that support a growing range of functions to strengthen the enterprise of evidence generation including the FDA-Catalyst and Biologics Effectiveness and Safety (BEST) initiatives.<sup>5,6</sup>

Sentinel System components rely on a shared data infrastructure based on the Sentinel Initiative's early pilot program, the Mini-Sentinel pilot. Launched in 2009, Mini-Sentinel was designed to test the feasibility of Sentinel's core function to access and analyze healthcare information from a variety of data sources, and to use that data to improve FDA decision making. These data sources are derived from Sentinel Data Partners consisting of national health insurance plans, large integrated delivery systems, and health care organizations.<sup>7</sup> They share summary information derived from longitudinal claims data, and growing amounts of clinical data derived from electronic health records (EHRs), through a common data model (CDM).

The CDM makes it possible to execute standardized programs developed by the SOC by allowing Data Partners to maintain and access data in a common data format. The CDM relies on existing standardized coding schema (e.g., ICD-9-CM, HCPCS/CPT, and NDC) to minimize technical challenges with interoperability and is continuously being maintained, updated, and enhanced by the SOC. The CDM was built based on the data elements found in electronic health care data from Data Partners, which includes administrative claims data, some outpatient and inpatient electronic health records, demographic information, outpatient pharmacy dispensings, and registry data.<sup>8</sup>

Once a safety question is prioritized, the FDA in partnership with the Sentinel Operations Center (SOC) develops standardized data queries, which are then distributed to Data Partners for analysis. There are a range of tools available under ARIA for executing safety assessment queries in support of signal detection, signal refinement, and signal evaluation. The output of these queries, typically in summary form, are then provided back to the SOC for analysis.

Data Partners maintain physical and operational control over their electronic health care data in their existing environments (i.e. behind their respective firewalls), which helps Sentinel Data Partners to adhere to all privacy provisions in concordance with the Health Insurance Portability and Accountability Act (HIPAA). In addition, the Office for Human Research Protections has determined that the regulations administered by that office (45 CFR part 46), which govern Institutional Review Board (IRB)

---

<sup>5</sup> Graphical depictions of Sentinel operations and organization of the data infrastructure are provided in Appendices B and C.

<sup>6</sup> A glossary of Sentinel terms is provided in Appendix D.

<sup>7</sup> A list of current Sentinel Data Partners is provided in Appendix E.

<sup>8</sup> A list of data variables presently captured by the CDM is provided in Appendix F.

requirements, do not apply to the activities under the Sentinel System.<sup>9</sup> Instead, these activities are considered public health and implemented under the agency's public health authority. However, not all Sentinel activities can be implemented in this manner. For example, studies conducted under the FDA-Catalyst initiative are considered research and must comply with IRB regulations.

---

<sup>9</sup> For more background on IRB compliance in Sentinel please see the report by Kristen Rosati et al. cited in Appendix G.

## Appendix B: Depiction of Query Execution in the Sentinel System

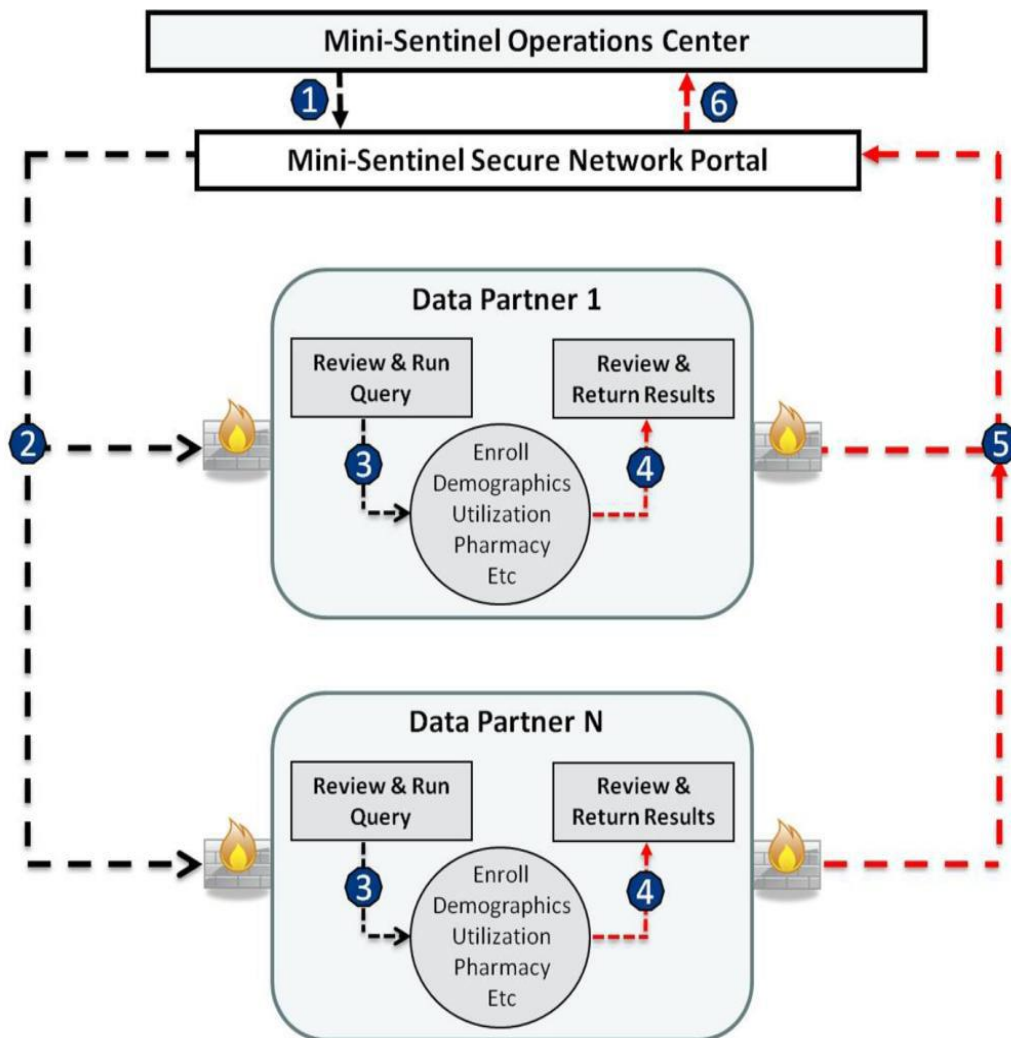


Figure 1. Overview of Mini Sentinel Distributed Network. (Source: Dr. Richard Platt, Harvard Pilgrim Health Care Institute).



## Appendix C: Organization of the Sentinel Initiative

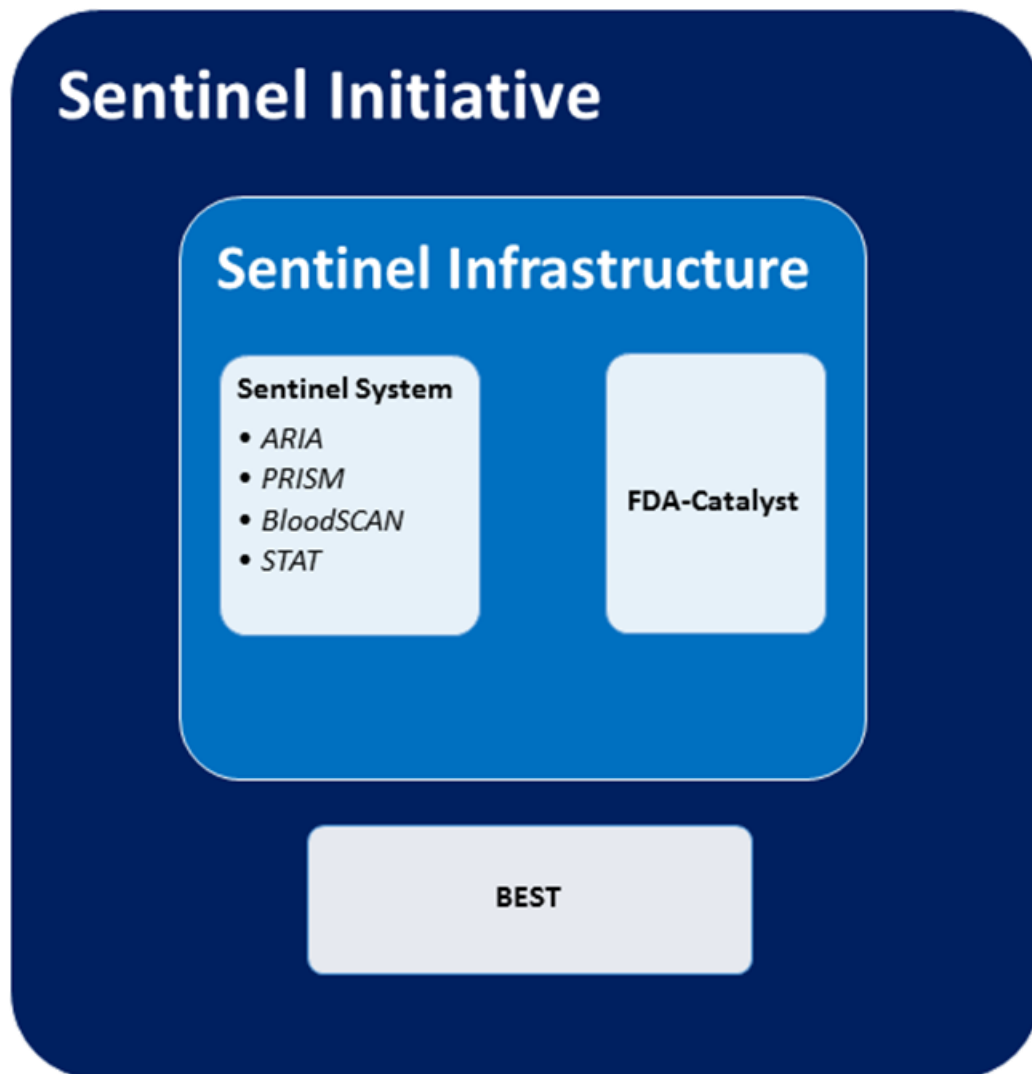


Figure 2. Organization of Sentinel Initiative. (Source: Adapted from FDA depiction obtained on the Sentinel Initiative website: <https://www.sentinelinitiative.org/sentinel/about>).



## Appendix D: Glossary of Sentinel Terms <sup>10</sup>

**Active Risk Identification and Analysis (ARIA):** The U.S. Food and Drug Administration's (FDA) active post-market risk identification and analysis system, which is comprised of pre-defined, parameterized, reusable routine querying tools, combined with the electronic data in the Sentinel Common Data Model. Because ARIA uses parameterized tools and a trusted multi-site distributed database that undergoes continuous quality checks and refreshes, safety analyses can be done more efficiently to conduct medical product safety surveillance to fulfill the mandate in the FDA Amendments Act of 2007.

**Biologics Effectiveness and Safety (BEST) Initiative:** A system currently being developed to expand and enhance FDA's Center for Biologics Evaluation and Research (CBER) post-market surveillance capabilities. The primary goals of BEST are a) to develop a system using electronic health records (EHR) and claims data sources covering a large proportion of the U.S. population, automated query tools, and additional infrastructure; b) to develop improved automated adverse events data collection, analysis, and reporting techniques for biologics by using methods such as natural language processing and machine learning.

**Blood Safety Surveillance Continuous Active Network (BloodSCAN):** A subcomponent of the CBER Sentinel Program focusing on surveillance and recipient safety evaluation of blood components and blood-derived products.

**Cohort Identification and Descriptive Analysis (CIDA) Tool:** CIDA serves as the foundation of the routine querying system, and is responsible for identifying, extracting, and characterizing cohorts of interest from the SDD based on the specification of a number of requester-defined options (e.g., continuous enrollment requirements, incidence criteria, inclusion/exclusion criteria).

**FDA-Catalyst:** Activities leverage the Sentinel Infrastructure by utilizing the data available through its data partners and supplementing it with data from interventions or interactions with members and/or providers.

**Post-licensure Rapid Immunization Safety Monitoring (PRISM):** A subcomponent of CBER Sentinel Program focusing on vaccine safety surveillance for evaluation of potential safety signals identified during pre-market and post-market reviews.

**Routine Querying Tools (Modular Programs):** Sentinel's routine querying tools include modular programs, summary tables, and software toolkits. Modular programs are grouped into three levels:

- **Level 1** modular program queries identify cohorts of interest and, for some cohorts, can perform unadjusted and minimally adjusted (i.e., by Data Partner, age group, sex, and year) analyses.

---

<sup>10</sup> Glossary taken from supplementary materials provided at the 10th Annual Sentinel Initiative Public Workshop convened by the Duke-Margolis Center for Health Policy on February 7, 2018. For more information please see: <https://healthpolicy.duke.edu/events/2018-sentinel-initiative-annual-public-workshop>.

- **Level 2** modular program queries identify cohorts of interest, perform more complex adjustment for confounding, and generate effect estimates and confidence intervals.
- **Level 3** modular program queries identify cohorts of interest and perform more complex adjustment for confounding repeatedly as part of prospective sequential analysis.

**Sentinel Collaborating Institutions:** A network of Data and Academic Partners that work with the FDA and Sentinel Coordinating Center to provide access to both healthcare data and scientific, technical, and organizational expertise.

**Sentinel Coordinating Center:** The Sentinel Coordinating Center includes the Sentinel Operations Center (SOC), comprised of the Applied Surveillance, Scientific Systems, and Administration Divisions housed at the Harvard Pilgrim Health Care Institute (HPHCI), and advisory groups. Both the Sentinel Coordinating Center and the SOC are led by the Sentinel Principal Investigator at HPHCI.

**Sentinel Data Partners:** Data Partners in the Sentinel System include a diverse group of organizations including academic medical centers, healthcare systems, and health insurance companies. Sentinel Data Partners maintain physical and operational control over electronic data in their existing environments.

**Sentinel Infrastructure:** The underlying data infrastructure created to enable analysis within the Sentinel System. The Sentinel Infrastructure involves: 1) a distributed data approach in which Data Partners maintain physical and operational control over electronic data in their existing environments; and 2) a Common Data Model consisting of standardized administrative and clinical information across Data Partners. The Sentinel Infrastructure has the potential to allow analysis of the data for other purposes besides safety for the FDA or those outside the FDA.

**Sentinel Initiative:** A multi-year effort beginning in 2008 to create a national electronic system for monitoring the performance of FDA-regulated medical products to improve the FDA's ability to identify and assess medical product safety issues.

**Sentinel System:** An active surveillance system that uses routine querying tools and pre-existing electronic healthcare data from multiple sources to monitor the safety of regulated medical products. Subcomponents of the Sentinel System include: ARIA, PRISM, BloodSCAN and STAT.

**Surveillance of Tissues and Advanced Therapeutics (STAT):** A subcomponent of the CBER Sentinel Program focusing on surveillance and recipient safety evaluation of human cell-, gene-, tissue-based products, other advanced therapies, and antivenins.

## Appendix E: List of Current Sentinel Data Partners <sup>11</sup>

\* Indicates Collaborating Institutions that are also Data Partners

- Aetna Informatics\*
- America's Health Insurance Plans: Clinical Affairs Department
- Blue Cross Blue Shield of Massachusetts\*
- Brigham and Women's Hospital: Division of Pharmacoepidemiology & Pharmacoeconomics in the Department of Medicine
- Department of Population Health Sciences, Duke University School of Medicine\*
- Harvard T.H. Chan School of Public Health
- HealthCore, Inc. Government & Academic Research\*
- Health Care Systems Research Network
  - Harvard Pilgrim Health Care Institute\*
  - HealthPartners Institute\*
  - Henry Ford Health System: Public Health Sciences Department
  - Marshfield Clinic Research Institute\*
  - Meyers Primary Care Institute, a joint endeavor of Fallon Community Health Plan\*
- Hospital Corporation of America\*
- Humana, Inc., Comprehensive Health Insights\*
- IQVIA
- Kaiser Permanente Center for Effectiveness and Safety Research
  - Kaiser Permanente Colorado Institute for Health Research\*
  - Kaiser Permanente Center for Health Research Hawaii\*
  - Kaiser Foundation Health Plan of the Mid-Atlantic States, Inc.\*
  - Kaiser Permanente Northern California, Division of Research\*
  - Kaiser Permanente Northwest Center for Health Research\*
  - Kaiser Permanente Washington Health Research Institute\*
- Optum\*
- Rutgers University: Center for Health Services Research on Pharmacotherapy, Chronic Disease Management and Outcomes at the Institute for Health, Health Care Policy and Aging Research
- University of Alabama at Birmingham: Center for Outcomes and Effectiveness Research and Education
- University of Florida College of Pharmacy: Department of Pharmaceutical Outcomes & Policy
- University of Illinois at Chicago: Department of Pharmacy Systems, Outcomes and Policy
- University of Iowa: Department of Epidemiology in the College of Public Health
- University of North Carolina Gillings School of Global Public Health: Center for Pharmacoepidemiology
- University of Pennsylvania School of Medicine: Center for Clinical Epidemiology and Biostatistics and Department of Biostatistics and Epidemiology
- Vanderbilt University School of Medicine, Department of Health Policy\*
- Weill Cornell Medicine, Healthcare Policy & Research

<sup>11</sup> List adapted from: <https://www.sentinelinitiative.org/collaborators>.

## Appendix F: Overview of the Sentinel Common Data Model

Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Enrollment start & end dates	Birth date	Dispensing date	Service date(s)	Service date(s)	Service date(s)
Drug coverage	Sex	National drug code (NDC)	Encounter ID	Encounter ID	Encounter ID
Medical coverage	ZIP code	Days supply	Encounter type & provider	Encounter type & provider	Encounter type & provider
Medical record availability	Etc.	Amount dispensed	Facility	Diagnosis code & type	Procedure code & type
			Etc.	Principal discharge diagnosis	Etc.

Lab Result	Vital Signs	Inpatient Pharmacy	Inpatient Transfusion	Death	Cause of Death
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Result and specimen collection dates	Measurement date and time	Administration date and time	Blood product code and type	Death date	Cause of death
Test type, immediacy & location	Height and weight	Encounter ID	Encounter ID	Source	Source
Logical Observation Identifiers Names and Codes (LOINC ®)	Diastolic & systolic BP	National Drug Code (NDC)	Blood type	Confidence	Confidence
Test result & unit	Tobacco use & type	Route	Administration start and end dates and times	Etc.	Etc.
Etc.	Etc.	Dose	Etc.		
		Etc.			

Figure 3. Overview of the Sentinel Common Data Model. (Source: Dr. Richard Platt, Harvard Pilgrim Health Care Institute).

## Appendix G: List of Reading Material

- Kennedy Adee et al. "Sentinel Medical Chart Review Gap Analysis Public Report." *Sentinel Methods*, (2018). Retrieved from [https://www.sentinelinitiative.org/sites/default/files/Methods/Chart\\_Review\\_Gap\\_Analysis\\_Report.pdf](https://www.sentinelinitiative.org/sites/default/files/Methods/Chart_Review_Gap_Analysis_Report.pdf)
- David S. Carrell et al. "Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence." *American Journal of Epidemiology*, (2014). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24488511>
- Glen T. Schumock et al. "Alternative Methods for Health Outcomes of Interest Validation." *Mini-Sentinel Methods*, (2013). Retrieved from [https://www.sentinelinitiative.org/sites/default/files/SurveillanceTools/ValidationsAndLiterature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation\\_0.pdf](https://www.sentinelinitiative.org/sites/default/files/SurveillanceTools/ValidationsAndLiterature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation_0.pdf)
- Jacqueline C. Kirby et al. "PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability." *Journal of American Medical Informatics Association*, (2016). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27026615>
- Kristen Rosati et al. "HIPAA and Common Rule Compliance in the Sentinel Initiative." *Sentinel Initiative Principles and Policies*, (2018). Retrieved from <https://www.sentinelinitiative.org/sites/default/files/Communications/Publications/HIPPA-Common-Rule-Compliance-in-Sentinel-Initiative.pdf>
- Lesley H. Curtis et al. "Design considerations, architecture, and use of the Mini-Sentinel distributed data system." *Pharmacoepidemiology and Drug Safety*, (2012). Retrieved from <https://onlinelibrary.wiley.com/doi/full/10.1002/pds.2336>
- Melissa A. Robb et al. "The US Food and Drug Administration's Sentinel Initiative: Expanding the horizons of medical product safety." *Pharmacoepidemiology and Drug Safety*, (2012). Retrieved from <https://onlinelibrary.wiley.com/doi/full/10.1002/pds.2311>
- Rachel L. Richesson et al. "A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications." *The Journal for Electronic Health Data and Methods*, (2016). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975566/>
- Robert Ball et al. "The FDA's Sentinel Initiative—A comprehensive approach to medical product surveillance." *Clinical Pharmacology & Therapeutics*, (2015). Retrieved from <https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.320>
- Taxiarchis Botsis et al. "Vaccine adverse event text mining system for extracting features from vaccine safety reports." *Journal of the American Medical Informatics Association*, (2012). Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22922172>

Vibhu Agarwal et al. "Learning statistical models of phenotypes using noisy labeled training data."  
*Journal of American Medical Informatics Association*, (2016) Retrieved from  
<https://www.ncbi.nlm.nih.gov/pubmed/27174893>