Sentinel Innovation Center Master Plan

Sentinel Innovation Center

Version 1

September, 2023

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Initiative is funded by the FDA through the Department of Health and Human Services (HHS) Contract number 75F40119D10037. The Sentinel Innovation Center is funded by the FDA through HHS Contract number 7540119D10037.

Sentinel Innovation Center Master Plan

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	9/27/2023	Original Version	Sentinel Innovation Center

Executive summary

To achieve its Sentinel System Five-Year Strategy 2019-2024, the US Food and Drug Administration (FDA) established three new Sentinel centers – the Operations Center, the Innovation Center, and the Community Building and Outreach Center. A main focus of the five-year strategy is incorporating emerging data science innovations, such as natural language processing and machine learning, and expanding access to and use of electronic health record (EHR) data. The Sentinel Innovation Center, in particular, is charged with addressing several high-potential innovation themes, including natural language processing, advanced analytics, novel data sources and data interoperability, and emerging disruptive technologies. The Innovation Center has established a robust infrastructure comprising four academic-based lead hubs and a broad network of collaborators in academia, industry, and other settings with deep expertise in these innovation themes. This document lays out the mission, vision, and the five-year Master Plan for the Innovation Center.

With a vision of improving public health, the mission of the Sentinel Innovation Center is to improve human health by optimizing the sufficiency of Sentinel's Active Risk Identification and Analysis (ARIA) capabilities to cost-effectively use electronic health care data sources for medical product safety surveillance and expanding the utility of real-world data for regulatory decision-making. The vision of the Center is to establish a query-ready distributed data network containing EHRs with accompanying methods and reusable analysis tools.

The Innovation Center – in collaboration with FDA and in consultation with the Operations Center and the Community Building and Outreach Center – has developed a Master Plan for the prioritization, development, and incorporation of innovative technologies and new data sources into the Sentinel System to help FDA achieve the five strategic aims laid out in the Sentinel System Five-Year Strategy 2019-2024. The Master Plan development was guided by specific needs and use cases identified by FDA, particularly a need to expand access to and use of EHRs and other data sources to address a lack of valid and robust computable phenotypes for many health outcomes of interest in Sentinel queries.

The Master Plan framework initially involved four key strategic priorities: (1) data infrastructure; (2) feature engineering; (3) causal inference; and (4) detection analytics, with the addition of a fifth priority area 'Innovation Incubator' to encourage research on topics that are more novel and broader than the four previously identified priority areas. Each strategic priority has a set of goals for achieving the Innovation Center vision. The first two strategic priorities, data infrastructure and feature engineering, focus on establishing a query-ready distributed data network containing EHRs; while causal inference and detection analytics focus on developing and evaluating methods and clarifying which approaches should be developed into reusable analysis tools. In this year, the Master Plan working group agreed to add a strategic priority area for Use Case Implementation "UC" in Year 5. Projects under the 'UC' strategic area will bring learnings from the other workstream together to showcase how the infrastructure and analytic capabilities developed by IC will be utilized solve typical Sentinel use cases.

The Master Plan defines goals and specific outputs for each strategic priority area. To achieve the goals, the Innovation Center Master Plan Workgroup outlined a set of initiatives aimed at generating outputs that will become the building blocks for the query-ready distributed data network containing EHRs and the reusable analysis tools. To oversee the development and conduct of the Master Plan initiatives and the successful development of their outputs, the Innovation Center has established four Innovation Cores aligned with the strategic priority areas. Each Core is co-led by two Innovation Center collaborators and a Sentinel Operations Center liaison. The Cores are responsible for ensuring that the appropriate projects are identified and initiated in order to generate the outputs that are necessary to achieve the goals of each strategic priority. Together, the Cores contribute to the success in achieving the overall vision of the Innovation Center.

Introduction

In 2019, the FDA established three new centers as part of the Sentinel System – the Operations Center, the Innovation Center, and the Community Building and Outreach Center. These centers were created to help FDA achieve the Sentinel System Five-Year Strategy 2019-2024, which focuses on incorporating emerging data science innovations, such as natural language processing and machine learning, and expanding access to and use of electronic health record data.

The Innovation Center, in particular, was created to increase and diversify the pathways for external investigators to engage with the Sentinel System for methods development, free up limited additional resources at the Operations Center to improve efficiency and production speed, enhance analytic tools, and accelerate novel data source acquisition and evaluation; provide mechanisms for broadening capacity enabling growth in the quantity and breadth of questions that can be addressed in the Sentinel System; and attract new data partnerships to improve system sustainability through continued diversification of the data network.

FDA outlined six strategic aims for Sentinel, and tasked the Innovation Center, together with the Operations Center, with addressing the three aims that are driven by specific legislative mandates: (1) optimizing the sufficiency of ARIA to cost-effectively use secondary electronic health care data sources for drug safety surveillance (FDA Amendments Act of 2007); (2) evaluating the use of real-world data for regulatory decision making (21st Century Cures Act); and (3) establishing a query-ready, quality-checked distributed data network containing electronic health records on at least 10 million lives with reusable analysis tools (RWE Data Enterprise). These mandates inform the mission and vision of the Innovation Center.

To achieve its mission and vision, the Innovation Center established a robust infrastructure comprising four academic-based lead hubs – the Brigham and Women's Hospital's Division of Pharmacoepidemiology and Pharmacoeconomics, the Duke Clinical Research Institute, Kaiser Permanente Washington Health Research Institute together with the University of Washington School of Public Health, and the Vanderbilt University Medical Center Department of Biomedical Informatics – and a broad network of collaborators in academia, industry, and other settings with deep expertise in the high-potential innovation themes outlined in the Sentinel System Five-Year Strategy 2019-2024, including natural language processing, advanced analytics, novel data sources, data interoperability, and emerging disruptive technologies.

The Innovation Center – in collaboration with FDA and in consultation with the Operations Center and the Community Building and Outreach Center – has developed a Master Plan for the prioritization, development, and incorporation of innovative technologies and new data sources into the Sentinel System to help FDA achieve the strategic aims laid out in the Sentinel System Five-Year Strategy 2019-2024:[1]

- Enhance the foundation of the Sentinel System (data, infrastructure, operations, technology)
- Further enhance safety analysis capabilities by leveraging advances in data science and signal detection
- Accelerate access to broader use of real-world data for generation of real-world evidence
- Create a national resource and further open the Sentinel System by broadening the Sentinel user base

- Disseminate knowledge and advance regulatory science to encourage innovation and meet Agency scientific needs

This report describes the mission and vision of the Innovation Center and its Master Plan, including key strategic priorities, the initiatives necessary for addressing the priorities, and the outputs of these initiatives that serve as the building blocks toward fulfilling the Innovation Center's vision and FDA's Sentinel System Five-Year Strategy 2019-2024.

Mission

The mission of the Sentinel Innovation Center is to improve human health by expanding ARIA capabilities to effectively use electronic health care data sources for medical product safety surveillance and increase confidence in and use of real-world data for regulatory decision-making.

Vision

The vision of the Sentinel Innovation Center is to establish a query-ready distributed data network containing electronic health records and accompanying methods and analysis tools.

Key FDA needs addressed by the Master Plan

ARIA insufficiency

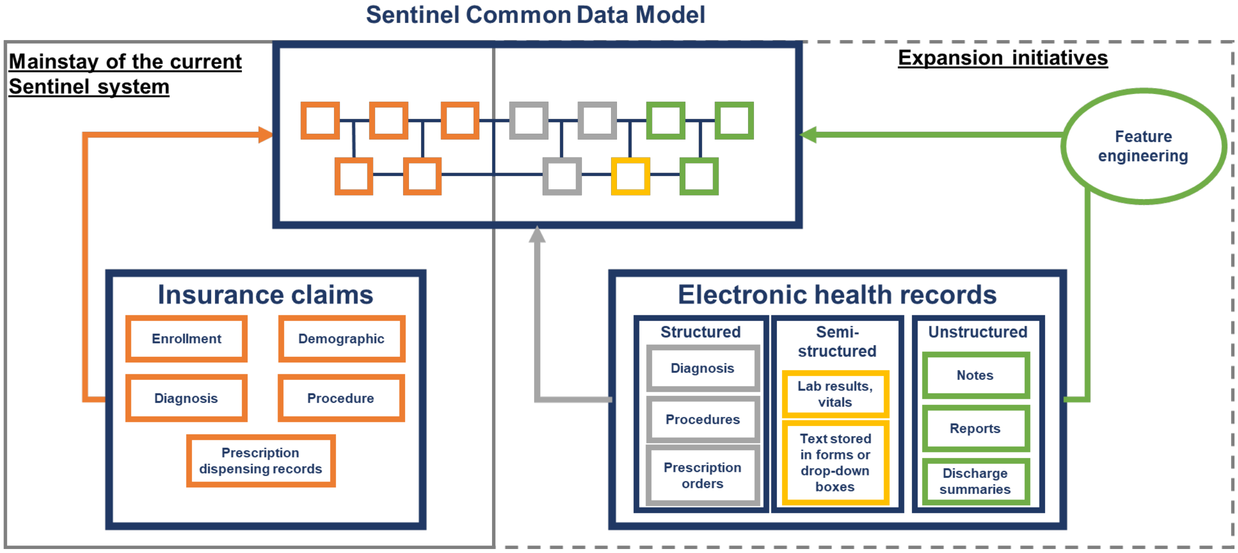
Sentinel's ARIA System comprises electronic healthcare data – including existing electronic health records – from Sentinel's data partners that are formatted in the Sentinel Common Data Model combined with Sentinel's parameterizable analytic tools that enable analyses to be done efficiently and at scale without the need for extensive *de novo* programming for each analysis.[3] At the end of 2019, FDA undertook an analysis of 211 medical product safety issues (i.e., product-outcome pairs) identified between Fall 2015 and November 2019 to determine whether the capabilities of ARIA (i.e., the electronic data in the Sentinel Common Data Model and the existing Sentinel analytic tools) were sufficient to meet the specific study purpose for each safety issue. FDA determined ARIA to be sufficient to address 113 (54%) of the product-outcome pairs.

In this recent analysis of ARIA capabilities, some of the most frequently cited reasons for the inability of using the current claims-based Sentinel for safety investigations included the lack of clinical details to accurately identify health outcomes, missing or inaccurate measures of important confounding variables, or unavailability of computable phenotyping algorithms to identify study populations with acceptable accuracy. Linking EHRs which contain granular information related to clinical parameters, with insurance claims will likely address some of the current gaps in Sentinel's capabilities.

Identifying and incorporating fit-for-purpose data sources

To expand Sentinel's access to EHR data, a first-order goal is to develop the organizational framework and establish the governance, harmonization, and quality assurance processes for ensuring high-fidelity, fit-for-purpose data to support queries of regulatory importance. To support causal conclusions, establishing a clear temporal sequence of events based on data from sources with near complete longitudinal capture is imperative. As most EHR sources in the US lack the ability to capture data when individuals receive care outside of the contributing healthcare systems, linkage of EHRs to insurance claims, which captures longitudinal data regardless of the care settings, is necessary to understand the completeness of longitudinal data. In establishing a query-ready distributed data network containing EHRs, the Sentinel Innovation Center will address key regulatory needs including determining where to source the EHR data with standing linkage to insurance claims and defining the minimum data elements necessary in order to address use cases that are currently difficult to address. Additionally, we will outline principles and strategies for determining how to organize both the structured, semi-structured, and unstructured EHR data alongside insurance claims data in a common data model to facilitate standardized query implementation (Figure 1).

Figure 1. Conceptual overview of integration of claims data and electronic health records in Sentinel



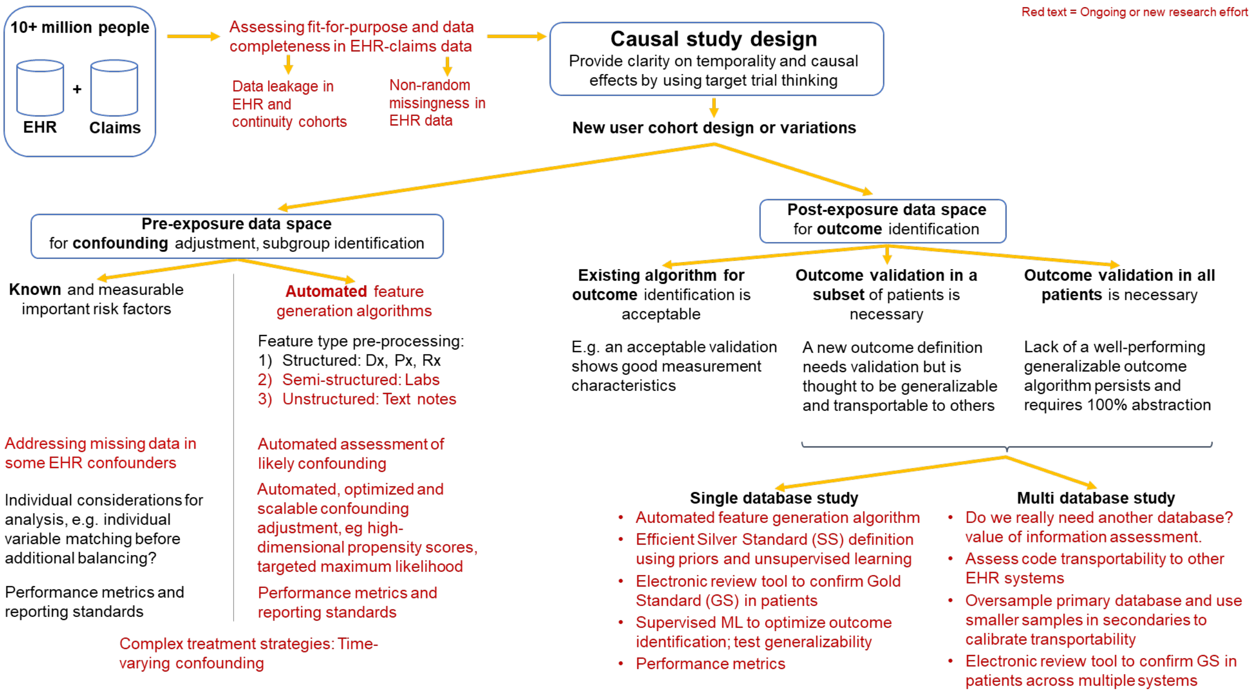
A system rooted in a causal analysis framework

To comprehensively monitor safety of marketed medical products, Sentinel investigations focus on both signal identification, which is intended to generate hypotheses regarding unsuspected adverse events, as well as signal refinement and evaluation, which is intended to test previously generated hypotheses and identify susceptible populations.[1] As regulatory decisions are predicated on inferring causal relations between a medical product exposure and adverse outcomes, a causal analysis framework fitting non-randomized treatment allocation and secondary data is needed.[2] Even for signal identification based on data mining methods, attention to principles including clearly established temporality of confounder assessment preceding exposure followed by outcome surveillance, paired with analytic strategies that reduce confounding, is required to limit the number of spurious signals.[3-5] For signal refinement and evaluation, in addition to these important principles, further considerations include explicitly specifying comparison groups, study populations, and specific outcome definitions ideally contemplated in a hypothetical ‘target’ trial which investigators then emulate in the secondary data of Sentinel.[6,7] Analyses further should be accompanied by robustness evaluations to address the consistency of evidence with respect to alternative investigator decisions in study design, analysis, or variable measurement. To clarify challenges in developing and applying causal methods that leverage both claims and EHR data sources, the Sentinel Innovation Center will develop a causal analysis framework proposing a stepwise process that systematically considers key choices with respect to design and analysis that influence the validity of studies conducted with non-randomized data. A standardized “industrial” process that will be outlined in this framework will serve as a valuable tool to inform the conduct and assessment of the quality of non-randomized studies of drug-outcome evaluation.

EHR data-specific innovation needs

While EHR data offer great promise for improving Sentinel's capabilities, including improvements in computable phenotyping and confounding control, they also bring a range of measurement challenges that must be addressed. Expanded measurement procedures and analytic approaches are outlined in **Figure 2,** identifying a range of Innovation Center projects. Briefly, we focus on distinct measurement challenges 1) to determine patient characteristics in the pre-exposure space (left side) where automated feature identification processes may feed directly into data adaptive confounding adjustment procedures,[8] and 2) to identify and expeditiously validate health outcomes of interest in the post-exposure space (right side) by augmenting claims-based algorithms with EHR-based data that may be structured, semi-structured, or unstructured.

Figure 2. Methodological research needs to support FDA safety decision-making with linked EHR-claims data


Red text = Ongoing or new research effort
10+ million people
EHR
+
Claims
Assessing fit-for-purpose and data completeness in EHR-claims data
Data leakage in EHR and continuity cohorts
Non-random missingness in EHR data
Causal study design
Provide clarity on temporality and causal effects by using target trial thinking
New user cohort design or variations
Pre-exposure data space for confounding adjustment, subgroup identification
Known and measurable important risk factors
Automated feature generation algorithms
Feature type pre-processing:
1) Structured: Dx, Px, Rx
2) Semi-structured: Labs
3) Unstructured: Text notes
Addressing missing data in some EHR confounders
Individual considerations for analysis, e.g. individual variable matching before additional balancing?
Performance metrics and reporting standards
Automated assessment of likely confounding
Automated, optimized and scalable confounding adjustment, eg high-dimensional propensity scores, targeted maximum likelihood
Performance metrics and reporting standards
Complex treatment strategies: Time-varying confounding
Post-exposure data space for outcome identification
Existing algorithm for outcome identification is acceptable
E.g. an acceptable validation shows good measurement characteristics
Outcome validation in a subset of patients is necessary
A new outcome definition needs validation but is thought to be generalizable and transportable to others
Outcome validation in all patients is necessary
Lack of a well-performing generalizable outcome algorithm persists and requires 100% abstraction
Single database study
• Automated feature generation algorithm
• Efficient Silver Standard (SS) definition using priors and unsupervised learning
• Electronic review tool to confirm Gold Standard (GS) in patients
• Supervised ML to optimize outcome identification; test generalizability
• Performance metrics
Multi database study
• Do we really need another database? value of information assessment.
• Assess code transportability to other EHR systems
• Oversample primary database and use smaller samples in secondaries to calibrate transportability
• Electronic review tool to confirm GS in patients across multiple systems

The Innovation Center has formed four main domains as strategic priority initiatives to deliver a query-ready, large-scale data infrastructure of combined EHR and claims data: data infrastructure, feature engineering, causal inference, detection analytics.

Strategic priority 1: Data infrastructure development

EHR data are heterogeneous in their content, structure, completeness, and quality. To enable efficient and timely querying of EHRs to support surveillance activities, pre-processing to make them query-ready is critical. To achieve this, the Sentinel Innovation Center is developing a principled approach to extend the Sentinel Common Data Model to include new data elements from structured and unstructured EHR data (Figure 1). While common data models can help to impose a standard organization of the data across multiple sites, mapping of source values into the model can lead to omissions, data errors, and other data consistency issues.[9] Even though a data model standardizes the name of data elements, a lack of semantic interoperability across sites may remain. To address this challenge, we will investigate approaches to detect and mitigate data consistency issues and to develop and harmonize data across multiple EHR data sites. Further, we will focus on development of a set of data quality metrics and approaches for integration of structured and unstructured data elements from EHR into a common data model to facilitate reliable analyses of medical product outcomes using these data.

Strategic priority 2: Feature engineering approaches

Health insurance claims data, which have been the primary source of data within the Sentinel System, are highly structured with all information stored using standard terminologies such as International Classification of Diseases (ICD) codes for diagnoses and National Drug Codes (NDC) for filled prescriptions. Some of the EHR-based data such as administrative data, medication orders, and most laboratory testing results are recorded as structured or semi-structured fields, but a vast amount of potentially useful information is stored in visit notes (e.g., narrative descriptions of a patient's signs and symptoms, family history, social history), radiology reports or images, and discharge summaries as unstructured data. Substantial engineering is needed to identify features from unstructured data that can be extracted and organized as structured data.

Natural language processing (NLP) and automated feature extraction are essential mechanisms to support scalable computable phenotyping from EHRs in signal detection and signal refinement activities. Development of automated feature extraction workflows that allow for time-contextualization is critical to enable determination of temporality in confounder, exposure, and outcome assessment in Sentinel queries. However, there are numerous challenges in using these approaches at scale in national consortia. NLP tool performance between sites has been a challenge due to systematic data changes. Tool development further needs to be tailored to eventual application of the derived features. For instance, identification of health outcomes of interest needs focused tools to ensure optimal performance characteristics including positive predicted value and specificity, while identification of potential confounders may utilize more flexible tools utilizing unsupervised approaches for high dimensional feature vector generation. To address these challenges, the Sentinel Innovation Center has initiated a range of activities. The first set of projects aims to develop and validate algorithms for identification of health outcomes of interest using focused NLP tools combined with machine learning approaches to identify complex concepts such as suicidal ideation. This work will expand on prior Sentinel activities that have demonstrated the proof of concept for use of algorithmic approaches in improving outcome identification[10-12] and focus on developing a general framework for efficient and standardized processes. Another set of activities aim to build a semi-automated system of confounding adjustment that uses generalized NLP tools to enable large-scale extraction of features that appear sequentially and may serve as indicators of patients' health trajectory. Finally, initiatives are also underway to improve generalizability and transportability of NLP approaches across sites using statistical learning approaches.[13]

Strategic priority 3: Causal inference methodology

Typical Sentinel investigations for drug safety conducted using insurance claims data sources have relied on restrictive study designs such as the active comparison new user designs to achieve internal validity given availability of large underlying populations; however, use of EHR-claims linked sources may require alternate design choices, such as prevalent new user design[14] to accommodate relatively smaller underlying populations. Tradeoffs when deviating from traditional design choices to accommodate available data assets need to be thoroughly investigated. Other unique challenges when using EHR for outcome and confounder measurement include non-random missingness, or similarly selective presence of data such as medical tests that may be ordered considering the patients' prognosis, or incompleteness that occurs when outcomes recorded outside of the care systems are not available ("data leakage"). The Sentinel Innovation Center will focus on characterizing these challenges and developing strategies, methods, and tools to address them.

Residual confounding due to selection into treatment groups driven by outcome risk factors is another salient challenge in non-randomized studies. The issue is accentuated when using data that lack clinical granularity such as insurance claims availability of EHR sources that contain richer clinical information on factors not readily available in claims hold promise to improve confounding adjustment in non-randomized studies. The Sentinel Innovation Center will evaluate the feasibility of improved confounding adjustment from EHR-based variables through a combination of automated feature generation algorithms and advanced statistical and machine learning approaches such as Super Learner and Targeted Maximum Likelihood Estimation (TMLE).[15,16] Super Learner, which is an ensemble algorithm for predictive modeling, can data-adaptively model confounder summary scores, such as the propensity score, and the outcome to address model misspecification in the setting of complex and high-dimensional data settings of EHRs. TMLE can incorporate data-driven methods for high-dimensional confounder selection to empirically identify confounder information not specified by investigators. Developing scalable tools to implement these innovative methods in real time will enhance the ability of Sentinel to address the common threat of confounding. Additionally, the Innovation Center will also investigate methods such as negative control outcomes or exposures and enhance existing tools for quantitative bias analyses[17] to better understand the robustness of findings, including the impact of residual confounding.

Availability of additional clinical information in EHR further opens new opportunities to 1) identify outcomes that are generally not identifiable with claims data alone or 2) to efficiently validate claims-based algorithms in a subset of patients with EHR data available. Use of advanced methods based on machine learning and NLP to expedite outcome identification or validation have the potential to increase the efficiency of traditional drug safety evaluations. For instance, data adaptive validation techniques where human experts review batches of patient charts for endpoint validation based on claims or EHR data could be iteratively used to train algorithms to inform selection of cases that have higher likelihood of being a true positive in the next batch.[18] Incorporating NLP-assisted technology that can sift through patient charts and present the most relevant chart based on pre-specified key terms to human expert reviewers can add further efficiency to the process. The Sentinel Innovation Center will consider methodologic research to improve outcome identification and validation, which is a vital requirement for reliable evaluation of medication safety in Sentinel.

Strategic priority 4: Detection analytics

Data mining approaches such as TreeScan™ have been developed in insurance claims data for signal detection in the Sentinel infrastructure based on grouping of ICD diagnosis codes into hierarchical levels.[3] EHRs offer a potentially promising complementary source of information for medication safety signal detection but may require tailored approaches to account for and leverage differences in data content and structure compared to insurance claims. Specifically, unstructured clinical narratives recorded in EHR may provide more complete capture of subtle adverse events that may not trigger formal coding or medical interventions, aspects that are observable in claims data. The addition of detailed information presents an opportunity to expand signal detection efforts that are currently used in Sentinel. While NLP-based identification of adverse events from unstructured clinical notes is feasible with currently available methods, relational identification of newly occurring adverse events in a temporal sequence to specific medication exposures is complex and the subject of active research.[19] The Sentinel Innovation Center plans to develop a methodological framework and conduct empirical evaluations to identify and test the most promising approaches for EHR-based signal detection.

Strategic priority 5: Innovation incubator

A key goal of the Innovation Center is to increase and diversify the pathways for external investigators to engage with the Sentinel System for methods development. Comprehensive engagement with a broad scientific user community will allow the Sentinel Innovation Center to support FDA-driven relevant use cases with a variety of data products. We seek to develop a community-based approach that encourages creativity and may serve as a pipeline to new tools, approaches, and study questions for the benefit of the larger Sentinel infrastructure. To achieve the goal of broad scientific community engagement, the Innovation Center has proposed to develop a 'Data Sandbox' and engagement environment under the strategic priority of 'Innovation incubator.'

Strategic priority 6: Use cases

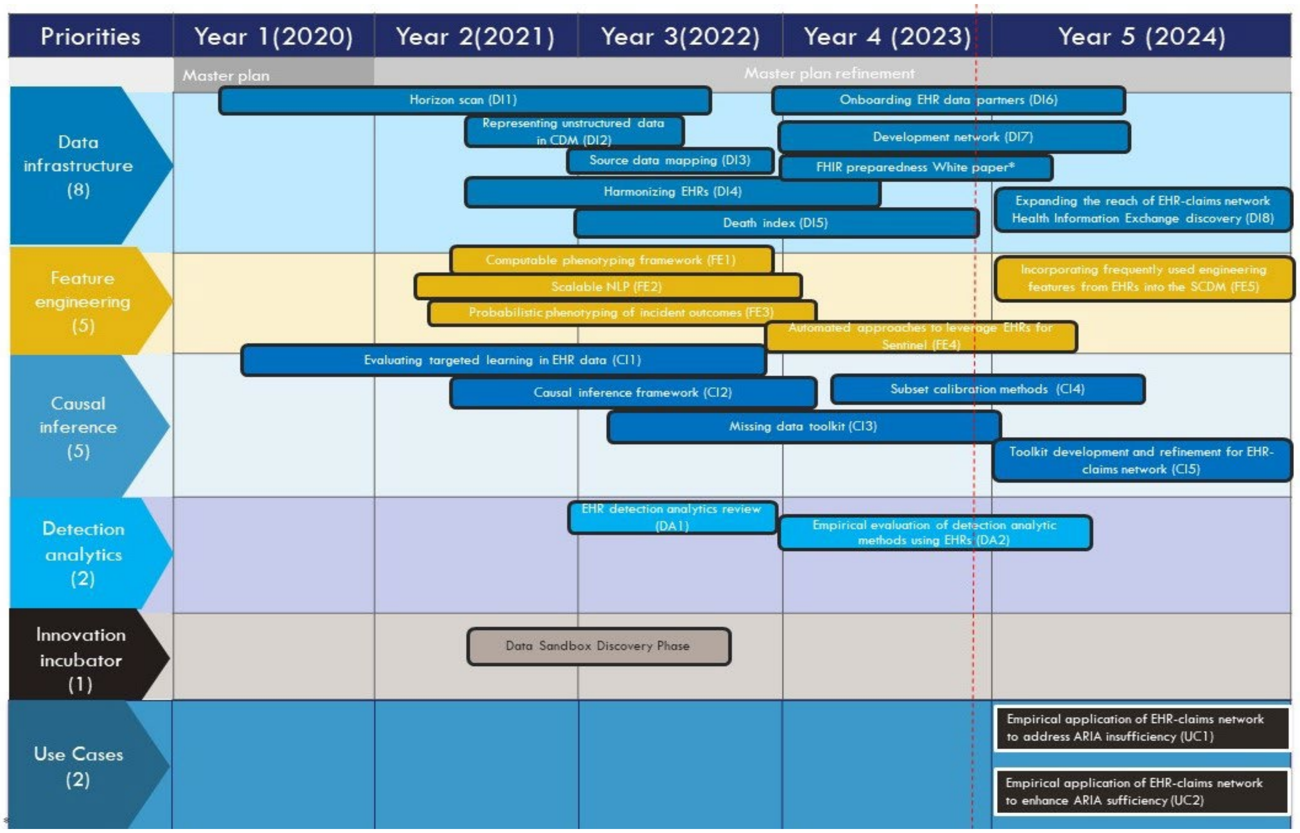
The primary goal for this priority area is to demonstrate the value of the EHR-claims data source to reduce ARIA insufficiencies and to provide further inputs for methods development as the system transitions into the production phase. Inferential requests deemed 'ARIA sufficient' are routinely conducted in the Sentinel Distributed Database to provide vital information to the FDA to aid in their decision making. However, in many circumstances, several uncertainties remain at the conclusion of an 'ARIA sufficient' analysis, which can be largely attributed to lack of data availability in insurance claims for critical variables pertaining to the research question. For instance, residual confounding by unmeasured clinical or lifestyle risk factors is often cited as a key limitation, threatening the validity of inferential analyses conducted in insurance claims data. Lack of validation for algorithms used to identify the outcome of interest is another example of frequently noted criticism for insurance claims-based studies. The primary purpose of this development network is to enable direct access to linked EHR-claims data including access to unstructured and structured information to Sentinel Investigators. Having unrestricted and timely access to granular EHR data from this development network can be critical for conducting supporting analysis to strengthen the validity of ARIA sufficient analyses conducted in the Sentinel Distributed Database using insurance claims data alone. Access to granular free-text data from EHRs is also necessary to address atypical use cases, for instance identifying use of exposures such as cannabis-derived products that are not recorded in prescribing or dispensing data sources. We will develop empirical case studies to address targeted use cases in the Innovation Center development network under this strategic priority area.

Summary

The FDA Sentinel Innovation Center has outlined a set of initiatives and a project portfolio to deliver a query-ready distributed data network containing EHR linked to claims and reusable analysis tools to enhance the capabilities of the current system. These initiatives will incorporate data science innovations, such as scalable natural language processing and machine learning, to maximize the potential of EHR for medical product safety surveillance.

Below we present the Master Plan roadmap (Figure 3), which details a timeline for the ongoing and imminent activities envisioned under the Master Plan. We will continue to refine our Master Plan based on learning from ongoing activities to eventually realize the vision of the Innovation Center of meaningfully enhancing the current Sentinel System.

Figure 3. Sentinel Innovation Center Master Plan roadmap


Priorities
Year 1(2020)
Year 2(2021)
Year 3(2022)
Year 4 (2023)
Year 5 (2024)
Master plan
Master plan refinement
Data infrastructure (8)
Horizon scan (DI1)
Representing unstructured data in CDM (DI2)
Source data mapping (DI3)
Harmonizing EHRs (DI4)
Death index (DI5)
Onboarding EHR data partners (DI6)
Development network (DI7)
FHIR preparedness White paper*
Expanding the reach of EHR-claims network Health Information Exchange discovery (DI8)
Feature engineering (5)
Computable phenotyping framework (FE1)
Scalable NLP (FE2)
Probabilistic phenotyping of incident outcomes (FE3)
Automated approaches to leverage EHRs for Sentinel (FE4)
Incorporating frequently used engineering features from EHRs into the SCDM (FE5)
Causal inference (5)
Evaluating targeted learning in EHR data (CI1)
Causal inference framework (CI2)
Missing data toolkit (CI3)
Subset calibration methods (CI4)
Toolkit development and refinement for EHR-claims network (CI5)
Detection analytics (2)
EHR detection analytics review (DA1)
Empirical evaluation of detection analytic methods using EHRs (DA2)
Innovation incubator (1)
Data Sandbox Discovery Phase
Use Cases (2)
Empirical application of EHR-claims network to address ARIA insufficiency (UC1)
Empirical application of EHR-claims network to enhance ARIA sufficiency (UC2)

References

1. Pray L, Robinson S. Challenges for the FDA: the future of drug safety. Paper presented at: Workshop summary. National Academies2007.
2. Schneeweiss S, Patorno E. Conducting real-world evidence studies on the clinical outcomes of diabetes treatments. *Endocr Rev*. 2021.
3. Wang SV, Maro JC, Baro E, et al. Data mining for adverse drug events with a propensity score matched tree-based scan statistic. *Epidemiology (Cambridge, Mass)*. 2018;29(6):895.
4. Wang SV, Maro JC, Gagne JJ, et al. A General Propensity Score for Signal Identification using Tree-Based Scan Statistics. *American Journal of Epidemiology*. 2021.
5. Nelson JC, Shortreed SM, Yu O, et al. Integrating database knowledge and epidemiological design to improve the implementation of data mining methods that evaluate vaccine safety in large healthcare databases. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2014;7(5):337-351.
6. Hernan MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American journal of epidemiology*. 2016;183(8):758-764.
7. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. *Circulation*. 2020.
8. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical epidemiology*. 2018;10:771.
9. Schneeweiss S, Brown JS, Bate A, Trifirò G, Bartels DB. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clinical Pharmacology & Therapeutics*. 2020;107(4):827-833.
10. Ball R, Toh S, Nolan J, Haynes K, Forshee R, Botsis T. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System. *Pharmacoepidemiology and drug safety*. 2018;27(10):1077-1084.
11. Bann MA, Carrell DS, Gruber S, et al. Identification and Validation of Anaphylaxis Using Electronic Health Data in a Population-based Setting. *Epidemiology*. 2021;32(3):439-443.
12. Gibson TB, Nguyen MD, Burrell T, et al. Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: Findings from a machine learning pilot project. *Journal of the American Medical Informatics Association*. 2021.
13. Shi X, Li X, Cai T. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*. 2020:1-12.
14. Suissa S, Moodie EE, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. *Pharmacoepidemiology and drug safety*. 2017;26(4):459-468.
15. Lendle SD, Fireman B, van der Laan MJ. Targeted maximum likelihood estimation in safety analysis. *Journal of clinical epidemiology*. 2013;66(8):S91-S98.
16. Wyss R, Schneeweiss S, Van Der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. 2018;29(1):96-106.
17. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. *American journal of public health*. 2016;106(7):1227-1230.
18. Collin LJ, MacLehose RF, Ahern TP, et al. Adaptive Validation Design: A Bayesian Approach to Validation Substudy Design With Prospective Data Collection. *Epidemiology (Cambridge, Mass)*. 2020;31(4):509.

19. Liu F, Jagannatha A, Yu H. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. *Drug safety*. 2019;42(1):95-97.