Sentinel Innovation Center Master Plan

Sentinel Innovation Center

Version 1.0

February 10, 2021

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	2/10/2021	Original Version	Sentinel Innovation Center

Executive summary

To achieve its Sentinel System Five-Year Strategy 2019-2023, 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 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 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 electronic health records with and accompanying methods and reusable analysis tools.

In its first year, the Innovation Center – in collaboration with FDA and in consultation with the Operations Center and the Community Building and Outreach Center – 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 2019-2023. 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 electronic health records 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 involves four key strategic priorities: (1) data infrastructure; (2) feature engineering; (3) causal inference; and (4) detection analytics. Each strategic priority has a set of goals for achieving the Innovation Center vision. The first two strategic priorities focus on establishing a query-ready distributed data network containing electronic health records and the latter two priorities focus on developing and evaluating methods and clarifying which approaches should be developed into reusable analysis tools.

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 electronic health records 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 will contribute to the success in achieving the overall vision of the Innovation Center.

Introduction

In 2019, the US Food and Drug Administration (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 2019-2023, 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.[1]

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 2019-2023, including natural language processing, advanced analytics, novel data sources, data interoperability, and emerging disruptive technologies.[1]

In its first year, the Innovation Center – in collaboration with FDA and in consultation with the Operations Center and the Community Building and Outreach Center – 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 2019-2023:[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 2019-2023.

Mission

The mission of the Sentinel Innovation Center is to improve human health by expanding Sentinel's Active Risk Identification and Analysis (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[2] 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.[3] FDA determined ARIA to be sufficient to address 113 (54%) of the product-outcome pairs.

FDA found that the inability to identify the health outcome of interest (i.e., the lack of a valid and robust computable phenotype) was the most common reason for insufficiency, although many safety issues involved multiple reasons for insufficiency. Other reasons for insufficiency included inadequate duration of follow-up for some outcome, the need for additional signal identification tools, and other methodological needs. In their analysis, FDA identified several key needs to increase the sufficiency of ARIA, including improved cause-of-death information, increased ascertainment of specific cancer outcomes, and improved pregnancy surveillance capabilities (**Table 1**). FDA also identified specific high-priority use cases that could be better addressed, at least in part, by addressing these needs. The development of the Master Plan was guided by these needs and example use cases.

Table 1. FDA-identified key needs and example use cases for the Sentinel Innovation Center

Key needs	Example use cases
Improved cause-of-death data	• Cardiovascular outcomes trials (need for capture of sudden cardiac death) • Opioid drug overdose and opioid-related death studies • 5-alpha reductase inhibitors and suicide
Improved cancer outcome data	• Immunosuppressive drugs and new onset, broad-based, cancer surveillance • Semaglutide and medullary thyroid cancer • Daclizumab and breast cancer • Burosumab_twza and tumor progression
Access to laboratory/microbiologic data	• Daclizumab (or tolvaptan, or anti-hepatitis C drugs) and drug induced liver injury • Anti-hepatitis C drugs and hepatitis B reactivation • Brodalumab and opportunistic infections
Improve pregnancy surveillance capabilities	• Pregnancy exposure registries • Pregnancy safety studies in healthcare claims databases
Inpatient/intraoperative exposure data, pediatric/ neonatal, long term follow-up, and specialized data	• Pediatric anesthesia and developmental outcomes • COVID-19 coagulopathy natural history and real-world evidence studies

Improving computable phenotyping in Sentinel

In a call for collaboration, FDA and Sentinel investigators have also outlined recommendations for improving computable phenotyping in Sentinel, and have identified the Innovation Center as a nexus for such collaboration, particularly around electronic health record data and emerging technologies to harness these data, including natural language processing and machine learning tools.[4] Addressing these areas may enhance Sentinel's ability to develop, validate, and utilize computable phenotyping algorithms more efficiently.

Some of the recommended areas to work on include:

1. Encouraging access to source data values within a common data model

   Availability of source data values will support transparency and the ability of analytic tools to make use of the source data values will maximize analytic flexibility. Because natural language processing and other technologies will be necessary to extract information from unstructured text in electronic health records, this can be accomplished by focusing on identifying, extracting, and standardizing data elements and clinical values informed by clinical contexts, rather than on extracting clinical concepts alone. For example, the data element 'abdominal pain' is relevant to diagnosing anaphylaxis, but only if it is *persistent* according to clinical diagnostic criteria.

2. Expanding efforts to standardize phenotype definitions

   When translating these data values into clinical concepts via computable phenotyping, the authors recommend expanding efforts to standardize phenotype definitions with the goal of sharing phenotypes, transparency, and reproducibility and establishing a network of learning laboratories that can quickly develop and test new computable phenotypes.

3. Develop approaches for rapid data linkage

Building capacity to rapidly link data as needed could reduce technical, operational, and governance complexity and time.

4. Developing standardized quality metrics

   Standardized, data-agnostic quality metrics could facilitate assessment of data fitness for purpose and contrasts across data sources.

Leading up to the formation of the Innovation Center, Sentinel began to address these topics in a series of projects aimed at creating a metadata table to allow re-use of validated outcomes, developing a framework for using machine learning and natural language processing to improve identification of health outcomes of interest, exploring the use of electronic health record data to better characterize key variables of interest, and assessing the feasibility of using linked electronic health record and claims data together with machine learning to validate health outcome of interest algorithms. The Master Plan strategic priority areas and goals build upon these activities.

Master Plan framework and roadmap

Below we present the Master Plan framework (**Figure 1**), which involves four key strategic priorities: (1) data infrastructure; (2) feature engineering; (3) causal inference; and (4) detection analytics. Each strategic priority has a set of goals for achieving the Innovation Center vision. The first two strategic priorities relate primarily to establishing a query-ready distributed data network containing electronic health records and the latter two priorities are concerned primarily with developing reusable analysis tools. We then present the Master Plan roadmap (**Figure 2**), which details the preliminary timeline for and sequencing of when initiatives intended to achieve the Master Plan outputs will be conducted. The roadmap links the initial outputs to the specific set of first initiatives outlined in the Master Plan framework.

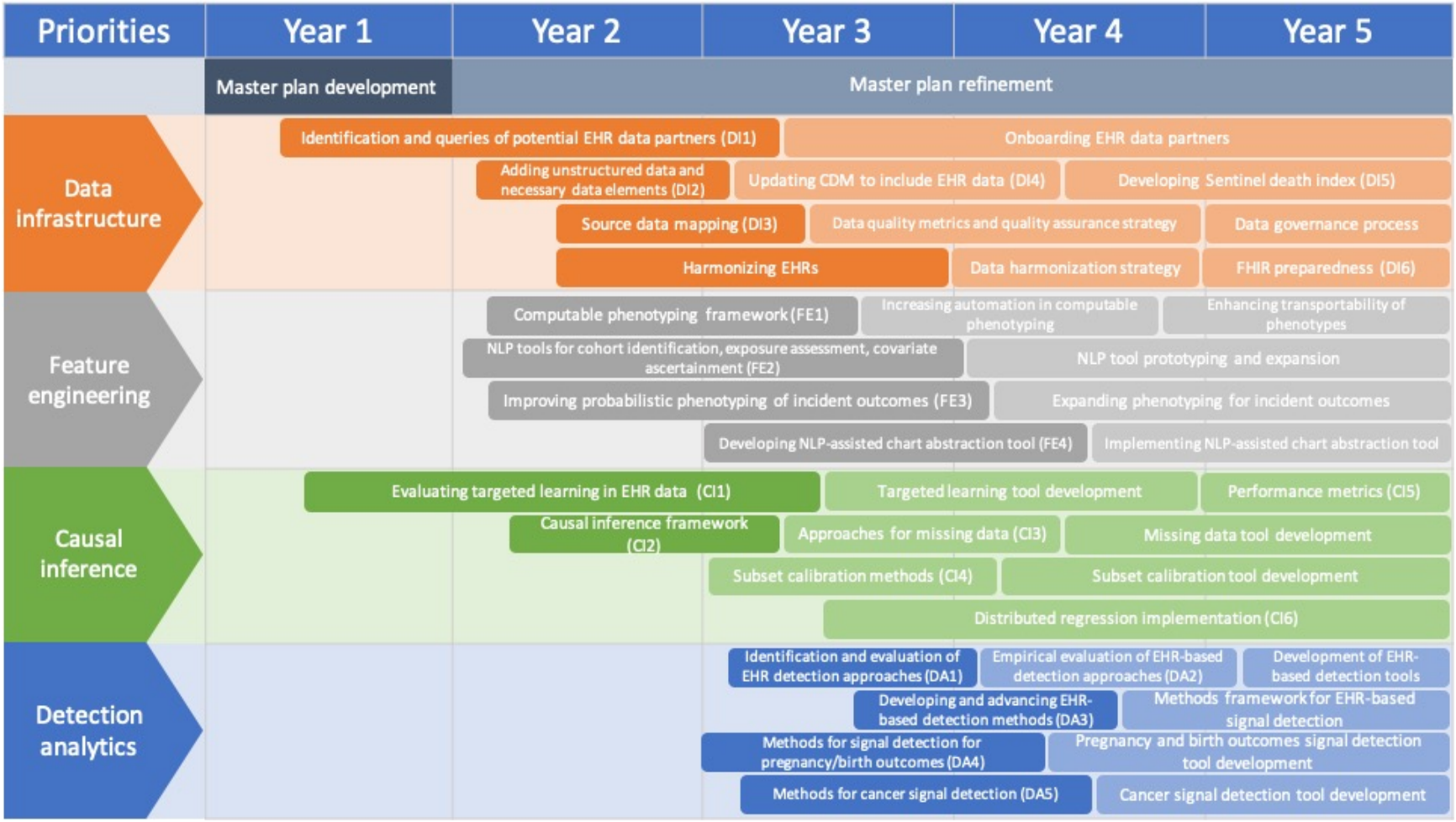
Following the figures, we provide a brief overview of the goals and outputs of each strategic priority area. To achieve these goals, the Innovation Center Master Plan Workgroup has defined a set of initiatives, or projects, aimed at generating outputs that will become the building blocks for the query-ready distributed data network containing electronic health records and the reusable analysis tools. These projects are listed in the Master Plan framework and those included in the initial set of proposed Master Plan initiatives that the Innovation Center seeks to launch are described in more detail in the Appendix. The alphanumeric codes in Figure 2 correspond to specific projects described in the Appendix.

Figure 1. Sentinel Innovation Center Master Plan framework

Priorities	Goals	Initiatives	Outputs
Data infrastructure	Establishing a Sentinel electronic health record (EHR) network requires determining where to source and how to structure the data, as well as implementation of robust governance, harmonization, and quality assurance (QA) processes.	• Horizon scan of EHR databases • Adding unstructured data to the Sentinel common data model • Identification and mitigation of source data mapping issues • Harmonizing EHRs • Updating common data model for EHR data • Developing and integrating approaches to identifying date and cause of death • FHIR implementation preparedness	• EHR data partners • Set of necessary EHR data elements • EHR common data model • Data governance process • Data harmonization and QA strategy • Data quality metrics • Sentinel death index • FHIR strategy
Feature engineering	Frameworks and tools are needed for extracting critical information from EHR data to enable and enhance EHR-based computable phenotyping and to support EHR-based descriptive, inferential, and detection queries in Sentinel.	• Extending machine learning methods development in Sentinel: follow-up analyses for anaphylaxis algorithm and formalization of a general phenotyping framework • Scalable automated natural language processing- (NLP-) assisted chart abstraction • Advancing scalable NLP approaches for unstructured EHR data • Improving probabilistic phenotyping of incident outcomes through enhanced ascertainment with NLP	• Computable phenotyping framework • NLP tools for cohort identification, exposure assessment, covariate ascertainment, and outcome identification • NLP-assisted chart abstraction tool • Chart review automation approaches • Automated feature extraction tool to improve confounding control in EHR data
Causal inference	Developing, evaluating, and implementing advanced epidemiologic and statistical methods will enable Sentinel to make best use of claims and EHR data to increase Active Risk Identification and Analysis (ARIA) sufficiency and expand the acceptance and use of real-world data for regulatory decision-making.	• Empirical evaluation of the causal inference effects of utilizing best practices for pharmacoepidemiologic studies • Enhancing causal inference in the Sentinel system: an evaluation of targeted learning and propensity scores • Approaches for handling missing laboratory data • Subset calibration for detecting and correcting for bias • Development of performance metrics and reporting standards • Advancing distributed regression in Sentinel	• Causal inference design and analysis framework • Super learner, target maximum likelihood estimation, missing data, subset calibration, and distributed regression tools • Inferential query performance metrics and reporting standards
Detection analytics	Building safety signal detection approaches for specific use cases and in EHR data, in general, will substantially enhance Sentinel's capabilities for ensuring medical product safety but requires special design and analytic methods.	• Evaluating existing approaches to EHR-based signal detection • Empirical comparison of EHR-based approaches to signal detection in Sentinel • Developing and advancing EHR-based signal detection methods • Advancing methods for safety signal detection for pregnancy and birth outcomes • Developing and evaluating a cancer signal detection tool	• Methodological framework for EHR-based signal detection • General safety signal detection tool for EHR data • Enhanced methods for signal detection for pregnancy and birth outcomes • Tool for cancer safety signal detection

ARIA, Active Risk Identification and Analysis; EHR, electronic health record; FHIR, fast healthcare interoperability standards; NLP, natural language processing; QA, quality assurance

Figure 2. Sentinel Innovation Center Master Plan roadmap



CDM, common data model; CI, causal inference; DA, detection analytics; DI, data infrastructure; FE, feature engineering; EHR, electronic health record; FHIR, fast healthcare interoperability standards; NLP, natural language processing. The alphanumeric codes correspond to specific projects described in the Appendix.

Data infrastructure

A first-order goal of the Innovation Center and of the data infrastructure priority area is to curate and establish the organizational framework for expanding access to electronic health record data and to establish the governance, harmonization, and quality assurance processes for ensuring high-fidelity, fit-for-purpose data to support FDA's Sentinel queries. In establishing a query-ready distributed data network containing electronic health records, key tasks include determining where to source the electronic health record data, defining the minimum data elements and their dimensions necessary from the electronic health records in order to address the key uses cases, determining how to organize both the structured and unstructured electronic health records data alongside claims data. As part of the Horizon scan of electronic health record databases project, the Innovation Center will engage with multiple electronic health record data providers to understand their data and capabilities and to assess their potential for addressing key uses cases and viability for being a long-term Sentinel data partner. This will determine potential partners to contribute to the distributed data network of electronic health records to meet the eventual goal of establishing a query-ready, quality-checked distributed data network containing electronic health records on at least 10 million lives.

Electronic health record data sources are heterogeneous in their content, structure, completeness, and quality. While a common data model 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 quality issues. Moreover, even though a data model standardizes the data elements, a lack of semantic interoperability across sites may remain. The Innovation Center plans to engage in multiple activities to detect and mitigate data quality issues and to develop and test approaches to harmonization across multiple electronic health record data sites.

Critical needs identified in FDA's analysis of ARIA insufficiency include improved access to cause-of-death, cancer outcomes, and long-term follow-up data. The Innovation Center will consider alternative data sources and innovative methods to address these needs. For example, the Innovation Center will evaluate the extent to which occurrence, date, and cause of death information can be triangulated from multiple sources to develop a "Sentinel death index" which might include information from publicly available sources (e.g., obituaries) as well as advanced machine learning approaches applied to electronic health record and claims data to predict cause of death.

The data infrastructure priority area also involves developing a set of data quality metrics and analyzing health information technology trends, including Fast Healthcare Interoperability Resources (FHIR) standardization and certification, to ensure that Sentinel is prepared for and well positioned within the future health information technology landscape.

Feature engineering

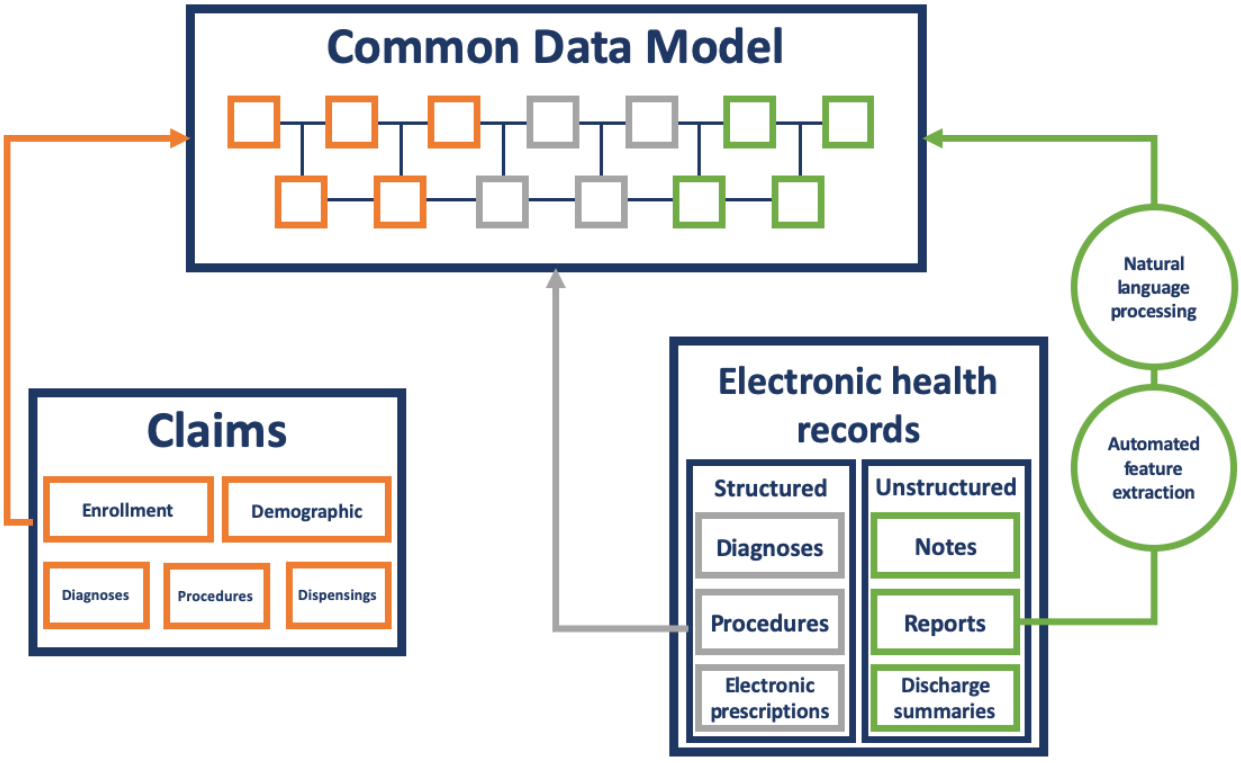
Feature engineering, which is a process to extract usable information (i.e "features") from the raw data, is critical to fully harness the power of electronic health records for safety evaluations in the Sentinel system. Whereas administrative claims, which have been the primary source of data within the Sentinel System, are highly structured, much of the potentially useful information contained within electronic health records are in the form of laboratory data, visit notes (e.g., narrative descriptions of a patient's signs and symptoms, family history, social history), radiology reports or images, and discharge summaries. This unstructured information requires substantial engineering to identify features that can be extracted and organized in the form of structured data. Natural language processing and automated feature extraction are essential mechanisms for engineering these features to support scalable computable phenotyping (i.e., the identification of health outcomes of interest in electronic health record data)[5] and automated confounding control, both of which are important for supporting both inferential and detection queries in Sentinel.

Outputs of the feature engineering priority area will involve creating a conceptual and methodological framework for computable phenotyping in Sentinel based on ongoing and new projects, advancing methods to address several key challenges to computable phenotyping in Sentinel and in the distributed data environment, and developing a set of tools for natural language processing and the automated engineering of features and algorithms to support computable phenotyping, inferential queries, and signal detect. The key methodological challenges to be addressed include:

- Developing and evaluating phenotypes in the absence of a true gold standard
- Transportability of phenotypes from one site to another (e.g., validly applying an outcome definition developed in one site to other sites)
- Increasing automation in chart review
- Making the leap from identifying prevalent conditions to incident events with automated computable phenotyping
- Incorporating unstructured EHR data into automated or semi-automated confounding control

The outputs of the data infrastructure and feature engineering priority areas serve as the building blocks for the query-ready distributed data network containing electronic health records, such as that illustrated in **Figure 3**. In particular, we envision a distributed data network containing electronic health records that utilizes an expanded Sentinel Common Data Model to organize structured data from both claims and electronic health records and to represent features extracted from the unstructured electronic health records via natural language processing including by automated feature extraction approaches.

Figure 3. Conceptual overview of integration of claims data, electronic health records, and feature engineering approaches in Sentinel



Causal inference

As part of its routine analytic framework, Sentinel has built a number of tools based on pharmacoepidemiology best practices to support inferential analyses.[6,7] Next generation pharmacoepidemiology studies will adopt the target trial paradigm in which a protocol is developed for the target trial that the observational analysis seeks to emulate.[8] Starting with the target trial paradigm, the Innovation Center will develop a causal inference design and analysis framework that will explicate and guide the role of advanced analytics and other approaches (e.g., use of positive and negative controls) in supporting future Sentinel queries in both claims and electronic health record data. Other initiatives, which will help to inform the causal inference framework, will evaluate the extent to which advanced statistical and machine learning approaches, such as target maximum likelihood estimation and super learner, can improve the validity of observational analyses, particularly in the setting of electronic health records, which contain rich clinical information and complex relations in the data that may not be fully harnessed using more traditional approaches.

While electronic health record data offer great promise for improving computable phenotyping and confounding control in Sentinel, they also bring new challenges. In particular, missing data is a critical topic that will need to be addressed when using electronic health record data in Sentinel queries to support regulatory decision-making. Methodological investigations of both design (e.g., establishing continuity cohorts[9]) and analysis (e.g., multiple imputation) approaches are needed to understand which methods are best in the setting of potentially large amounts of missing data, which will, in turn, also inform the causal inference framework that will be developed. In many cases, future Sentinel queries might be conducted primarily within a population in which full claims information is available, but in which richer clinical information from electronic health record data are available in a subset of patients. Two-stage calibration approaches, such as propensity score calibration,[10] have been proposed to use the additional information in the data-rich subgroup to correct for bias in measure of association estimated in the full claims-based population. The data-rich subgroup can also be used to evaluate whether there is evidence of imbalances between treatment groups in outcome risk factors available in the electronic health record data but not the claims data, which could be a marker of residual confounding and an indication for propensity score calibration.[11]

Another form of missing data is the lack of long follow-up information for many patients in the Sentinel distributed database as they move from one health insurance provider or integrated delivery system to another, or as they seek care in different health care settings that capture data in separate electronic health record databases. Major administrative barriers may prevent direct deterministic linkage of patient information across multiple data sources. Privacy-preserving record linkage and distributed regression and inference methods can enable continued follow-up of patients as they change insurance providers and can be used within the context of inferential analyses, but such approaches require further methodological work and evaluation to optimize these approaches in the Sentinel setting. Tools to implement the methodological approaches that are incorporated into the causal inference framework (e.g., tools to implement missing data methods, subset calibration methods, and distributed regression and inference methods) will need to be created.

Detection analytics

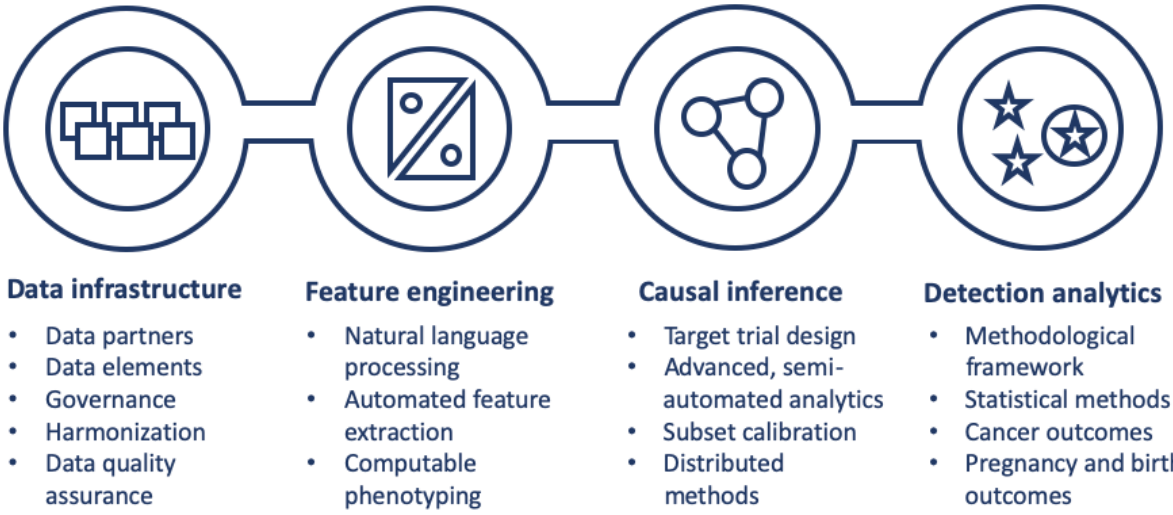
Sentinel has made important strides in advancing signal detection methods and capabilities for administrative claims data.[12] Electronic health record data offer a potentially promising complementary source of information for medical project safety signal detection but may require different signal detection approaches to account for and leverage differences in data content and structure. Specific use cases relating to signal detection for pregnancy and birth outcomes as well as for cancer outcomes require special considerations, such as focusing only on a certain set of

potential health outcomes of interest and because specific data linkages may be necessary to examine these outcomes. The Innovation Center will develop a methodological framework for electronic health record-based signal detection to address general safety use cases as well as the specific pregnancy, birth outcomes, and cancer use cases. We will conduct a horizon scan and review of emerging approaches that have been proposed or tested for signal detection in electronic health record data. Empirical evaluations will consider these and other approaches (including those that Sentinel is evaluating for claims data) to identify and test the most promising approaches, which will inform the methodological framework for electronic health record-based signal detection and be prioritized for tool development.

Innovation Center Cores

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 the respective strategic priority and contribute to the success in achieving the overall vision of the Innovation Center. **Figure 4** outlines the four Cores and the content areas that they cover.

Figure 4. Innovation Center Cores



GLOSSARY OF TERMS

Term	Definition
Active Risk Identification and Analysis (ARIA)	A component of FDA's Sentinel System comprising electronic healthcare data from Sentinel's data partners that are formatted in the Sentinel Common Data Model combined with Sentinel's parameterizable tools.
Common data model (CDM)	A standardized format in which data are stored, maintained and accessed.
Computable phenotyping	Ascertaining a condition, disease, or clinical characteristic from electronic healthcare data.
Distributed database	A network of two or more databases that reside in different locations, such as at different data partner sites.
Fast healthcare interoperability resources (FHIR)	A standard describing data formats and elements and an application programming interface for exchanging electronic health records.
Feature engineering	A process for extracting or transforming raw data into elements or fields for use in subsequent analyses.
Real world evidence (RWE)	Information obtained from real world data, which are observational data obtained outside the context of randomized controlled trials and generated during routine clinical practice.
Target trial	A hypothetical randomized trial that would answer a question of interest if resource constraints or ethics did not preclude conducting it.
US Food and Drug Administration (FDA)	A federal agency of the Department of Health and Human Service responsible for protecting and promoting public health through the control and supervision of food safety, medications, vaccines, biopharmaceuticals, and other products.

References

1. US Food and Drug Administration: Sentinel System Five-Year Strategy 2019-2023. January 2019. Available at: https://www.fda.gov/media/120333/download Accessed: September 30, 2020.
2. US Food and Drug Administration: Sentinel Common Data Model. Available at: https://www.sentinelinitiative.org/about/sentinel-common-data-model Accessed: September 30, 2020.
3. Nguyen MD, Maro JC, De Marco N, Money D, Pinheiro S, Kashoki M, Dal Pan G, Ball R. When is real world data good enough to assess drug safety? Lessons learned from the FDA's Sentinel System. *In press*
4. Brown JS, Maro JC, Nguyen M, Ball R. Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system. *J Am Med Inform Assoc* 2020;27:793-7.
5. Straub L, Gagne JJ, Maro JC, Nguyen MD, Beaulieu N, Brown JS, Kennedy A, Johnson M, Wright A, Zhou L, Wang SV. Evaluation of use of technologies to facilitate medical chart review. *Drug Saf* 2019;42:1071-80.
6. US Food and Drug Administration: Sentinel Routine Querying Tools. Available at: https://www.sentinelinitiative.org/methods-surveillance-tools/routine-querying-tools Accessed: September 30, 2020.
7. Connolly JG, Wang SV, Fuller CC, Toh S, Panozzo CA, Cocoros N, Zhou M, Gagne JJ, Maro JC. Development and application of two semi-automated tools for targeted medical product surveillance in a distributed data network. *Curr Epidemiol Rep* 2017;4:298-306.
8. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;79:70-5.
9. Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, Schneeweiss S. Identifying patients with high data completeness to improve validity of comparative effectiveness research in electronic health records data. *Clin Pharmacol Ther* 2018;103:899-905.
10. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;162:279-89.
11. Patorno E, Gopalakrishnan C, Franklin JM, Brodovicz KG, Masso-Gonzalez E, Bartels DB, Liu J, Schneeweiss S. Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical variables only observed in electronic health records. *Diabetes Obes Metab* 2018;20:974-84.
12. Wang SV, Maro JC, Baro E, Izem R, Dashevsky I, Rogers JR, Nguyen M, Gagne JJ, Patorno E, Huybrechts KF, Major JM, Zhou E, Reidy M, Cosgrove A, Schneeweiss S, Kulldorff M. Data mining for adverse drug events with a propensity score-matched tree-based scan statistic. *Epidemiology* 2018;29:895-903.

Appendix: Initial set of proposed Sentinel Innovation Center Master Plan initiatives (Ongoing projects)

This Appendix outlines the initial set of proposed Master Plan initiatives that the Innovation Center seeks to launch. These initiatives are subject to change as statements of work are developed and reviewed by FDA, which will lead to revisions to specific activity plans. FDA has ultimate decision-making authority with respect to the scopes for each of these initiatives and whether each is funded. The initial set of proposed Sentinel Innovation Center Master Plan initiatives includes:

Data infrastructure
• Horizon scan of EHR databases (DI1)
Feature engineering
• Extending machine learning methods development in Sentinel: follow-up analyses for anaphylaxis algorithm and formalization of a general phenotyping framework (FE1) • Advancing scalable NLP approaches for unstructured EHR data (FE2) • Improving probabilistic phenotyping of incident outcomes through enhanced ascertainment with NLP (FE3)
Causal inference
• Enhancing causal inference in the Sentinel system: an evaluation of targeted learning and propensity scores (CI1)

Note: non-italicized initiatives indicate initial projects that are further described in the Appendix. The preliminary set of future initiatives includes those in italics.

In addition to these initial project proposals and the preliminary set of future initiatives, the Innovation Cores will continue to identify and propose initiatives needed to generate the outputs to achieve the goals of the respective strategic priority and contribute to the success in achieving the overall vision of the Innovation Center.

Horizon scan of electronic health record databases (DI1)	
Strategic priority area	Data infrastructure
What this project adds	This project will lead to the identification of the most promising electronic record databases for incorporation into Sentinel and an understanding of the organizations' data and capabilities.

Activity background

To FDA has tasked the Innovation Center and Operations Center with establishing a query-ready distributed data network containing electronic health records for at least 10 million lives with reusable analysis tools. A key first step in establishing such a network is identifying and assessing potential partners that could contribute the necessary data for this system, including existing Data, Expansion, and Innovation Partners as well as data sources not currently included in Sentinel.

Activity description

The purpose of the Horizon scan of electronic health record databases project is to identify viable electronic health records sources for achieving a query-ready distributed data network containing electronic health records; conduct interviews with these potential partners to understand their data and capabilities for meeting the needs of this system; and conduct empirical queries of the most promising electronic health records databases to understand the data available from potential partners, to evaluate the potential partners' processes and readiness to perform the queries and to identify potential barriers to integrating the potential partners into Sentinel.

Activity plan

This project involves six key activities:

- Literature review and environmental scan to identify recently published papers that use electronic health records data;
- Interviews with representatives of the most promising data sources identified;
- Presentations by selected electronic health records representatives to the Innovation Center, Operations Center, FDA Workgroup members;
- Designing empirical queries to explore potential electronic health records data partners' data and capabilities;
- Working with the potential electronic health records data partners to implement those queries by potential data partners; and
- Reviewing the query results and the potential electronic health records data partners' processes for implementing the queries.

Extending machine learning methods development in Sentinel: follow-up analyses for anaphylaxis algorithm and formalization of a general phenotyping framework (FE1)	
Strategic priority area	Feature engineering
What this project adds	This project will further enhance methods for computable phenotyping in Sentinel, including evaluation of a full-automated phenotyping approach, and will lead to a computable phenotyping framework to guide future activities in Sentinel.

Activity background

Sentinel is currently conducting a pilot series of work to develop a framework to use machine learning and natural language processing techniques to improve health outcome of interest identification algorithms that may later be used in Sentinel. The work include two ongoing activities: (1) machine learning algorithm development for anaphylaxis; and (2) machine learning algorithm development for acute pancreatitis and multi-site adaptation for anaphylaxis algorithm. This work has led to the identification of additional important questions that need to be addressed and highlights the importance of developing a high-level computable phenotyping framework to guide future activities in Sentinel.

Activity description

The overall purpose of this activity is to continue to learn how to improve the accuracy with which we identify health outcomes of interest in Sentinel using electronic data. This activity is building on ongoing computable phenotyping work in Sentinel to:

- Expand the ongoing anaphylaxis outcome analysis plan and conduct secondary analyses;
- Develop and conduct a more scalable automated natural language processing feature engineering process (i.e., compare a PheNorm-like automated model to the current model based on manual feature curation); and
- Develop a computable phenotyping framework to provide formalized and comprehensive guidance for Sentinel.

Activity plan

The Workgroup will develop statistical analyses plans to assess generalizability across sites by training the anaphylaxis model on data from the second site and evaluating it at the first site; develop a prediction model using pooled data from the two sites; and Develop a two-stage model to reduce misclassification at one (or two) of the selected cut-points for the model. To address the second aim, the Workgroup will implement all aspects of automated algorithm development for the acute pancreatitis phenotype resulting in a final PheNorm algorithm and apply the algorithm to presumptive acute pancreatitis events to evaluate performance of the algorithm using gold standard outcome labels. In the third aim, the Workgroup will develop a comprehensive roadmap document for future electronic phenotyping work in Sentinel.

Advancing scalable natural language processing approaches for unstructured electronic health record data (FE2)

Strategic priority area	Feature engineering
What this project adds	Using COVID-19 as a use case, this project will develop scalable natural language processing tools to identify cohorts, ascertain exposures, assess covariates, and identify outcomes in electronic health record data.

Activity background

In recent years, the biomedical research community has demonstrated the benefits of marshalling unstructured clinical data to better understand disease etiology and improve care delivery. Often referred to as precision phenotyping, advanced analytic methods are being applied to structured and unstructured electronic health record data to characterize the timing, severity, and complexity of patients' health conditions. This prior work suggests that ubiquitous, unstructured electronic health record data are beneficial for generating information on populations, exposures, health outcomes of interest, and covariates, which can help achieve Sentinel's medical product safety surveillance objectives, in general, and in response to the COVID-19 pandemic, in particular.

Activity description

This project will develop approaches toward an initial framework for integrating unstructured electronic health record data into Sentinel. These approaches will be designed with multi-site interoperability and scalability in mind. The project has three specific aims:

- Identify study populations: In addition to using structured diagnosis codes and laboratory results, we will develop methods and algorithms to identify relevant COVID-19-positive patient populations from unstructured electronic health record data.
- Extract select clinical features: We will develop methods and algorithms to capture a select subset of COVID-19-relevant clinical features (exposures, health outcomes of interest, and covariates) from unstructured electronic health record data sources.
- Evaluation: To assess the benefits of utilizing unstructured data, we will compare results of using structured data only versus structured plus unstructured data for identifying study populations and capturing sample features.

Activity plan

Planned activities include: review of existing methods, identification of structured and unstructured data relevant to identifying COVID-19-positive patients; identifying and prioritizing symptoms, exposures, outcomes, and covariates of interest; developing natural language processing methods and assembling clinical corpora; creating gold standard annotation; developing algorithms; specifying strategies and requirements for scaling natural language processing tools to multiple sites, and evaluating the algorithms.

Improving probabilistic phenotyping of incident outcomes through enhanced ascertainment with natural language processing (FE3)	
Strategic priority area	Feature engineering
What this project adds	This project will test an approach to extending computable phenotyping from prevalent conditions to incident events for a specific outcome of interest and examine the generalizability of the approach to another outcome.

Activity background

A number of outcomes for which ARIA is currently insufficient exhibit features that make defining and evaluating outcomes (or phenotypes) difficult, including the need to rely predominantly on unstructured, rather than structured, electronic health record data; a lack of clearly defined reference standard; and the need for near-real-time natural language processing to identify emergent cases. Investigators at Vanderbilt University Medical Center have previously developed and implemented suicide attempt risk prediction models within electronic health record data. While these models are scalable and accurate and provide real-time estimates of risk of suicidal ideation and attempt, they currently rely on structured diagnostic coding for outcome ascertainment and identify *lifetime* suicidal ideation and suicide attempt. Work is needed to develop approaches to identify suicidality in individual notes to capture incident, rather than prevalent, suicidality. Probabilistic phenotyping may also be a solution for developing a Sentinel death index.

Activity description

The main goals of this activity are to advance computable phenotyping approaches in the setting of outcomes that rely predominantly on unstructured data, that lack a clearly defined reference standard, and that require near-real-time natural language processing to identify incident cases. The specific activity aims are to:

- Adapt natural language processing methods that identify lifetime suicidality to classify suicidality at the level of an individual clinical note (i.e., to identify incident suicidality);
- Validate the probabilistic phenotype for identifying suicide as cause of death against National Death Index records; and
- Characterize neuropsychiatric events' applicability to this approach and validate developed approach on an exemplar neuropsychiatric event.

Activity plan

The first aim will involve developing and evaluating a novel enhancement to a previously develop natural language processing approach to classify the likelihood that a note asserts positively suicidal ideation or self-harm with suicidal intent. The evaluation will be performed against medical chart review. The second aim will evaluate the performance of the natural language processing and machine learning approach to probabilistic phenotyping algorithm for correctly identifying suicide as the cause of death. The third aim will evaluate the portability of the approach to a different outcome – namely incident neuropsychiatric events.

Enhancing causal inference in the Sentinel system: an evaluation of targeted learning and propensity scores for confounding control in drug safety (CI1)	
Strategic priority area	Causal inference
What this project adds	This project will determine whether advanced machine learning methods and structured and unstructured electronic health record data can improve confounding control as compared to traditional propensity score analysis of claims data.

Activity background

Current causal inference approaches implemented in the Sentinel system include methods based on propensity scores for control of confounding. Targeted Learning, which incorporates machine learning methods for causal inference, has been proposed as an alternative and potentially better approach. The goal of this project is to evaluate the potential use of Targeted Learning methods within the Sentinel system compared to existing approaches. As Sentinel expands its data resources to include electronic health records, the project focuses on comparing the performance of Targeted Learning methods with causal inference approaches currently available to Sentinel in data environments that include linked insurance claims and electronic health record data.

Activity description

This project addresses the performance of Targeted Learning methods within linked insurance claims and electronic health record data by investigating targeted maximum likelihood estimation (TMLE) and Super Learner versus traditional propensity score methods available in Sentinel. The project has three specific aims:

- Use plasmode simulation to tailor implementation of TMLE to two use cases of interest;
- Compare the performance of propensity score-based approaches with TMLE in two empirical studies using randomized control trial results as a benchmark; and
- Compare the performance of propensity score-based approaches with TMLE in two empirical studies, where the evaluation will include confounders based on features automatically extracted from unstructured electronic health record data.

Activity plan

We will select two use cases where some ground truth is known or can be assumed; prior claims-based studies have not produced results consistent with ground truth; and data from electronic health records are expected to reduce confounding. Simulated data will be generated based on these use cases, which will then be used to tailor the implementation of TMLE. The use cases will also be implemented in actual electronic health record data to compare TMLE with existing propensity-score based approaches in Sentinel. The comparison will consider only unstructured electronic health record data (second aim) as well as features that are automatically extracted from unstructured data (third aim).