DIA

# Real-World Evidence Conference

October 16-17
Baltimore, MD

# Unexpected Issues in Pharmacoepidemiology Studies Applying Natural Language Processing to Clinical Notes

***Multi-source Observational Safety Study for Advanced Information Classification using NLP***

Sentinel Operations Center/Harvard
Cerner Enviza
John Snow Labs

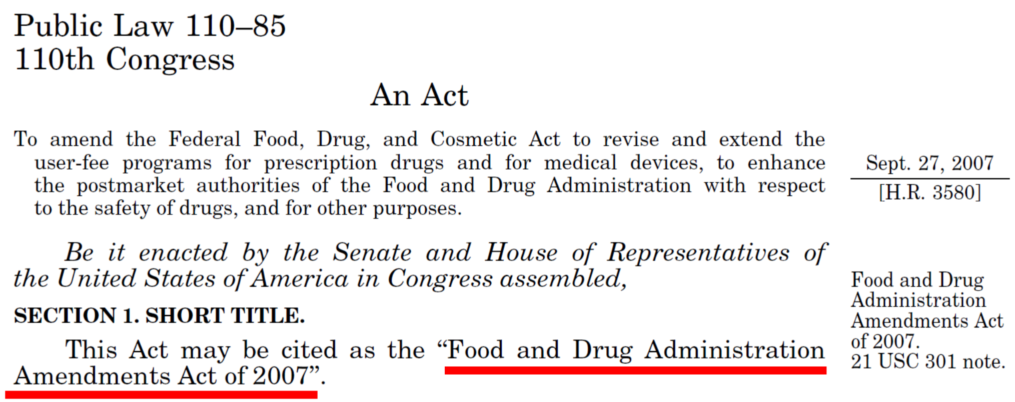
# Disclaimer

The views and opinions expressed in the following PowerPoint slides are those of the individual presenter and should not be attributed to DIA, its directors, officers, employees, volunteers, members, chapters, councils, Communities or affiliates.

This presentation is incomplete without accompanying verbal commentary.

# The FDA Sentinel System

Darren Toh, ScD
DPM Endowed Professor
Department of Population Medicine
Harvard Medical School and Harvard Pilgrim Health Care Institute



Public Law 110–85
110th Congress

An Act

To amend the Federal Food, Drug, and Cosmetic Act to revise and extend the user-fee programs for prescription drugs and for medical devices, to enhance the postmarket authorities of the Food and Drug Administration with respect to the safety of drugs, and for other purposes.

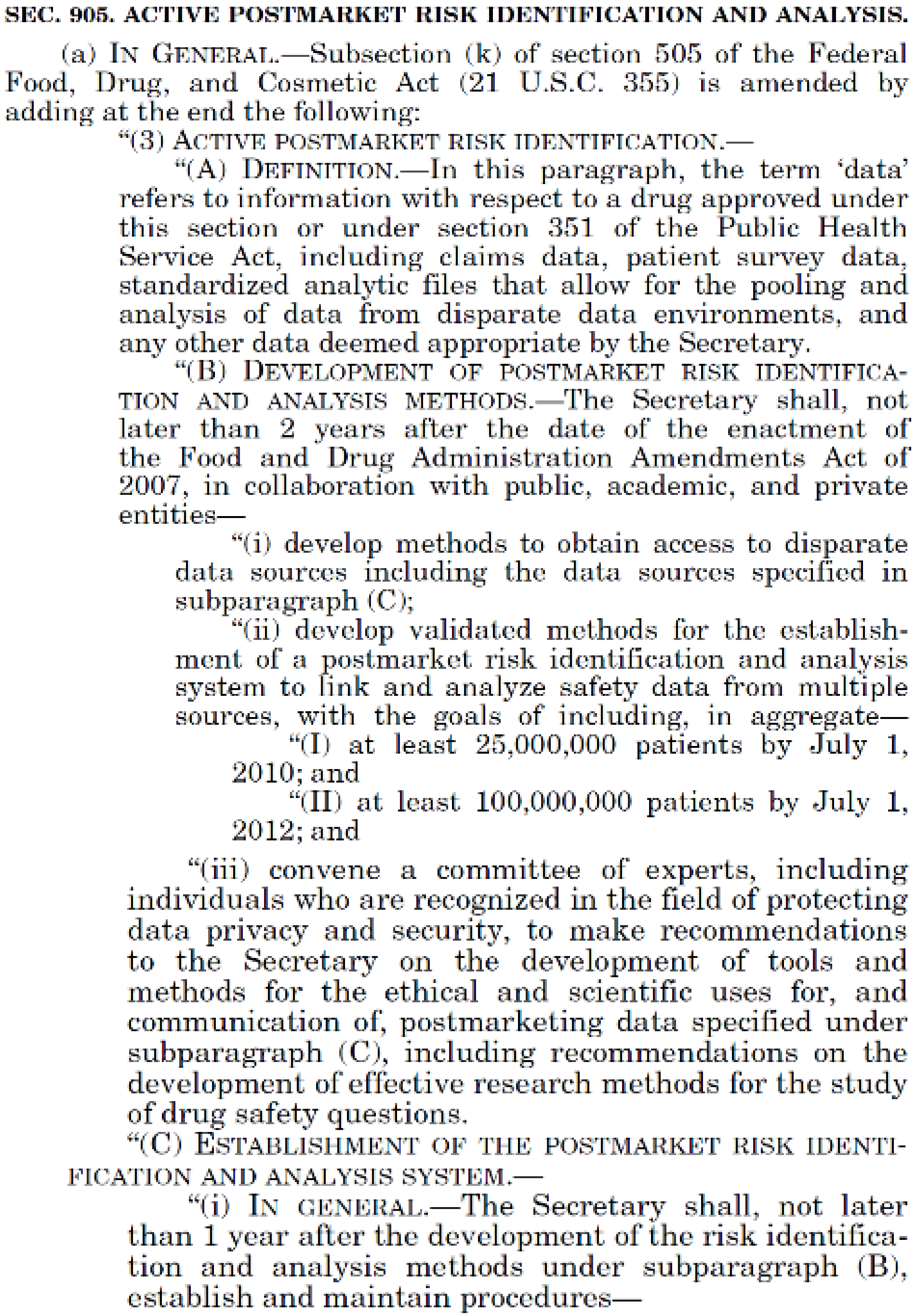
Sept. 27, 2007
[H.R. 3580]

*Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,*

Food and Drug Administration Amendments Act of 2007.
21 USC 301 note.

**SECTION 1. SHORT TITLE.**

This Act may be cited as the "Food and Drug Administration Amendments Act of 2007".



**SEC. 905. ACTIVE POSTMARKET RISK IDENTIFICATION AND ANALYSIS.**

(a) IN GENERAL.—Subsection (k) of section 505 of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 355) is amended by adding at the end the following:

"(3) ACTIVE POSTMARKET RISK IDENTIFICATION.—

"(A) DEFINITION.—In this paragraph, the term 'data' refers to information with respect to a drug approved under this section or under section 351 of the Public Health Service Act, including claims data, patient survey data, standardized analytic files that allow for the pooling and analysis of data from disparate data environments, and any other data deemed appropriate by the Secretary.

"(B) DEVELOPMENT OF POSTMARKET RISK IDENTIFICATION AND ANALYSIS METHODS.—The Secretary shall, not later than 2 years after the date of the enactment of the Food and Drug Administration Amendments Act of 2007, in collaboration with public, academic, and private entities—

"(i) develop methods to obtain access to disparate data sources including the data sources specified in subparagraph (C);

"(ii) develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources, with the goals of including, in aggregate—

"(I) at least 25,000,000 patients by July 1, 2010; and

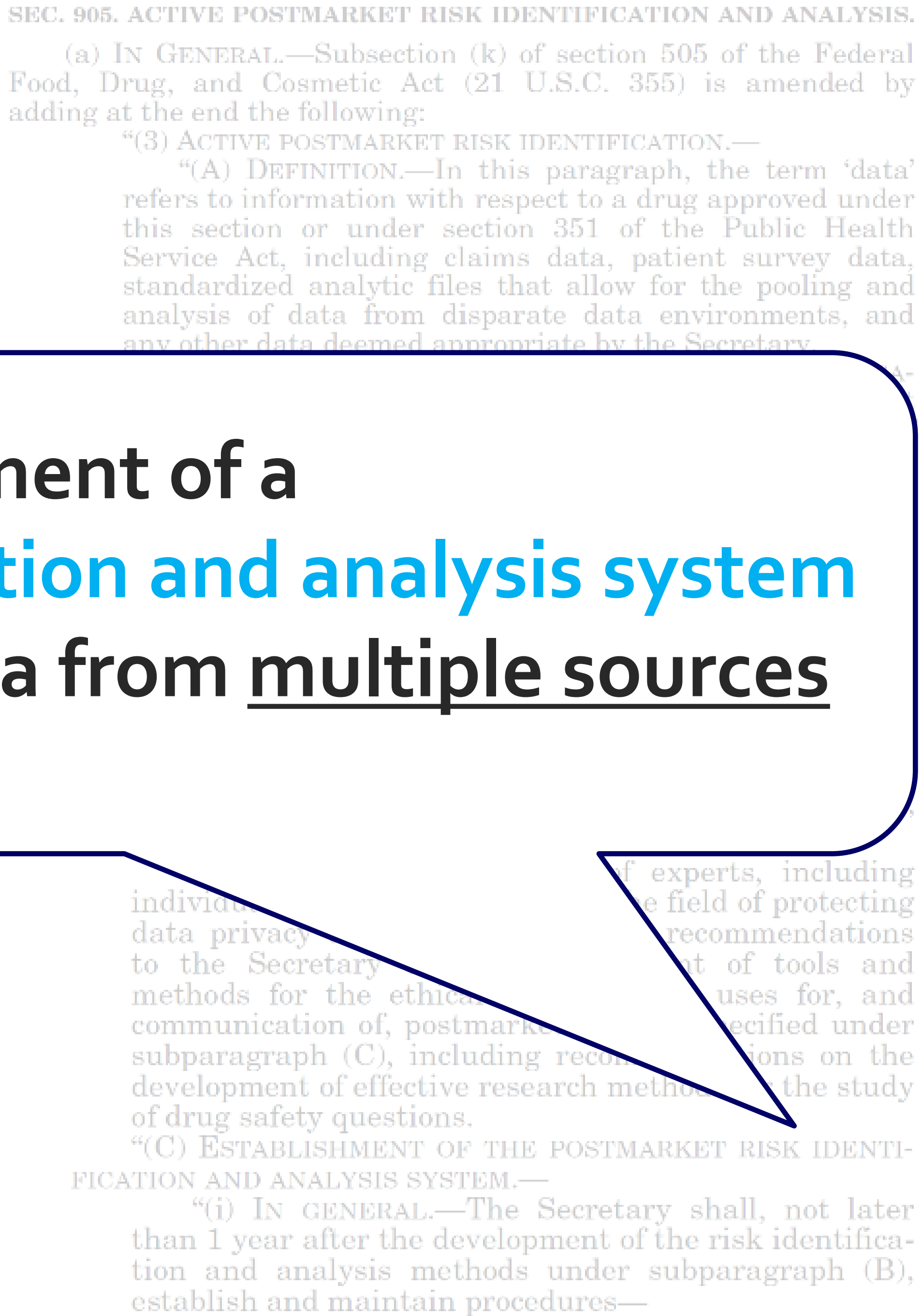
"(II) at least 100,000,000 patients by July 1, 2012; and

"(iii) convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data specified under subparagraph (C), including recommendations on the development of effective research methods for the study of drug safety questions.

"(C) ESTABLISHMENT OF THE POSTMARKET RISK IDENTIFICATION AND ANALYSIS SYSTEM.—

"(i) IN GENERAL.—The Secretary shall, not later than 1 year after the development of the risk identification and analysis methods under subparagraph (B), establish and maintain procedures—

https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf



SEC. 905. ACTIVE POSTMARKET RISK IDENTIFICATION AND ANALYSIS.

(a) IN GENERAL.—Subsection (k) of section 505 of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 355) is amended by adding at the end the following:

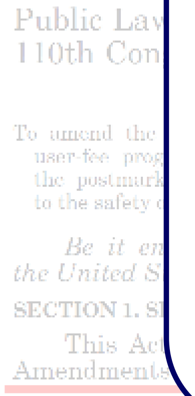
"(3) ACTIVE POSTMARKET RISK IDENTIFICATION.—

"(A) DEFINITION.—In this paragraph, the term 'data' refers to information with respect to a drug approved under this section or under section 351 of the Public Health Service Act, including claims data, patient survey data, standardized analytic files that allow for the pooling and analysis of data from disparate data environments, and any other data deemed appropriate by the Secretary.

f experts, including
individu
e field of protecting
data privacy
recommendations
to the Secretary
t of tools and
methods for the ethica
uses for, and
communication of, postmarke
ecified under
subparagraph (C), including reco
ions on the
development of effective research metho
the study
of drug safety questions.

"(C) ESTABLISHMENT OF THE POSTMARKET RISK IDENTIFICATION AND ANALYSIS SYSTEM.—

"(i) IN GENERAL.—The Secretary shall, not later than 1 year after the development of the risk identification and analysis methods under subparagraph (B), establish and maintain procedures—



Public Law
110th Con

To amend the
user-fee prog
the postmark
to the safety

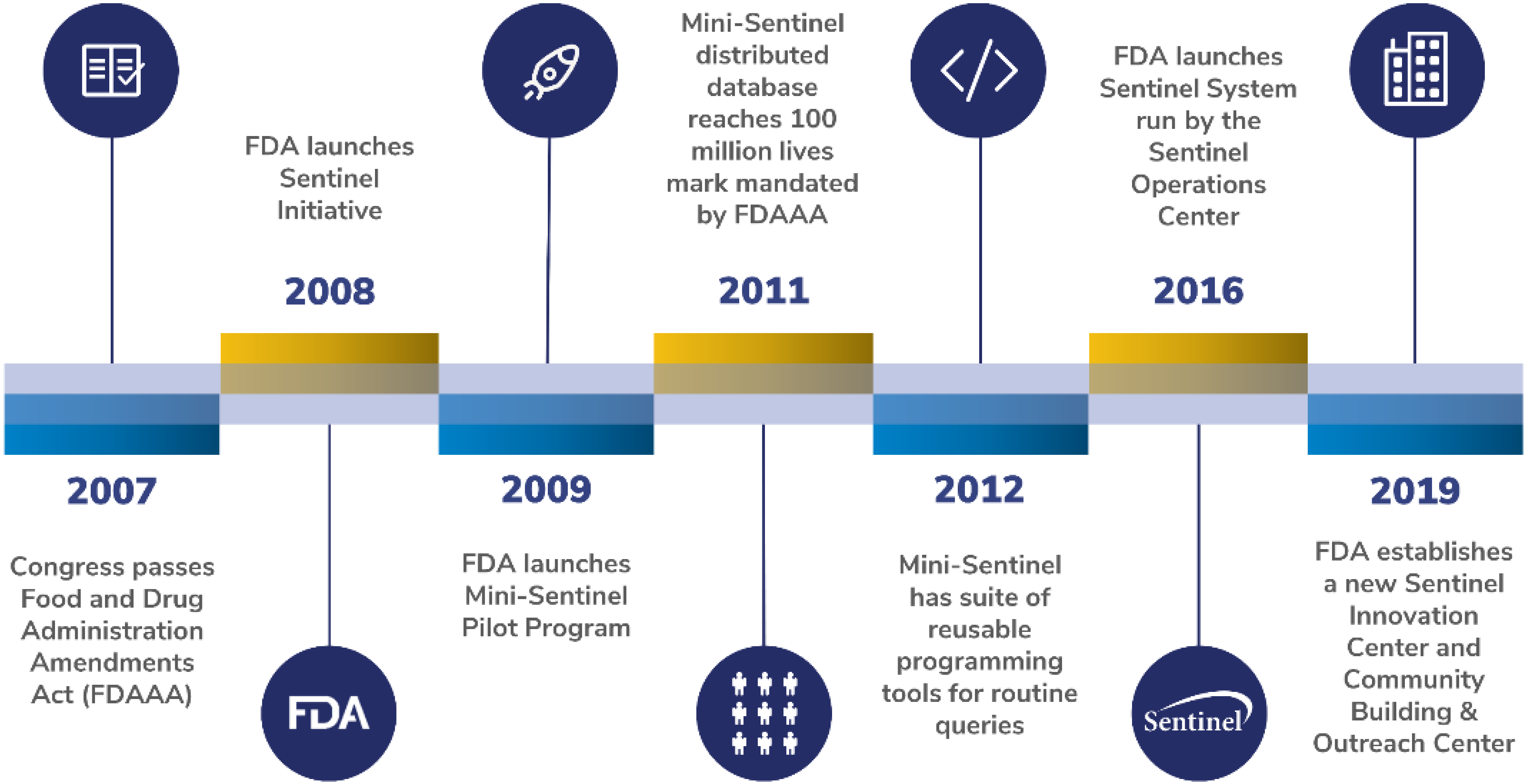
*Be it en*
*the United S*

SECTION 1. S

This Act
Amendments

## Establishment of a postmarket risk identification and analysis system to link analyze safety data from multiple sources

https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf



**2007**

Congress passes Food and Drug Administration Amendments Act (FDAAA)

**2008**

FDA launches Sentinel Initiative

**2009**

FDA launches Mini-Sentinel Pilot Program

**2011**

Mini-Sentinel distributed database reaches 100 million lives mark mandated by FDAAA

**2012**

Mini-Sentinel has suite of reusable programming tools for routine queries

**2016**

FDA launches Sentinel System run by the Sentinel Operations Center

**2019**

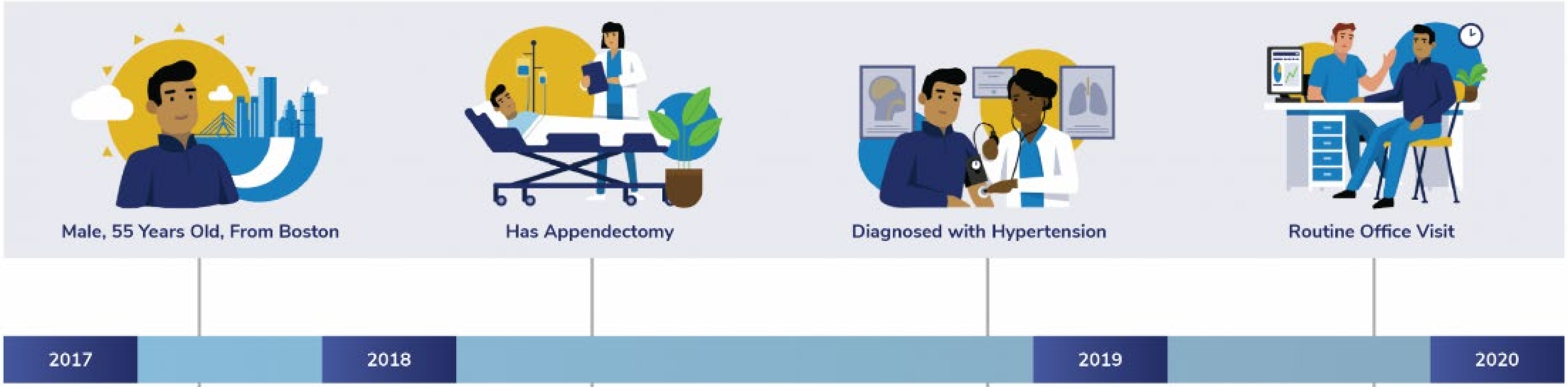
FDA establishes a new Sentinel Innovation Center and Community Building & Outreach Center

https://www.sentinelinitiative.org/about

# Collaborating Institutions

**Sentinel Operations Center Lead:** Harvard Pilgrim Health Care Institute

- Brigham and Women's Hospital: Division of Pharmacoepidemiology & Pharmacoeconomics in the Department of Medicine
- Carelon Research/Elevance Health
- CVS Health (Aetna)
- Duke University School of Medicine, Department of Population Health Sciences (Medicare Fee-for-Service and Medicaid data)
- Harvard T.H. Chan School of Public Health
- HCA Healthcare
- Health Partners Institute
- HealthVerity
- Humana Healthcare Research
- Kaiser Permanente Colorado
- Kaiser Permanente Hawaii
- Kaiser Permanente Mid-Atlantic
- Kaiser Permanente Northwest
- Kaiser Permanente Washington
- Marshfield Clinic Research Institute
- Merative
- Meyers Health Care Institute
- Optum
- TriNetX
- University of Florida College of Pharmacy, Department of Pharmaceutical Outcomes and Policy
- University of North Carolina Gillings School of Global Public Health
- University of Pennsylvania Perelman School of Medicine, Center for Clinical Epidemiology and Biostatistics
- University of Washington School of Public Health
- Vanderbilt University Medical Center (Tennessee Medicaid data)



Male, 55 Years Old, From Boston

1/1/2017

**Encounter**

**Office Visit Diagnosis:**
Influenza with pneumonia

**Dispensings**

**Prescription:**
Antibiotic

Has Appendectomy

3/15/2018

**Encounters**

**Emergency Department Procedure:**
Appendectomy

3/15/2018 - 3/18/2018

**Hospital:**
Inpatient Stay

Diagnosed with Hypertension

12/11/2018

**Encounter**

**Office Visit Diagnosis:**
Hypertension

**Dispensings**

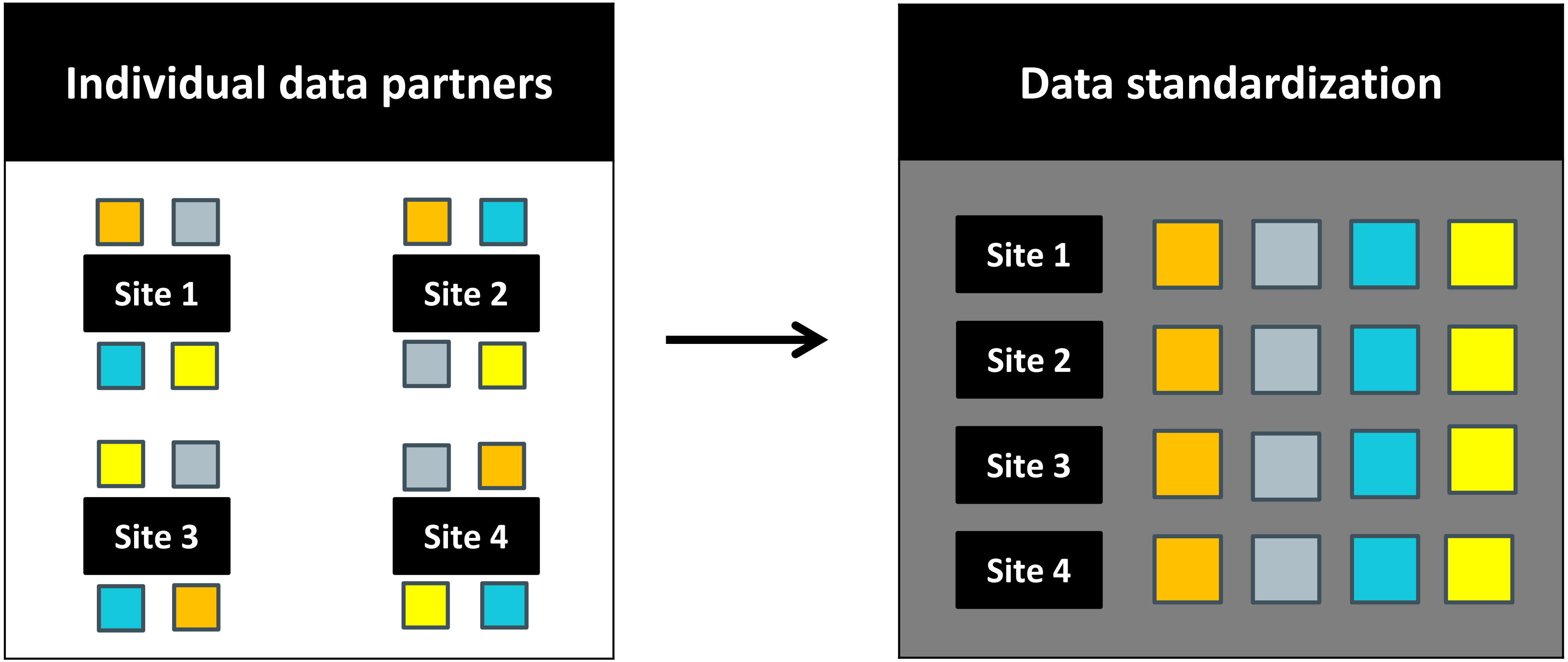
**Prescription:**
Anti-hypertensive

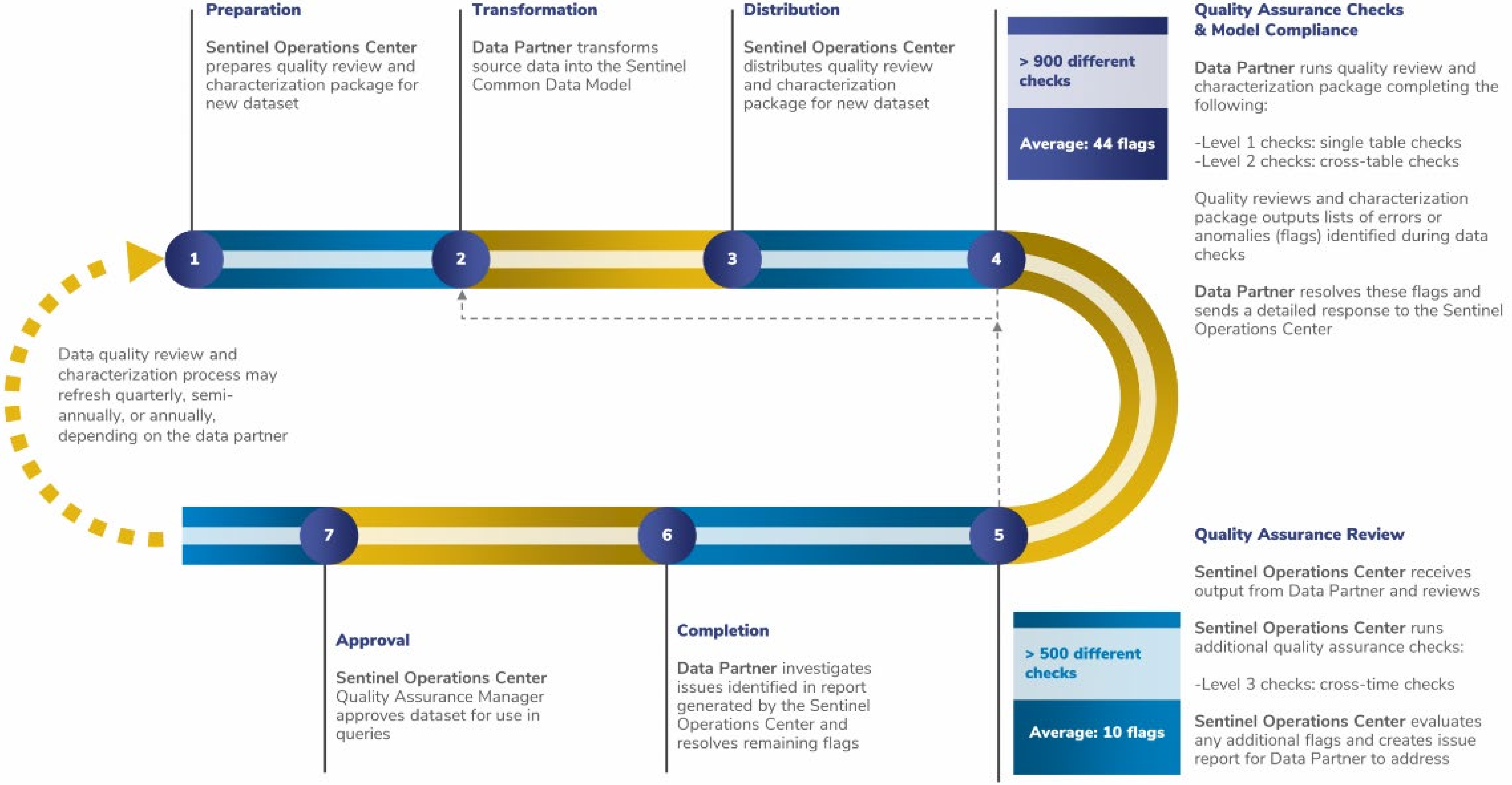
Routine Office Visit

10/31/2019

**Encounter**

**Office Visit Diagnosis:**
Hypertension


Individual data partners
Site 1
Site 2
Site 3
Site 4
Data standardization
Site 1
Site 2
Site 3
Site 4



**Preparation**

**Sentinel Operations Center** prepares quality review and characterization package for new dataset

**Transformation**

**Data Partner** transforms source data into the Sentinel Common Data Model

**Distribution**

**Sentinel Operations Center** distributes quality review and characterization package for new dataset

**> 900 different checks**

**Average: 44 flags**

**Quality Assurance Checks & Model Compliance**

**Data Partner** runs quality review and characterization package completing the following:

-Level 1 checks: single table checks
-Level 2 checks: cross-table checks

Quality reviews and characterization package outputs lists of errors or anomalies (flags) identified during data checks

**Data Partner** resolves these flags and sends a detailed response to the Sentinel Operations Center

**Quality Assurance Review**

**Sentinel Operations Center** receives output from Data Partner and reviews

**Sentinel Operations Center** runs additional quality assurance checks:

-Level 3 checks: cross-time checks

**Sentinel Operations Center** evaluates any additional flags and creates issue report for Data Partner to address

**> 500 different checks**

**Average: 10 flags**

**Completion**

**Data Partner** investigates issues identified in report generated by the Sentinel Operations Center and resolves remaining flags

**Approval**

**Sentinel Operations Center** Quality Assurance Manager approves dataset for use in queries

Data quality review and characterization process may refresh quarterly, semi-annually, or annually, depending on the data partner

# Types of Data Quality Checks and Examples

**Level 1 Checks:**
Single table checks

✓ **Completeness**
Admission date is not missing value

✓ **Validity**
Admission date is in date format

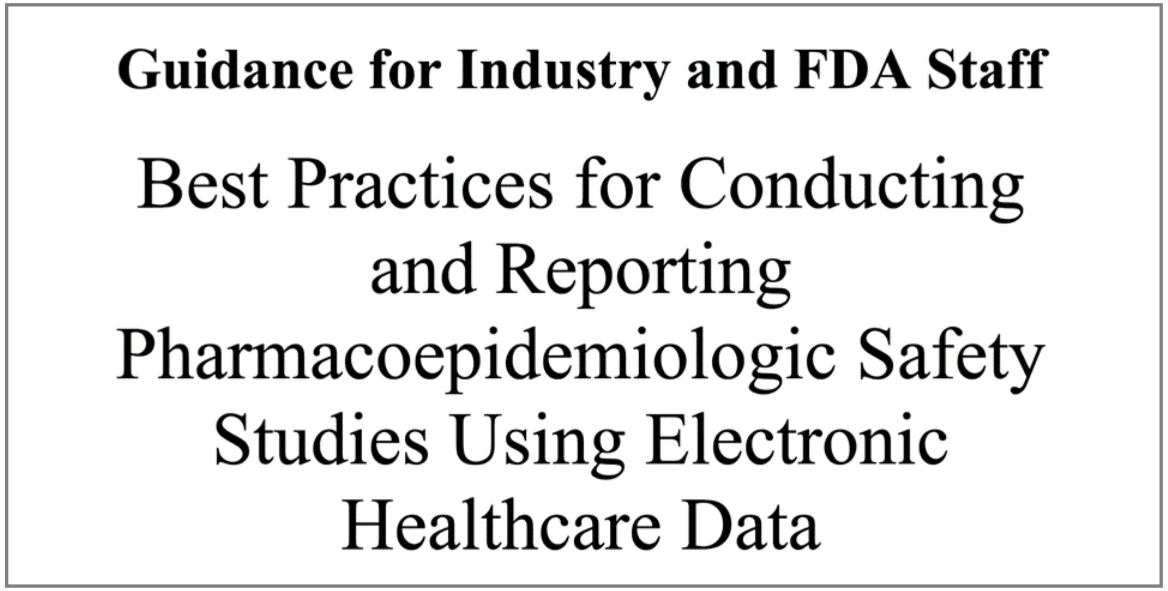
**Level 2 Checks:**
Cross-table checks

✓ **Accuracy**
Admission date occurs before the patient's discharge

✓ **Integrity**
Admission date occurs within the patient's active enrollment period

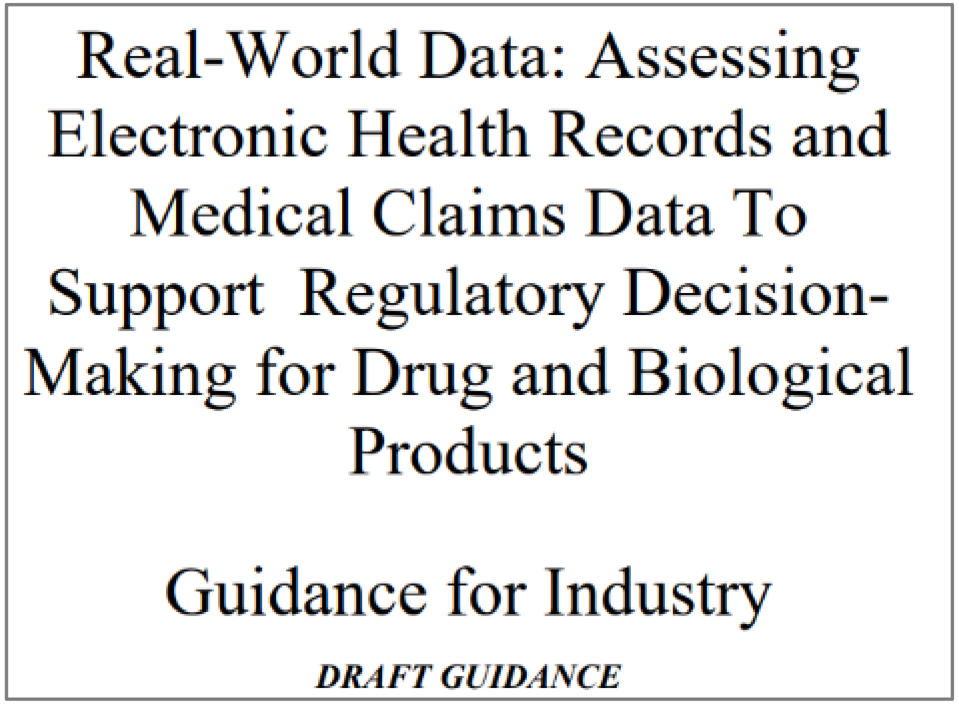
**Level 3 Checks:**
Cross-time checks

✓ **Consistency of Trends**
There is no sizable percent change in admission date record counts by month-year



**Guidance for Industry and FDA Staff**

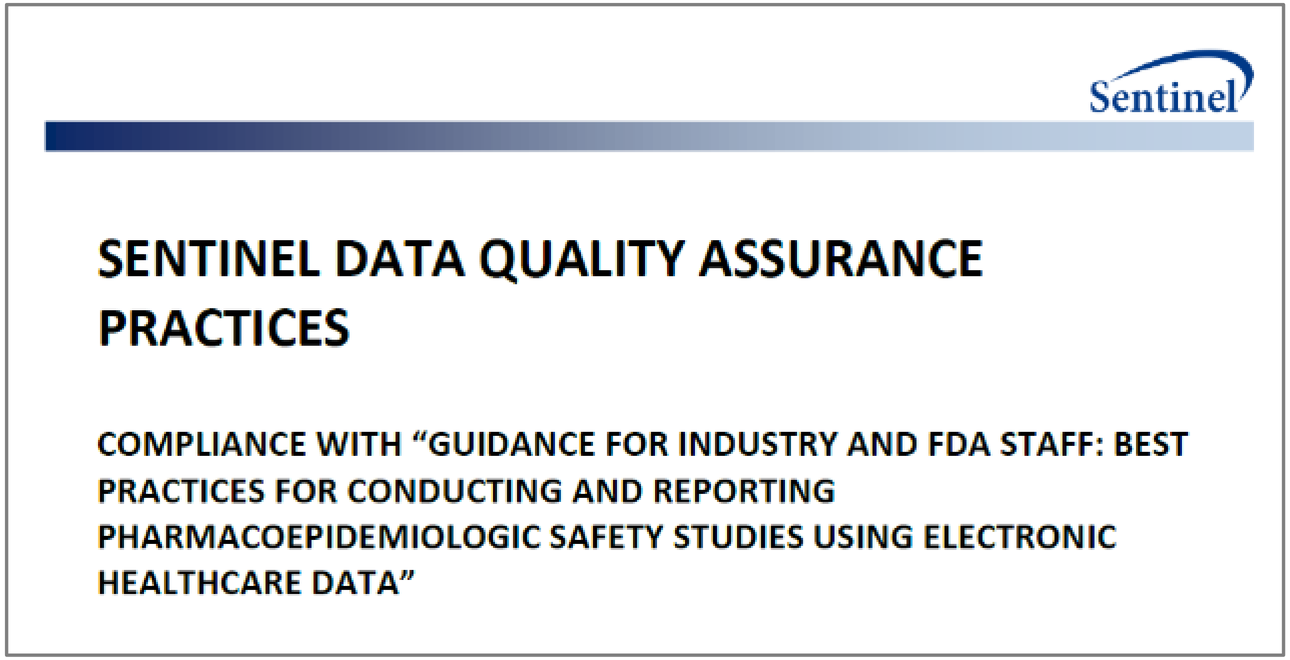
Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data



Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

***DRAFT GUIDANCE***



**SENTINEL DATA QUALITY ASSURANCE PRACTICES**

COMPLIANCE WITH "GUIDANCE FOR INDUSTRY AND FDA STAFF: BEST PRACTICES FOR CONDUCTING AND REPORTING PHARMACOEPIDEMIOLOGIC SAFETY STUDIES USING ELECTRONIC HEALTHCARE DATA"

https://www.fda.gov/downloads/drugs/guidances/ucm243537.pdf
https://www.fda.gov/media/152503/download
https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf

# Sentinel Common Data Model

## Administrative Data

| Enrollment | Demographic | Dispensing | Encounter | Diagnosis | Procedure | Prescribing |
|---|---|---|---|---|---|---|
| Patient ID | Patient ID | Patient ID | Patient ID | Patient ID | Patient ID | Patient ID |
| Enrollment Start & End Dates | Birth Date | Provider ID | Encounter ID & Type | Encounter ID & Type | Encounter ID & Type | Encounter ID |
| Medical Coverage | Sex | Dispensing Date | Service Date(s) | Provider ID | Provider ID | Provider ID |
| Drug Coverage | Postal Code | Rx | Facility ID | Service Date(s) | Service Date(s) | Order Date |
| Medical Record Availability | Race | Rx Code Type | Etc. | Diagnosis Code & Type | Procedure Code & Type | Rx |
|  | Etc. | Days Supply |  | Principal Discharge Diagnosis | Etc. | Days Supply |
|  |  | Amount Dispensed |  |  |  | Rx Route of Delivery |
|  |  |  |  |  |  | Etc. |

## Mother-Infant Linkage Data

| Mother-Infant Linkage |
|---|
| Mother ID |
| Mother Birth Date |
| Encounter ID & Type |
| Mother Admission & Discharge Date |
| Child ID |
| Childbirth Date |
| Mother-Infant Match Method |
| Etc. |

## Auxiliary Data

| Facility | Provider |
|---|---|
| Facility ID | Provider ID |
| Facility Location | Provider Specialty & Specialty Code Type |

## Registry Data

| Death | Cause of Death | State Vaccine* |
|---|---|---|
| Patient ID | Patient ID | Patient ID |
| Death Date | Cause of Death | Vaccination Date |
| Date Imputed Flag | Source | Admission Date |
| Source | Confidence | Vaccine Code & Type |
| Confidence | Etc. | Provider |
| Etc. |  | Etc. |

## Inpatient Data

| Inpatient Pharmacy | Inpatient Transfusion |
|---|---|
| Patient ID | Patient ID |
| Encounter ID | Encounter ID |
| Rx Administration Date & Time | Transfusion Administration ID |
| National Drug Code (NDC) | Administration Start & End Date & Time |
| Rx ID | Transfusion Product Code |
| Route | Blood Type |
| Dose | Etc. |
| Etc. |  |

## Clinical Data

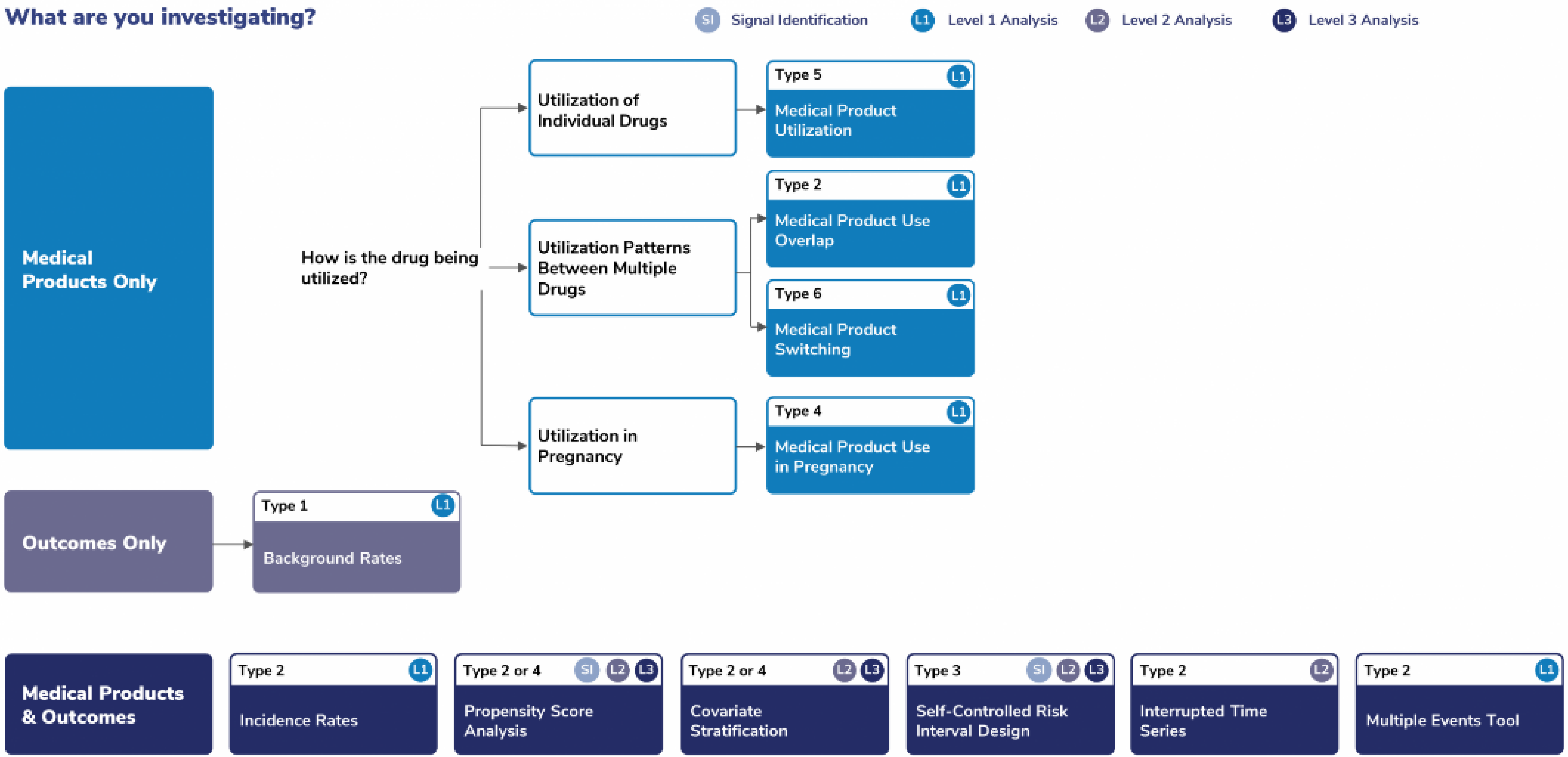
| Lab Result | Vital Signs |
|---|---|
| Patient ID | Patient ID |
| Result & Specimen Collection Dates | Measurement Date & Time |
| Test Type, Immediacy & Location | Height & Weight |
| Logical Observation Identifiers Names and Codes (LOINC®) | Diastolic & Systolic BP |
| Etc. | Tobacco Use & Type |
|  | Etc. |

## Patient-Reported Measures (PRM) Data

| PRM Survey | PRM Survey Response |
|---|---|
| Measure ID | Patient ID |
| Survey ID | Encounter ID |
| Question ID | Measure ID |
| Etc. | Survey ID |
|  | Question ID |
|  | Response Text |
|  | Etc. |

**The State Vaccine table has not been in use since SCDM v6.0.*

## What are you investigating?



SI Signal Identification L1 Level 1 Analysis L2 Level 2 Analysis L3 Level 3 Analysis

Medical Products Only

How is the drug being utilized?

Utilization of Individual Drugs

Type 5 L1
Medical Product Utilization

Utilization Patterns Between Multiple Drugs

Type 2 L1
Medical Product Use Overlap

Type 6 L1
Medical Product Switching

Utilization in Pregnancy

Type 4 L1
Medical Product Use in Pregnancy

Outcomes Only

Type 1 L1
Background Rates

Medical Products & Outcomes

Type 2 L1
Incidence Rates

Type 2 or 4 SI L2 L3
Propensity Score Analysis

Type 2 or 4 L2 L3
Covariate Stratification

Type 3 SI L2 L3
Self-Controlled Risk Interval Design

Type 2 L2
Interrupted Time Series

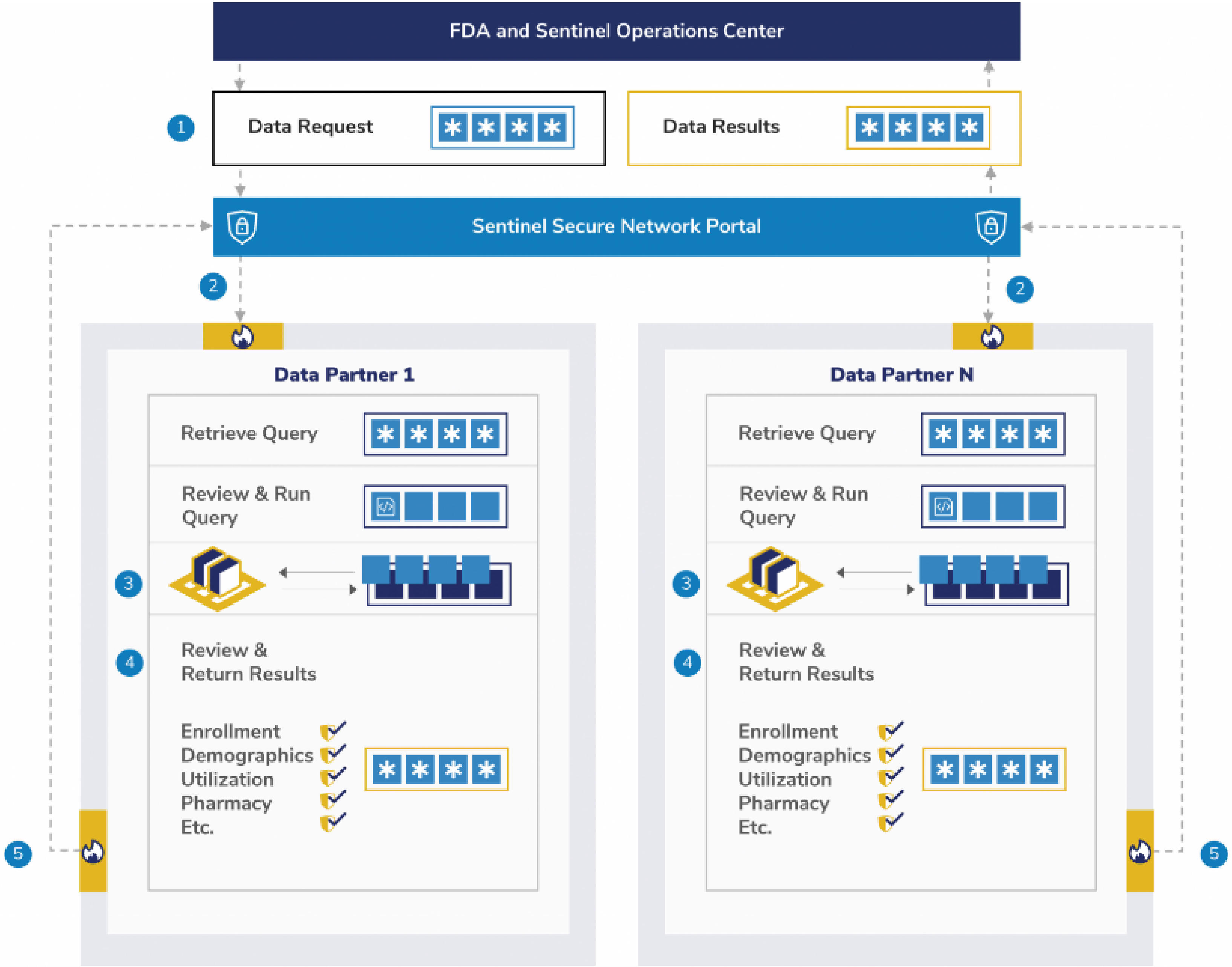
Type 2 L1
Multiple Events Tool

1. FDA data request sent to Data Partners via FISMA-compliant secure network portal
2. Data Partners retrieve query
3. Data Partners review and run query against their local data behind their firewalls
4. Data Partners review results for accuracy and privacy compliance
5. Data Partners return de-identified results to SOC via secure portal

Firewall

Local Data

Privacy Compliance



**463 million**
unique patient identifiers (2000-2023)

**1.1 billion**
person-years of data*

**113 million**
members currently accruing data*

**20 billion**
pharmacy dispensing*

**20 billion**
medical encounters*

**8 million**
deliveries with mom-baby linkage

* Among individuals with both medical and drug coverage

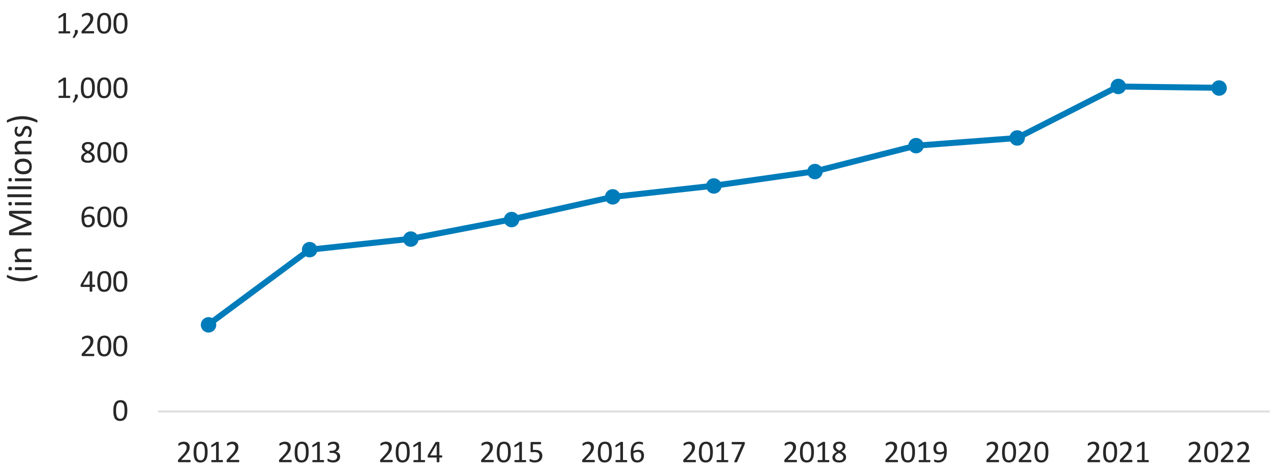
| Table | DP Count | Member Count | Record Count |
|---|---:|---:|---:|
| Laboratory Results | 11 | 99,358,668 | 8,857,509,772 |
| Vital Signs | 7 | 10,636,075 | 368,812,494 |
| Prescribing | 3 | 3,271,299 | 162,101,760 |

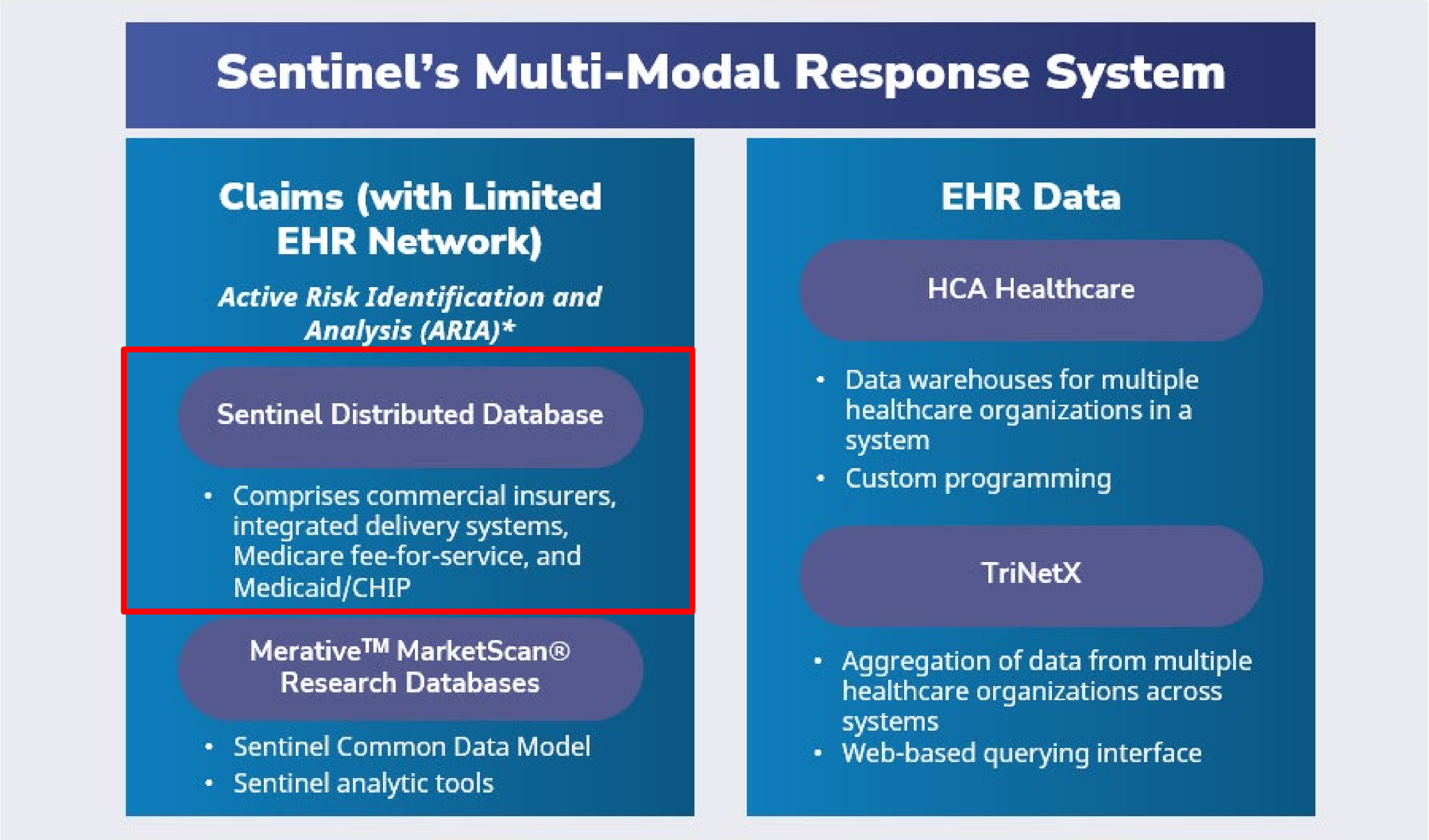
**Members with Medical and Drug Coverage who Have at least One Vital Sign Measurement, by Vital Sign Measure**

| Vital Sign | Member Count |
|---|---:|
| Diastolic Blood Pressure | 6,253,679 |
| Systolic Blood Pressure | 6,254,628 |
| Weight | 6,416,934 |
| Height | 5,942,271 |

## Growth in Laboratory Result Data By Year

Total Laboratory Result Records





## Sentinel's Multi-Modal Response System

### Claims (with Limited EHR Network)

*Active Risk Identification and Analysis (ARIA)**

**Sentinel Distributed Database**

- Comprises commercial insurers, integrated delivery systems, Medicare fee-for-service, and Medicaid/CHIP

**Merative™ MarketScan® Research Databases**

- Sentinel Common Data Model
- Sentinel analytic tools

### EHR Data

**HCA Healthcare**

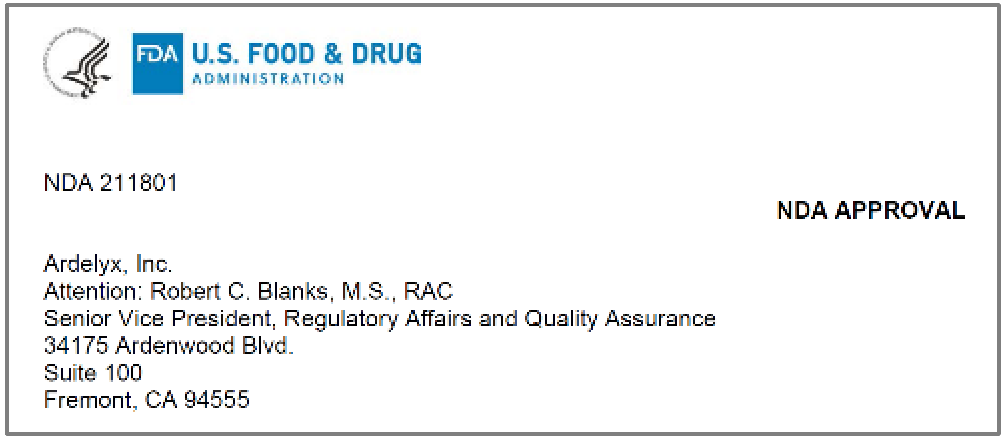
- Data warehouses for multiple healthcare organizations in a system
- Custom programming

**TriNetX**

- Aggregation of data from multiple healthcare organizations across systems
- Web-based querying interface

**Note*: The Active Risk Identification and Analysis (ARIA) System is comprised of the Sentinel Distributed Database, the Sentinel Common Data Model, and Sentinel analytic tools.*

# Conduct studies for safety concerns that arise during the review of an application for a new drug or biologic

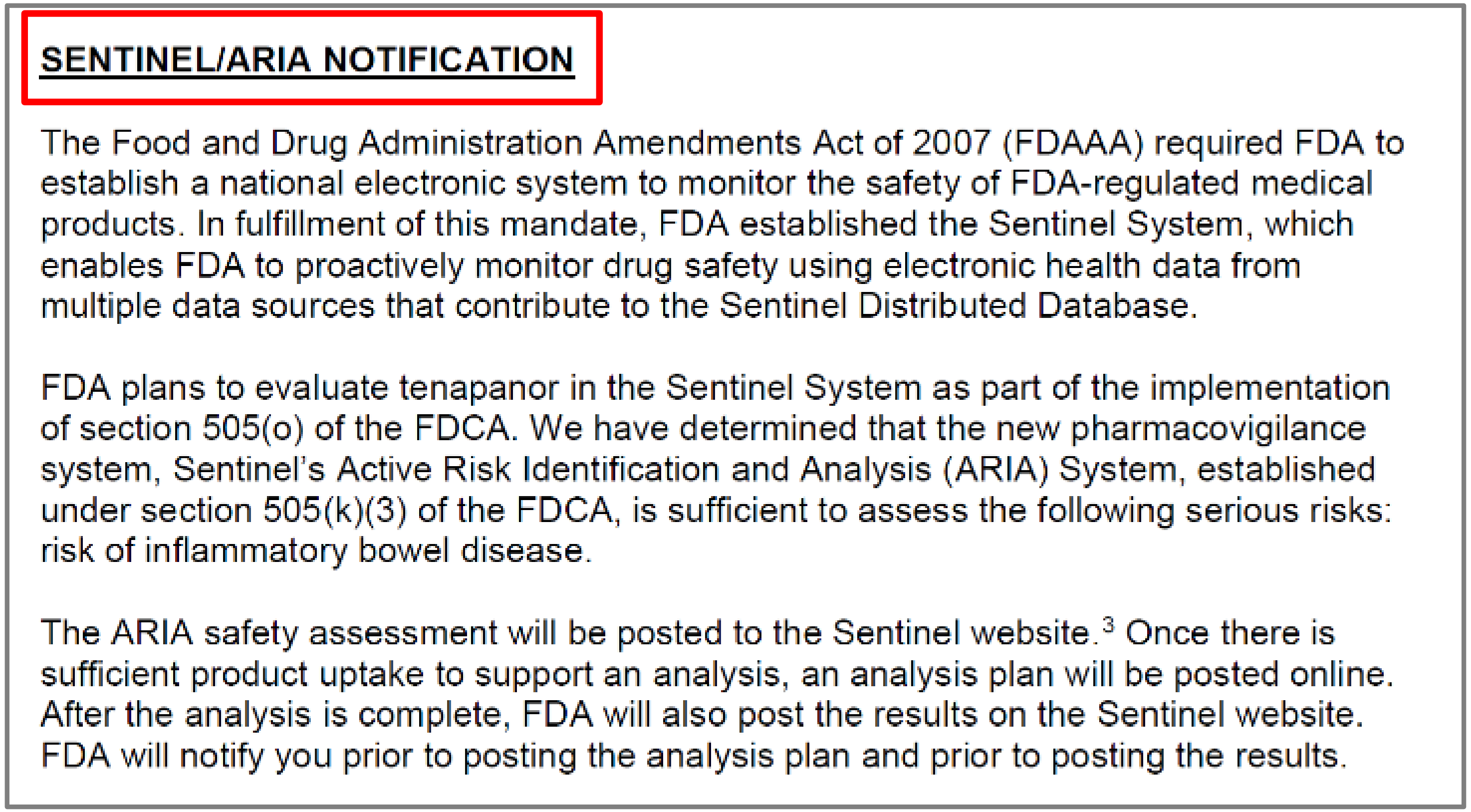


FDA U.S. FOOD & DRUG ADMINISTRATION

NDA 211801

**NDA APPROVAL**

Ardelyx, Inc.
Attention: Robert C. Blanks, M.S., RAC
Senior Vice President, Regulatory Affairs and Quality Assurance
34175 Ardenwood Blvd.
Suite 100
Fremont, CA 94555



**SENTINEL/ARIA NOTIFICATION**

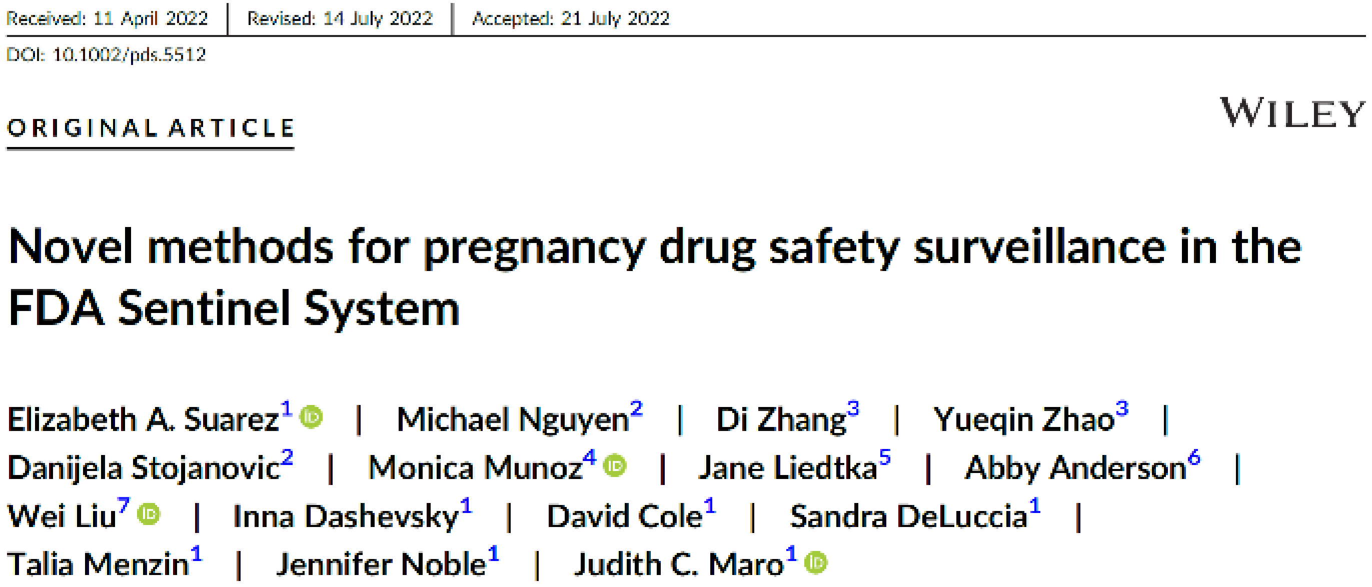
The Food and Drug Administration Amendments Act of 2007 (FDAAA) required FDA to establish a national electronic system to monitor the safety of FDA-regulated medical products. In fulfillment of this mandate, FDA established the Sentinel System, which enables FDA to proactively monitor drug safety using electronic health data from multiple data sources that contribute to the Sentinel Distributed Database.

FDA plans to evaluate tenapanor in the Sentinel System as part of the implementation of section 505(o) of the FDCA. We have determined that the new pharmacovigilance system, Sentinel's Active Risk Identification and Analysis (ARIA) System, established under section 505(k)(3) of the FDCA, is sufficient to assess the following serious risks: risk of inflammatory bowel disease.

The ARIA safety assessment will be posted to the Sentinel website.[3] Once there is sufficient product uptake to support an analysis, an analysis plan will be posted online. After the analysis is complete, FDA will also post the results on the Sentinel website. FDA will notify you prior to posting the analysis plan and prior to posting the results.

https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2019/211801Orig1s000ltr.pdf

# Examine medication safety during pregnancy



Received: 11 April 2022 | Revised: 14 July 2022 | Accepted: 21 July 2022

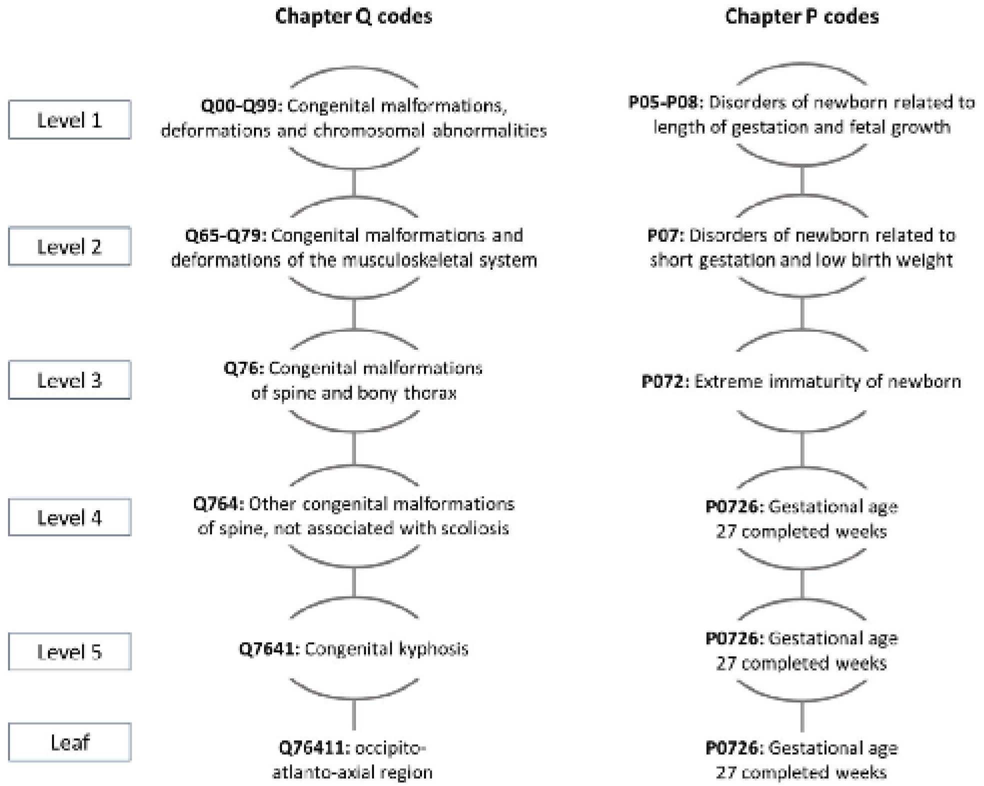
DOI: 10.1002/pds.5512

ORIGINAL ARTICLE

WILEY

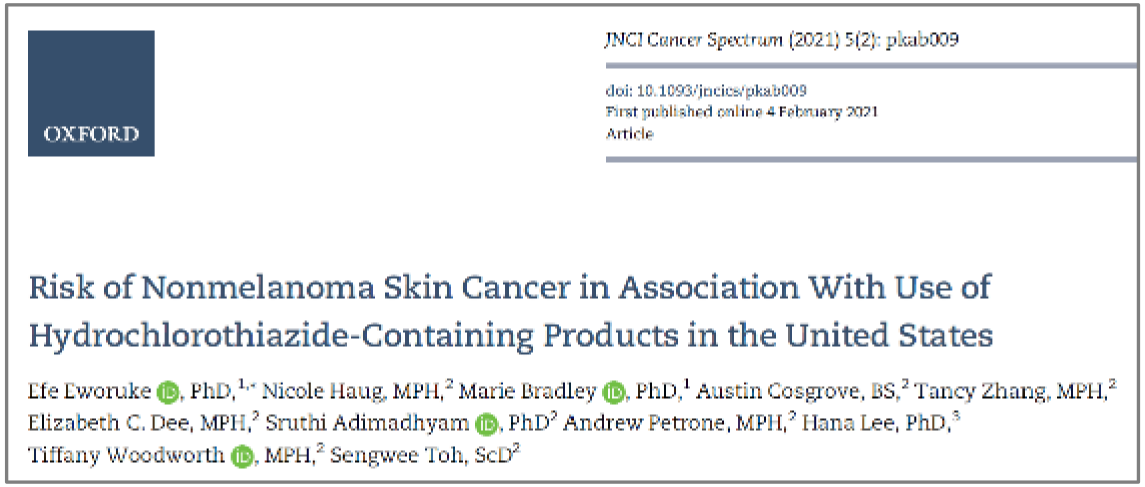
## Novel methods for pregnancy drug safety surveillance in the FDA Sentinel System

Elizabeth A. Suarez[1] | Michael Nguyen[2] | Di Zhang[3] | Yueqin Zhao[3] | Danijela Stojanovic[2] | Monica Munoz[4] | Jane Liedtka[5] | Abby Anderson[6] | Wei Liu[7] | Inna Dashevsky[1] | David Cole[1] | Sandra DeLuccia[1] | Talia Menzin[1] | Jennifer Noble[1] | Judith C. Maro[1]



| | Chapter Q codes | Chapter P codes |
|---|---|---|
| Level 1 | **Q00-Q99:** Congenital malformations, deformations and chromosomal abnormalities | **P05-P08:** Disorders of newborn related to length of gestation and fetal growth |
| Level 2 | **Q65-Q79:** Congenital malformations and deformations of the musculoskeletal system | **P07:** Disorders of newborn related to short gestation and low birth weight |
| Level 3 | **Q76:** Congenital malformations of spine and bony thorax | **P072:** Extreme immaturity of newborn |
| Level 4 | **Q764:** Other congenital malformations of spine, not associated with scoliosis | **P0726:** Gestational age 27 completed weeks |
| Level 5 | **Q7641:** Congenital kyphosis | **P0726:** Gestational age 27 completed weeks |
| Leaf | **Q76411:** occipito-atlanto-axial region | **P0726:** Gestational age 27 completed weeks |

# Inform label change



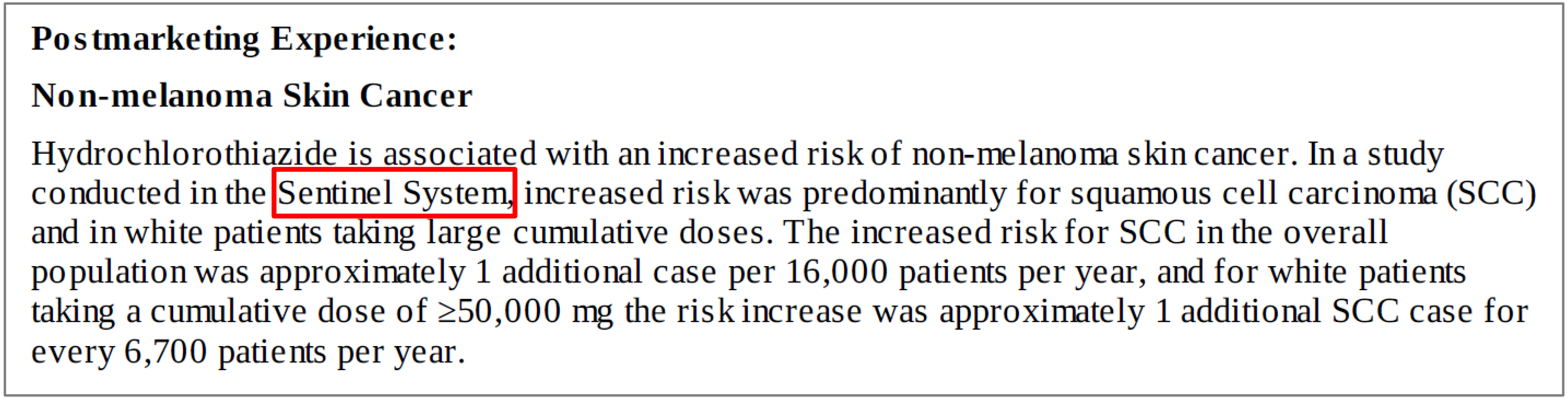
OXFORD

JNCI Cancer Spectrum (2021) 5(2): pkab009

doi: 10.1093/jncics/pkab009
First published online 4 February 2021
Article

Risk of Nonmelanoma Skin Cancer in Association With Use of Hydrochlorothiazide-Containing Products in the United States

Efe Eworuke, PhD,[1,*] Nicole Haug, MPH,[2] Marie Bradley, PhD,[1] Austin Cosgrove, BS,[2] Tancy Zhang, MPH,[2] Elizabeth C. Dee, MPH,[2] Sruthi Adimadhyam, PhD[2] Andrew Petrone, MPH,[2] Hana Lee, PhD,[3] Tiffany Woodworth, MPH,[2] Sengwee Toh, ScD[2]



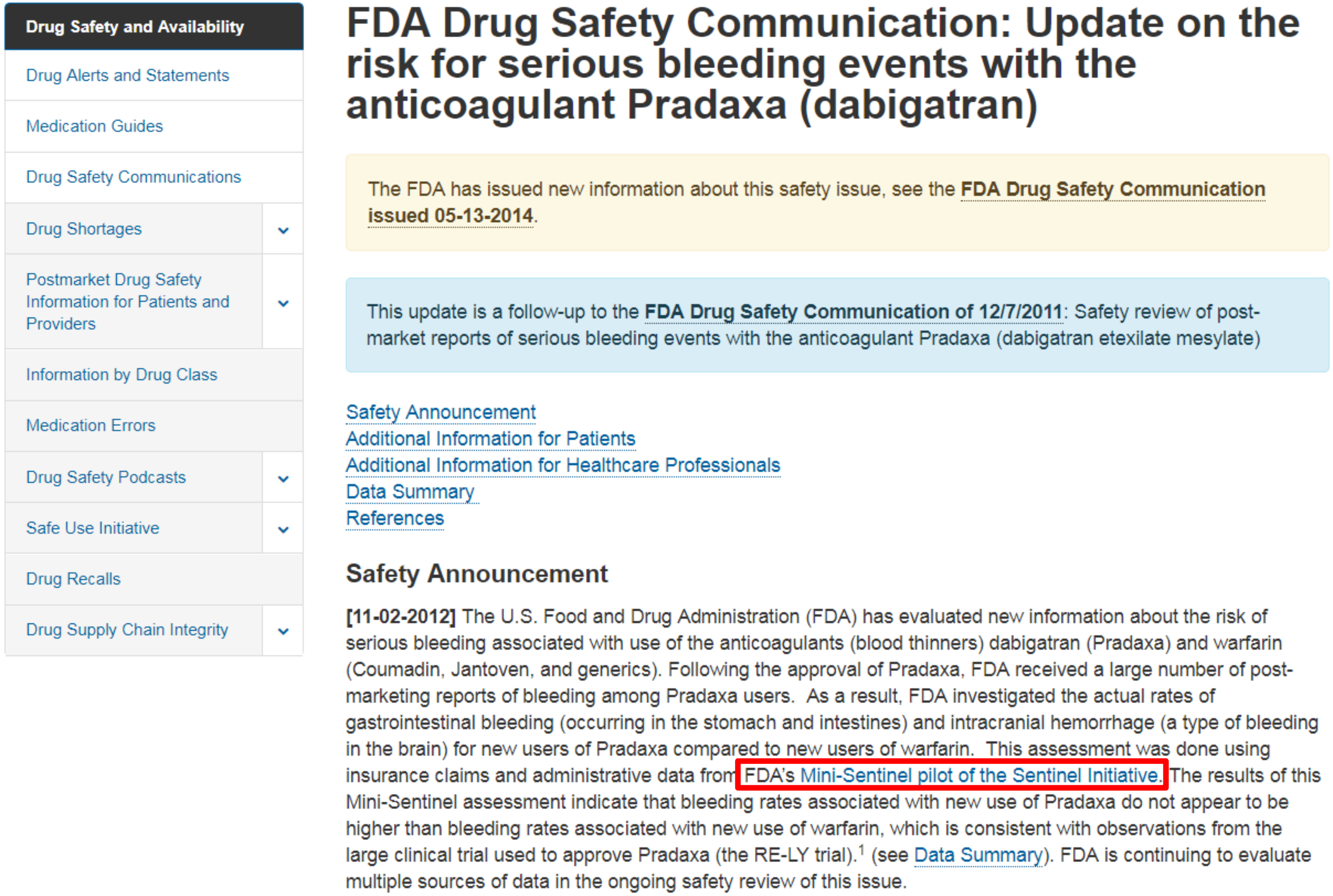
**Postmarketing Experience:**

**Non-melanoma Skin Cancer**

Hydrochlorothiazide is associated with an increased risk of non-melanoma skin cancer. In a study conducted in the Sentinel System, increased risk was predominantly for squamous cell carcinoma (SCC) and in white patients taking large cumulative doses. The increased risk for SCC in the overall population was approximately 1 additional case per 16,000 patients per year, and for white patients taking a cumulative dose of ≥50,000 mg the risk increase was approximately 1 additional SCC case for every 6,700 patients per year.

https://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=02e96a51-1d56-460c-8c20-3d6f37e0ce46

# Contribute to FDA Drug Safety Communication


Drug Safety and Availability
Drug Alerts and Statements
Medication Guides
Drug Safety Communications
Drug Shortages
Postmarket Drug Safety Information for Patients and Providers
Information by Drug Class
Medication Errors
Drug Safety Podcasts
Safe Use Initiative
Drug Recalls
Drug Supply Chain Integrity

## FDA Drug Safety Communication: Update on the risk for serious bleeding events with the anticoagulant Pradaxa (dabigatran)

The FDA has issued new information about this safety issue, see the **FDA Drug Safety Communication issued 05-13-2014**.

This update is a follow-up to the **FDA Drug Safety Communication of 12/7/2011**: Safety review of post-market reports of serious bleeding events with the anticoagulant Pradaxa (dabigatran etexilate mesylate)

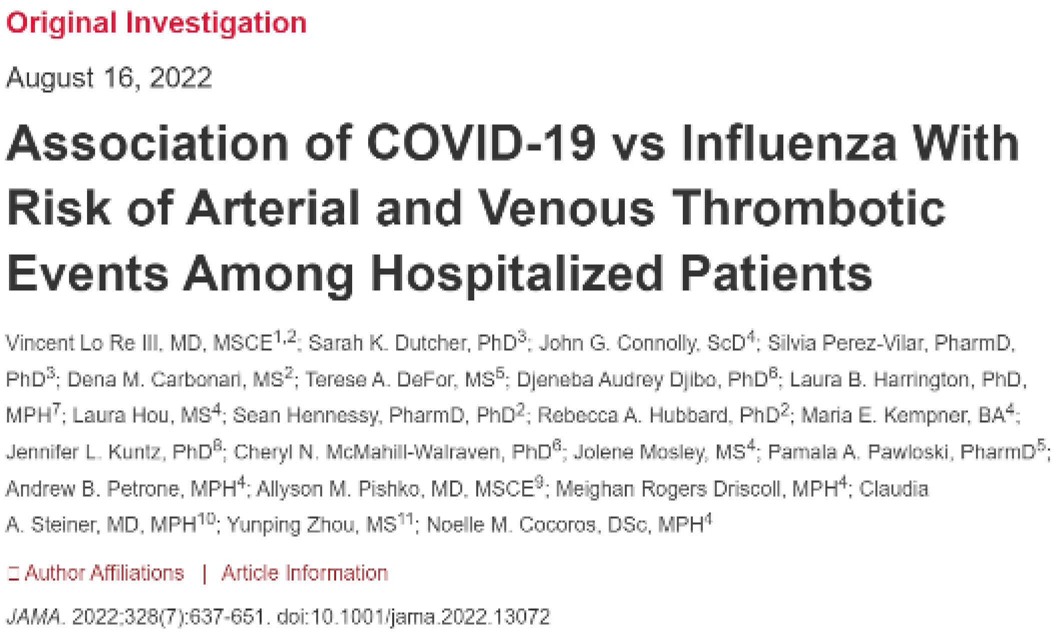
Safety Announcement
Additional Information for Patients
Additional Information for Healthcare Professionals
Data Summary
References

### Safety Announcement

**[11-02-2012]** The U.S. Food and Drug Administration (FDA) has evaluated new information about the risk of serious bleeding associated with use of the anticoagulants (blood thinners) dabigatran (Pradaxa) and warfarin (Coumadin, Jantoven, and generics). Following the approval of Pradaxa, FDA received a large number of post-marketing reports of bleeding among Pradaxa users. As a result, FDA investigated the actual rates of gastrointestinal bleeding (occurring in the stomach and intestines) and intracranial hemorrhage (a type of bleeding in the brain) for new users of Pradaxa compared to new users of warfarin. This assessment was done using insurance claims and administrative data from FDA's Mini-Sentinel pilot of the Sentinel Initiative. The results of this Mini-Sentinel assessment indicate that bleeding rates associated with new use of Pradaxa do not appear to be higher than bleeding rates associated with new use of warfarin, which is consistent with observations from the large clinical trial used to approve Pradaxa (the RE-LY trial).[1] (see Data Summary). FDA is continuing to evaluate multiple sources of data in the ongoing safety review of this issue.

Southworth et al. N Engl J Med 2013;368:1272-1274
https://wayback.archive-it.org/7993/20170112031650/http:/www.fda.gov/Drugs/DrugSafety/ucm326580.htm

# Generate timely evidence during pandemic



Original Investigation

August 16, 2022

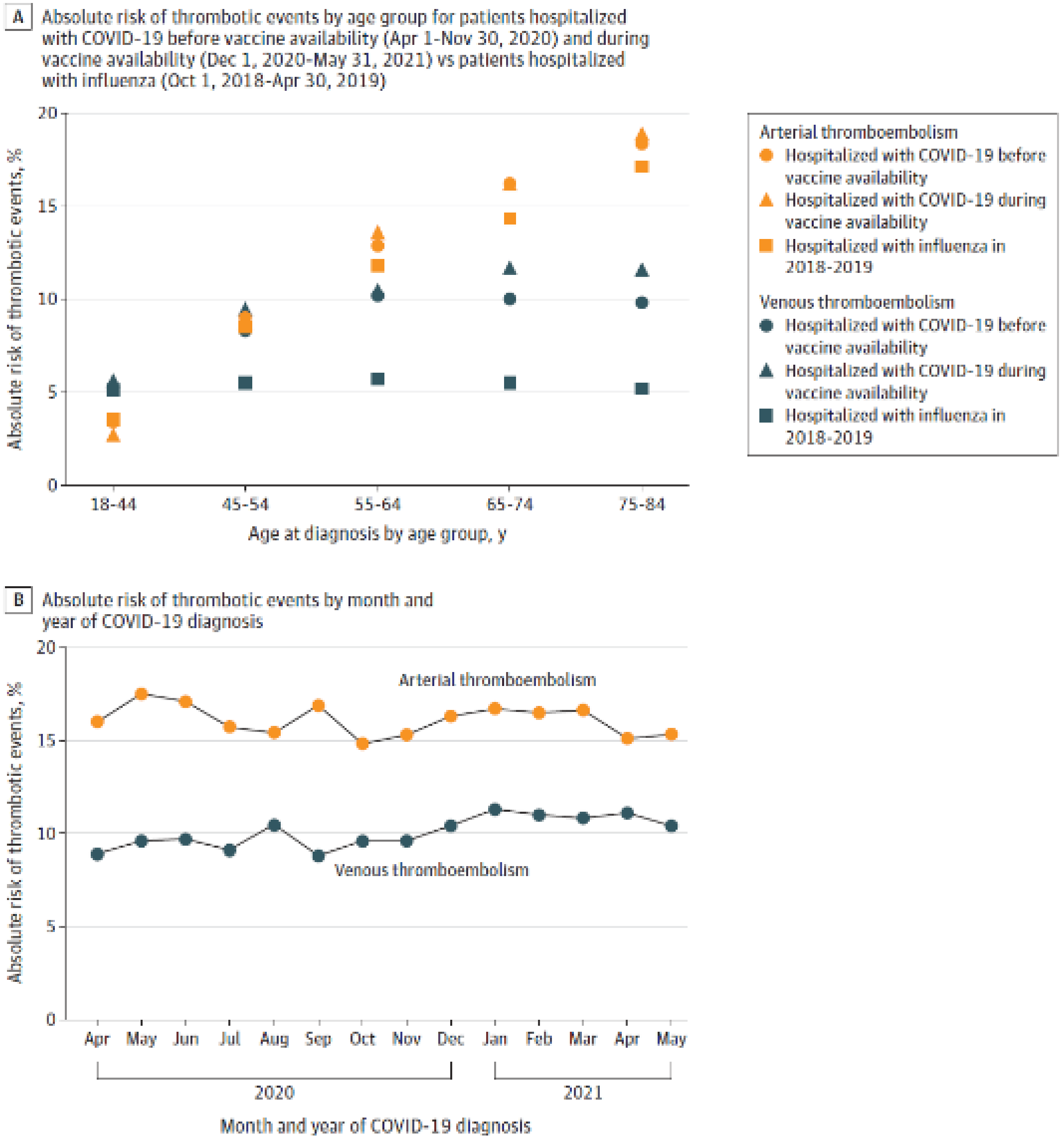
## Association of COVID-19 vs Influenza With Risk of Arterial and Venous Thrombotic Events Among Hospitalized Patients

Vincent Lo Re III, MD, MSCE[1,2]; Sarah K. Dutcher, PhD[3]; John G. Connolly, ScD[4]; Silvia Perez-Vilar, PharmD, PhD[3]; Dena M. Carbonari, MS[2]; Terese A. DeFor, MS[5]; Djeneba Audrey Djibo, PhD[6]; Laura B. Harrington, PhD, MPH[7]; Laura Hou, MS[4]; Sean Hennessy, PharmD, PhD[2]; Rebecca A. Hubbard, PhD[2]; Maria E. Kempner, BA[4]; Jennifer L. Kuntz, PhD[8]; Cheryl N. McMahill-Walraven, PhD[6]; Jolene Mosley, MS[4]; Pamala A. Pawloski, PharmD[5]; Andrew B. Petrone, MPH[4]; Allyson M. Pishko, MD, MSCE[9]; Meighan Rogers Driscoll, MPH[4]; Claudia A. Steiner, MD, MPH[10]; Yunping Zhou, MS[11]; Noelle M. Cocoros, DSc, MPH[4]

□ Author Affiliations | Article Information

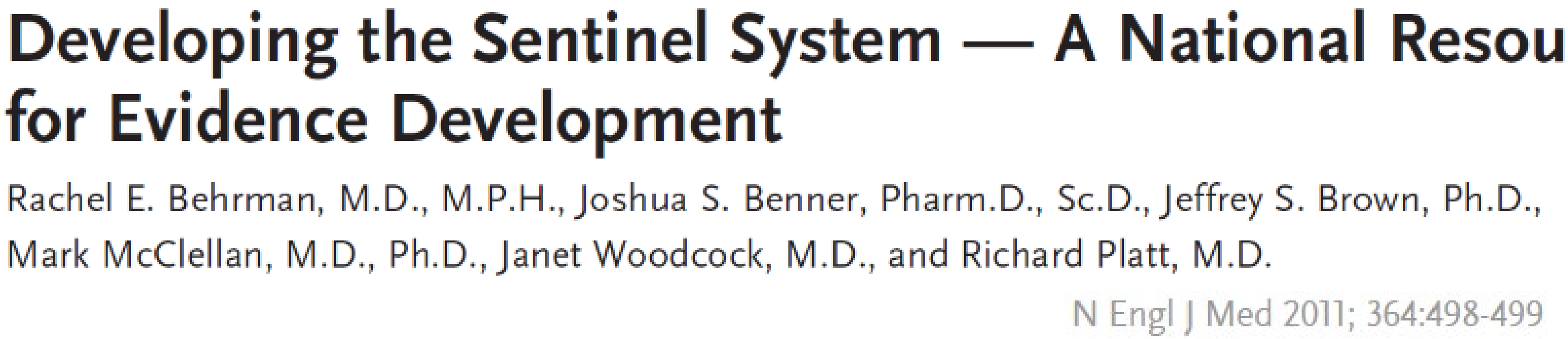
*JAMA*. 2022;328(7):637-651. doi:10.1001/jama.2022.13072

Figure. Absolute Risk of Inpatient Arterial and Venous Thrombotic Events



A Absolute risk of thrombotic events by age group for patients hospitalized with COVID-19 before vaccine availability (Apr 1-Nov 30, 2020) and during vaccine availability (Dec 1, 2020-May 31, 2021) vs patients hospitalized with influenza (Oct 1, 2018-Apr 30, 2019)

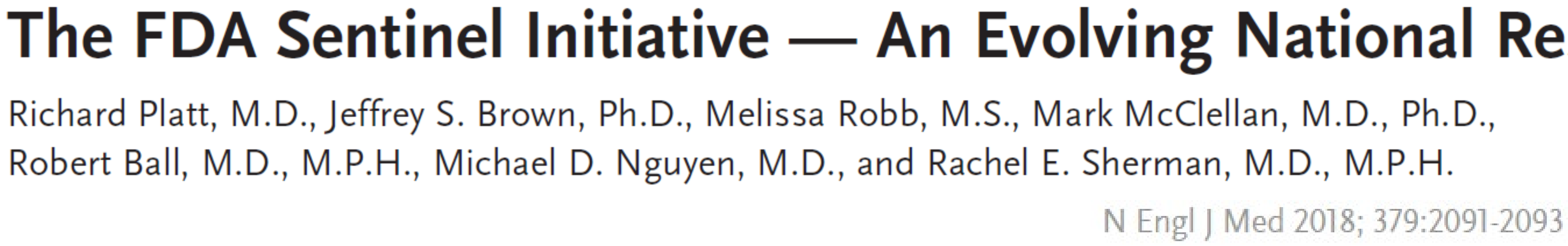
B Absolute risk of thrombotic events by month and year of COVID-19 diagnosis



# Developing the Sentinel System — A National Resource for Evidence Development

Rachel E. Behrman, M.D., M.P.H., Joshua S. Benner, Pharm.D., Sc.D., Jeffrey S. Brown, Ph.D.,
Mark McClellan, M.D., Ph.D., Janet Woodcock, M.D., and Richard Platt, M.D.

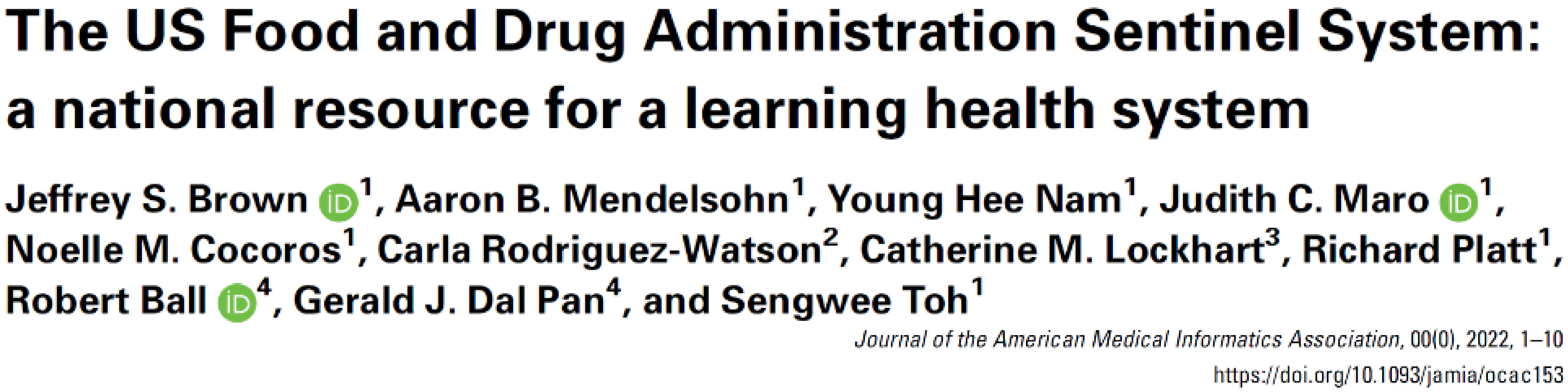
N Engl J Med 2011; 364:498-499



# The FDA Sentinel Initiative — An Evolving National Resource

Richard Platt, M.D., Jeffrey S. Brown, Ph.D., Melissa Robb, M.S., Mark McClellan, M.D., Ph.D.,
Robert Ball, M.D., M.P.H., Michael D. Nguyen, M.D., and Rachel E. Sherman, M.D., M.P.H.

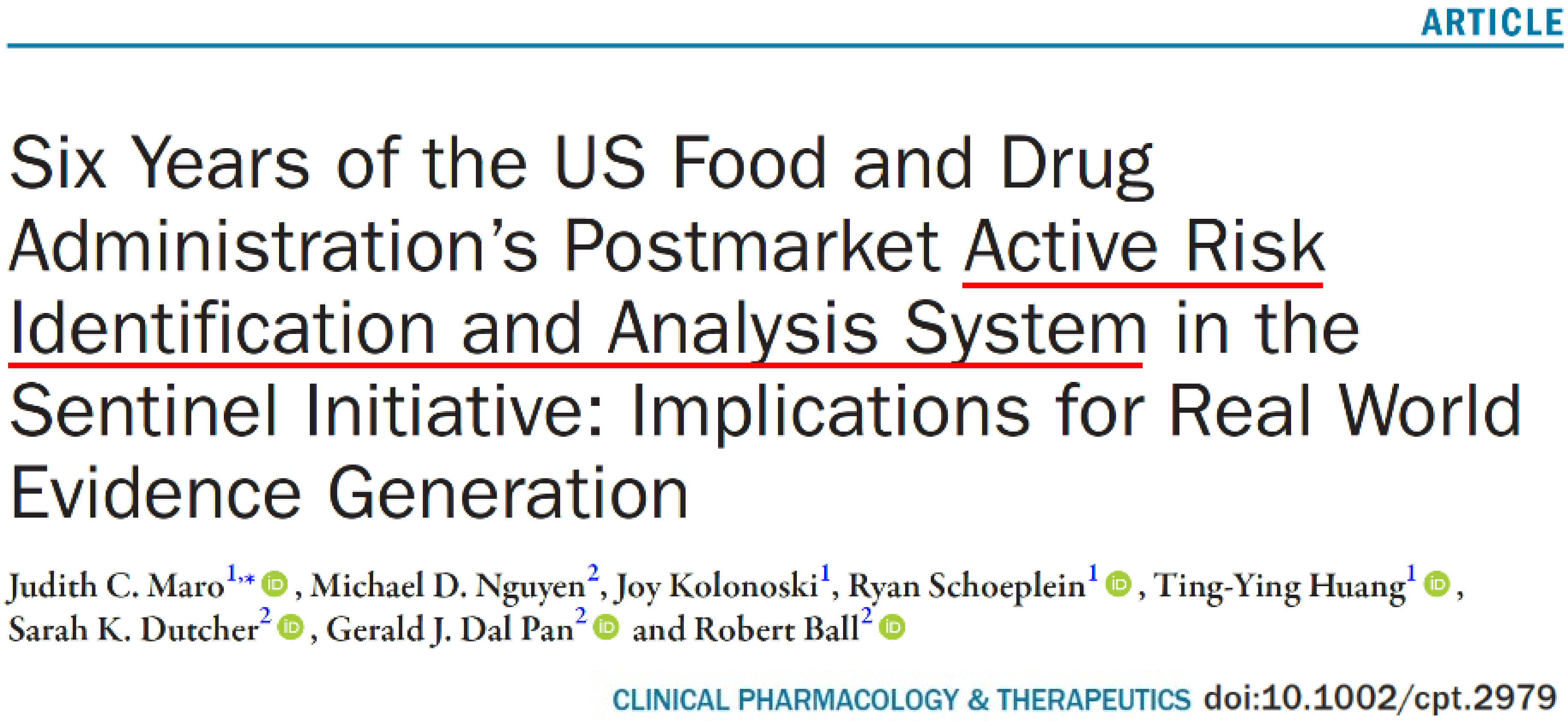
N Engl J Med 2018; 379:2091-2093



# The US Food and Drug Administration Sentinel System: a national resource for a learning health system

Jeffrey S. Brown[1], Aaron B. Mendelsohn[1], Young Hee Nam[1], Judith C. Maro[1],
Noelle M. Cocoros[1], Carla Rodriguez-Watson[2], Catherine M. Lockhart[3], Richard Platt[1],
Robert Ball[4], Gerald J. Dal Pan[4], and Sengwee Toh[1]

*Journal of the American Medical Informatics Association*, 00(0), 2022, 1–10
https://doi.org/10.1093/jamia/ocac153

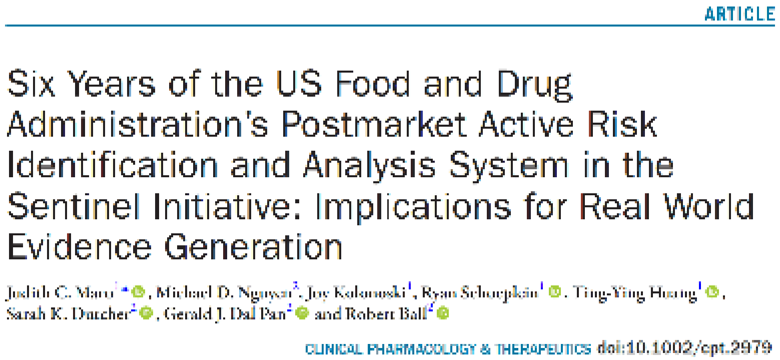


ARTICLE

# Six Years of the US Food and Drug Administration's Postmarket Active Risk Identification and Analysis System in the Sentinel Initiative: Implications for Real World Evidence Generation

Judith C. Maro[1,*], Michael D. Nguyen[2], Joy Kolonoski[1], Ryan Schoeplein[1], Ting-Ying Huang[1], Sarah K. Dutcher[2], Gerald J. Dal Pan[2] and Robert Ball[2]

CLINICAL PHARMACOLOGY & THERAPEUTICS doi:10.1002/cpt.2979



ARTICLE

# Six Years of the US Food and Drug Administration's Postmarket Active Risk Identification and Analysis System in the Sentinel Initiative: Implications for Real World Evidence Generation

Judith C. Maro[1,*], Michael D. Nguyen[2], Joy Kolonoski[1], Ryan Schoeplein[1], Ting-Ying Huang[1], Sarah K. Dutcher[2], Gerald J. Dal Pan[2] and Robert Ball[2]

CLINICAL PHARMACOLOGY & THERAPEUTICS doi:10.1002/cpt.2979

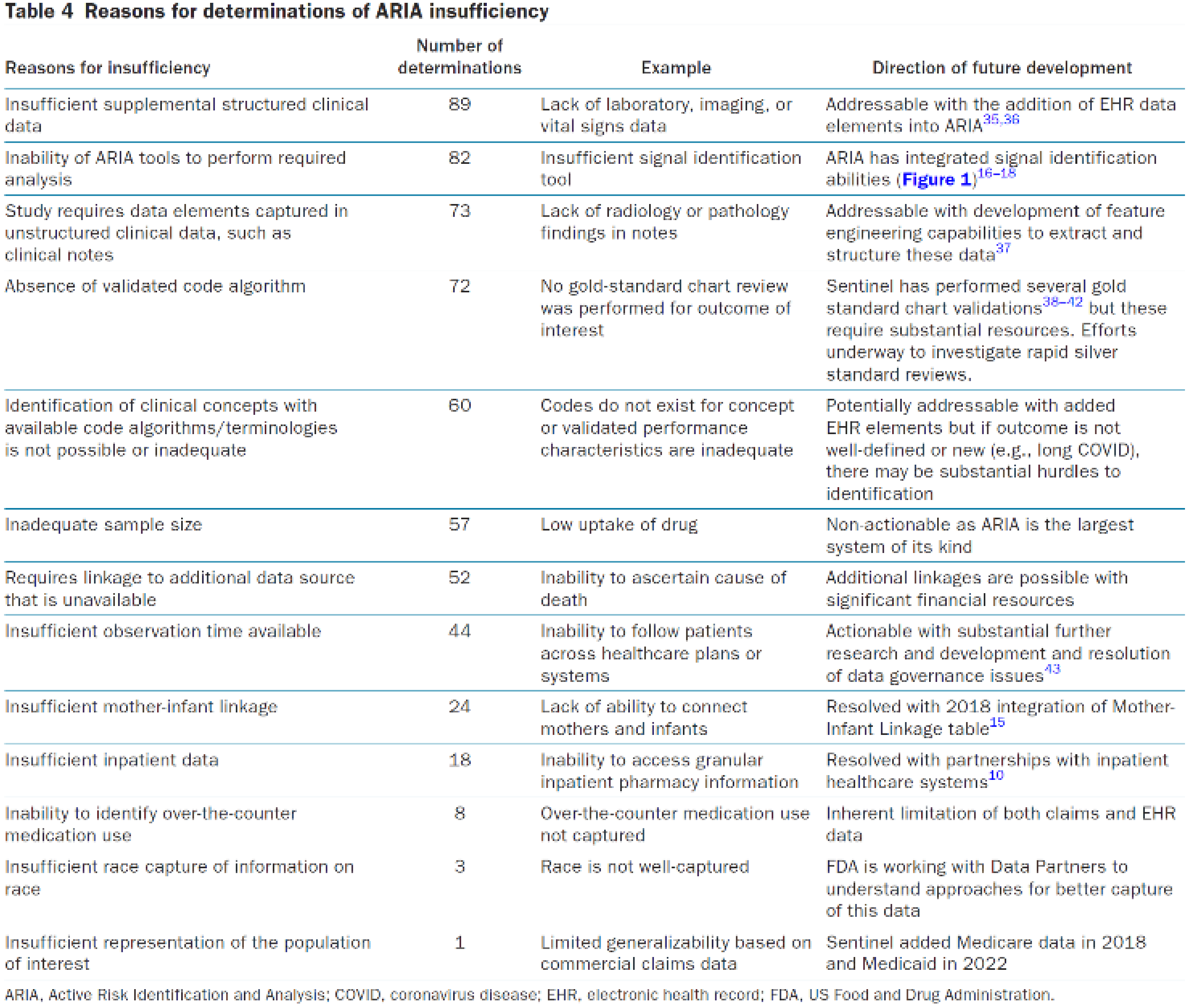
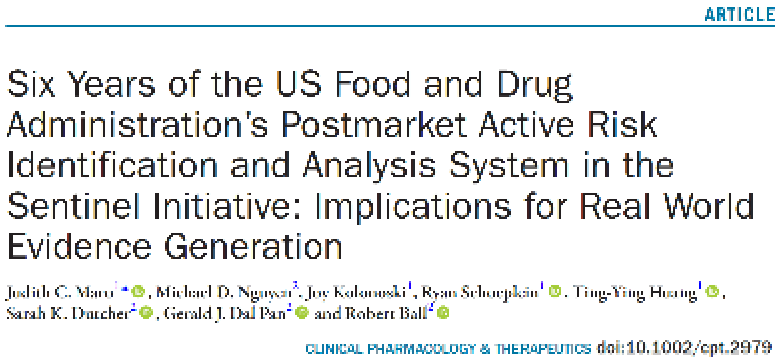


**Table 4 Reasons for determinations of ARIA insufficiency**

| Reasons for insufficiency | Number of determinations | Example | Direction of future development |
|---|---|---|---|
| Insufficient supplemental structured clinical data | 89 | Lack of laboratory, imaging, or vital signs data | Addressable with the addition of EHR data elements into ARIA[35,36] |
| Inability of ARIA tools to perform required analysis | 82 | Insufficient signal identification tool | ARIA has integrated signal identification abilities (**Figure 1**)[16–18] |
| Study requires data elements captured in unstructured clinical data, such as clinical notes | 73 | Lack of radiology or pathology findings in notes | Addressable with development of feature engineering capabilities to extract and structure these data[37] |
| Absence of validated code algorithm | 72 | No gold-standard chart review was performed for outcome of interest | Sentinel has performed several gold standard chart validations[38–42] but these require substantial resources. Efforts underway to investigate rapid silver standard reviews. |
| Identification of clinical concepts with available code algorithms/terminologies is not possible or inadequate | 60 | Codes do not exist for concept or validated performance characteristics are inadequate | Potentially addressable with added EHR elements but if outcome is not well-defined or new (e.g., long COVID), there may be substantial hurdles to identification |
| Inadequate sample size | 57 | Low uptake of drug | Non-actionable as ARIA is the largest system of its kind |
| Requires linkage to additional data source that is unavailable | 52 | Inability to ascertain cause of death | Additional linkages are possible with significant financial resources |
| Insufficient observation time available | 44 | Inability to follow patients across healthcare plans or systems | Actionable with substantial further research and development and resolution of data governance issues[43] |
| Insufficient mother-infant linkage | 24 | Lack of ability to connect mothers and infants | Resolved with 2018 integration of Mother-Infant Linkage table[15] |
| Insufficient inpatient data | 18 | Inability to access granular inpatient pharmacy information | Resolved with partnerships with inpatient healthcare systems[10] |
| Inability to identify over-the-counter medication use | 8 | Over-the-counter medication use not captured | Inherent limitation of both claims and EHR data |
| Insufficient race capture of information on race | 3 | Race is not well-captured | FDA is working with Data Partners to understand approaches for better capture of this data |
| Insufficient representation of the population of interest | 1 | Limited generalizability based on commercial claims data | Sentinel added Medicare data in 2018 and Medicaid in 2022 |

ARIA, Active Risk Identification and Analysis; COVID, coronavirus disease; EHR, electronic health record; FDA, US Food and Drug Administration.



ARTICLE

# Six Years of the US Food and Drug Administration's Postmarket Active Risk Identification and Analysis System in the Sentinel Initiative: Implications for Real World Evidence Generation

Judith C. Maro[1,*], Michael D. Nguyen[2], Joy Kolonoski[1], Ryan Schoeplein[1], Ting-Ying Huang[1], Sarah K. Dutcher[2], Gerald J. Dal Pan[2] and Robert Ball[2]

CLINICAL PHARMACOLOGY & THERAPEUTICS doi:10.1002/cpt.2979

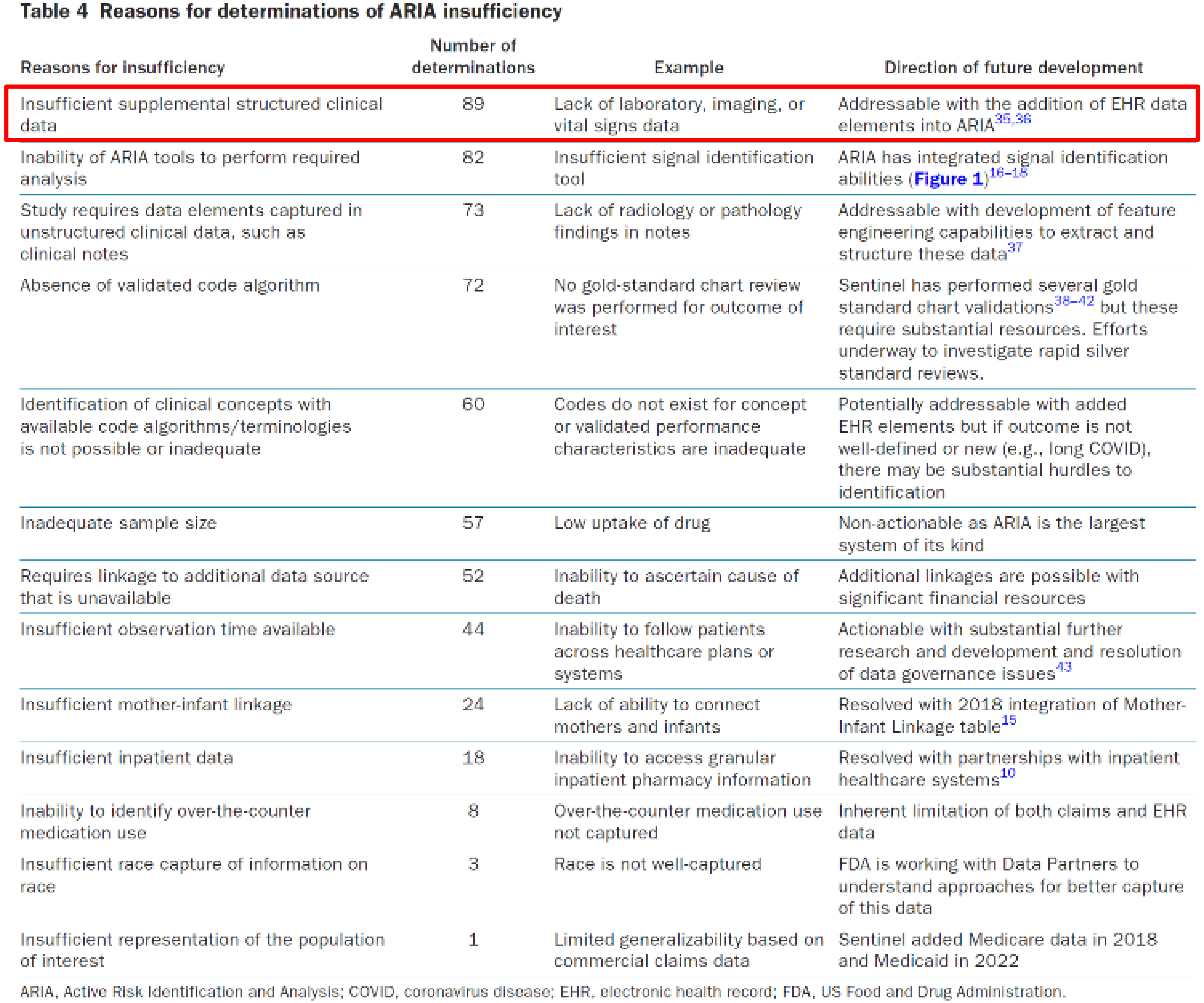
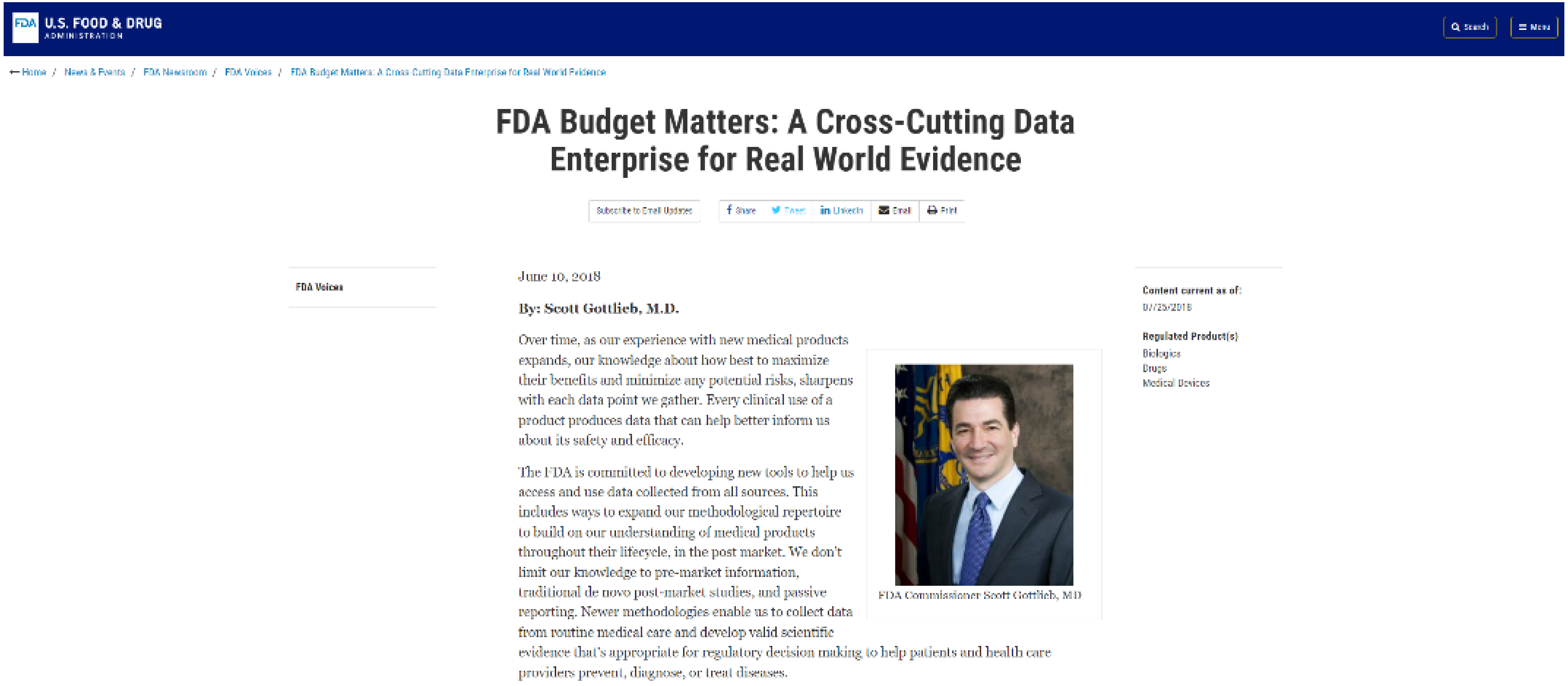


**Table 4 Reasons for determinations of ARIA insufficiency**

| Reasons for insufficiency | Number of determinations | Example | Direction of future development |
|---|---|---|---|
| Insufficient supplemental structured clinical data | 89 | Lack of laboratory, imaging, or vital signs data | Addressable with the addition of EHR data elements into ARIA[35,36] |
| Inability of ARIA tools to perform required analysis | 82 | Insufficient signal identification tool | ARIA has integrated signal identification abilities (Figure 1)[16–18] |
| Study requires data elements captured in unstructured clinical data, such as clinical notes | 73 | Lack of radiology or pathology findings in notes | Addressable with development of feature engineering capabilities to extract and structure these data[37] |
| Absence of validated code algorithm | 72 | No gold-standard chart review was performed for outcome of interest | Sentinel has performed several gold standard chart validations[38–42] but these require substantial resources. Efforts underway to investigate rapid silver standard reviews. |
| Identification of clinical concepts with available code algorithms/terminologies is not possible or inadequate | 60 | Codes do not exist for concept or validated performance characteristics are inadequate | Potentially addressable with added EHR elements but if outcome is not well-defined or new (e.g., long COVID), there may be substantial hurdles to identification |
| Inadequate sample size | 57 | Low uptake of drug | Non-actionable as ARIA is the largest system of its kind |
| Requires linkage to additional data source that is unavailable | 52 | Inability to ascertain cause of death | Additional linkages are possible with significant financial resources |
| Insufficient observation time available | 44 | Inability to follow patients across healthcare plans or systems | Actionable with substantial further research and development and resolution of data governance issues[43] |
| Insufficient mother-infant linkage | 24 | Lack of ability to connect mothers and infants | Resolved with 2018 integration of Mother-Infant Linkage table[15] |
| Insufficient inpatient data | 18 | Inability to access granular inpatient pharmacy information | Resolved with partnerships with inpatient healthcare systems[10] |
| Inability to identify over-the-counter medication use | 8 | Over-the-counter medication use not captured | Inherent limitation of both claims and EHR data |
| Insufficient race capture of information on race | 3 | Race is not well-captured | FDA is working with Data Partners to understand approaches for better capture of this data |
| Insufficient representation of the population of interest | 1 | Limited generalizability based on commercial claims data | Sentinel added Medicare data in 2018 and Medicaid in 2022 |

ARIA, Active Risk Identification and Analysis; COVID, coronavirus disease; EHR, electronic health record; FDA, US Food and Drug Administration.


U.S. FOOD & DRUG ADMINISTRATION
Search
Menu
Home / News & Events / FDA Newsroom / FDA Voices / FDA Budget Matters: A Cross Cutting Data Enterprise for Real World Evidence
Subscribe to Email Updates
Share
Tweet
LinkedIn
Email
Print
FDA Voices

# FDA Budget Matters: A Cross-Cutting Data Enterprise for Real World Evidence

June 10, 2018

**By: Scott Gottlieb, M.D.**

Over time, as our experience with new medical products expands, our knowledge about how best to maximize their benefits and minimize any potential risks, sharpens with each data point we gather. Every clinical use of a product produces data that can help better inform us about its safety and efficacy.

The FDA is committed to developing new tools to help us access and use data collected from all sources. This includes ways to expand our methodological repertoire to build on our understanding of medical products throughout their lifecycle, in the post market. We don't limit our knowledge to pre-market information, traditional de novo post-market studies, and passive reporting. Newer methodologies enable us to collect data from routine medical care and develop valid scientific evidence that's appropriate for regulatory decision making to help patients and health care providers prevent, diagnose, or treat diseases.

FDA Commissioner Scott Gottlieb, MD

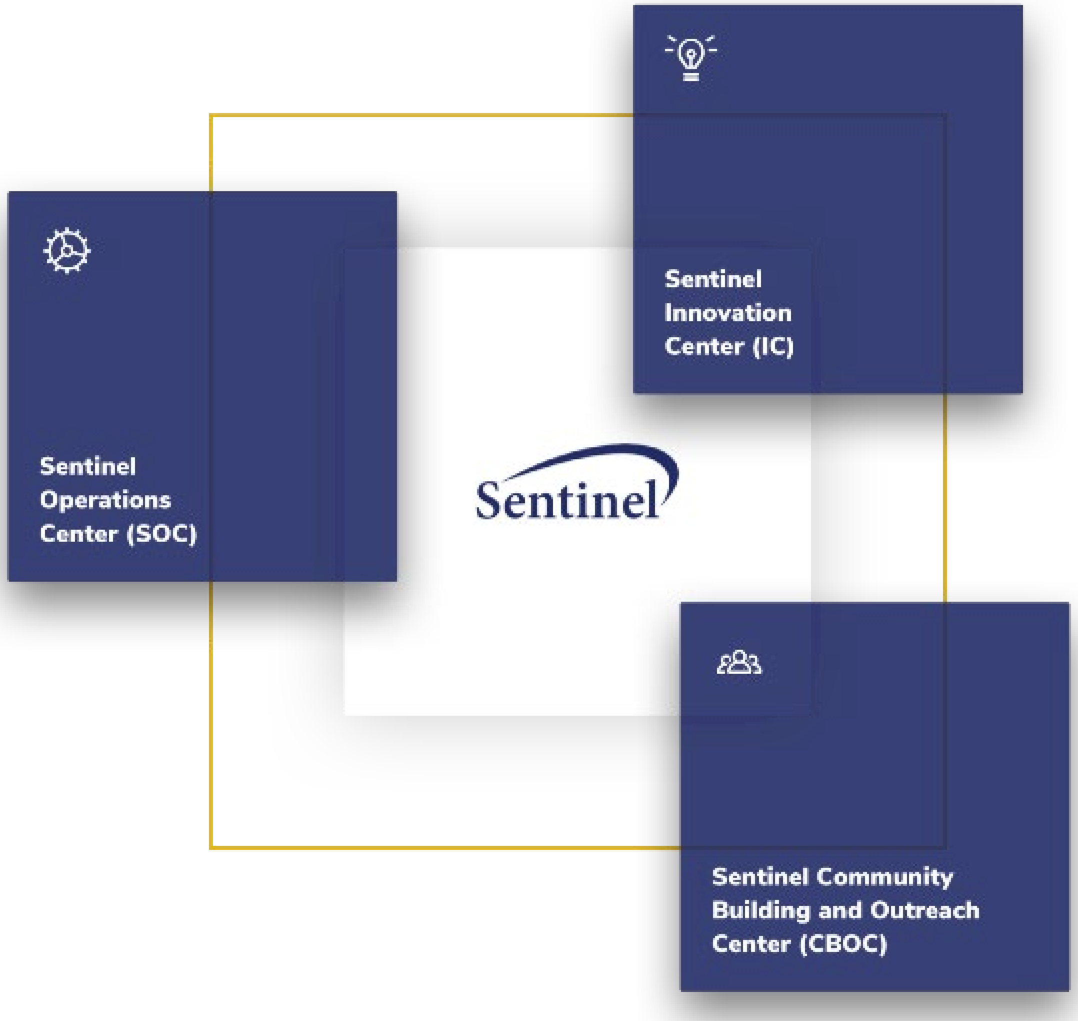
**Content current as of:**

07/25/2018

**Regulated Product(s)**

Biologics
Drugs
Medical Devices

https://www.fda.gov/news-events/fda-voices/fda-budget-matters-cross-cutting-data-enterprise-real-world-evidence



Sentinel Operations Center (SOC)

Sentinel Innovation Center (IC)

Sentinel Community Building and Outreach Center (CBOC)

https://www.sentinelinitiative.org/about/sentinel-structure

Sentinel

# Sentinel Innovation Center Master Plan

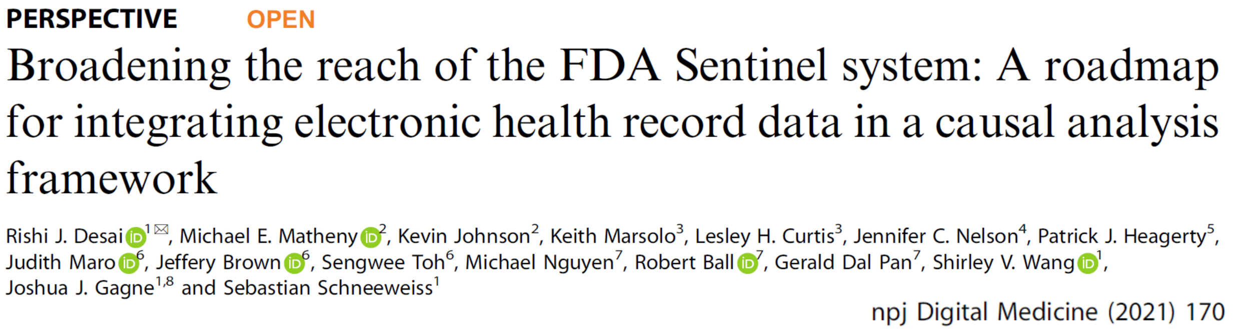
*Sentinel Innovation Center*

Version 1.1

June 17, 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.

https://www.sentinelinitiative.org/sites/default/files/documents/IC-Master-Plan_Updated_0.pdf

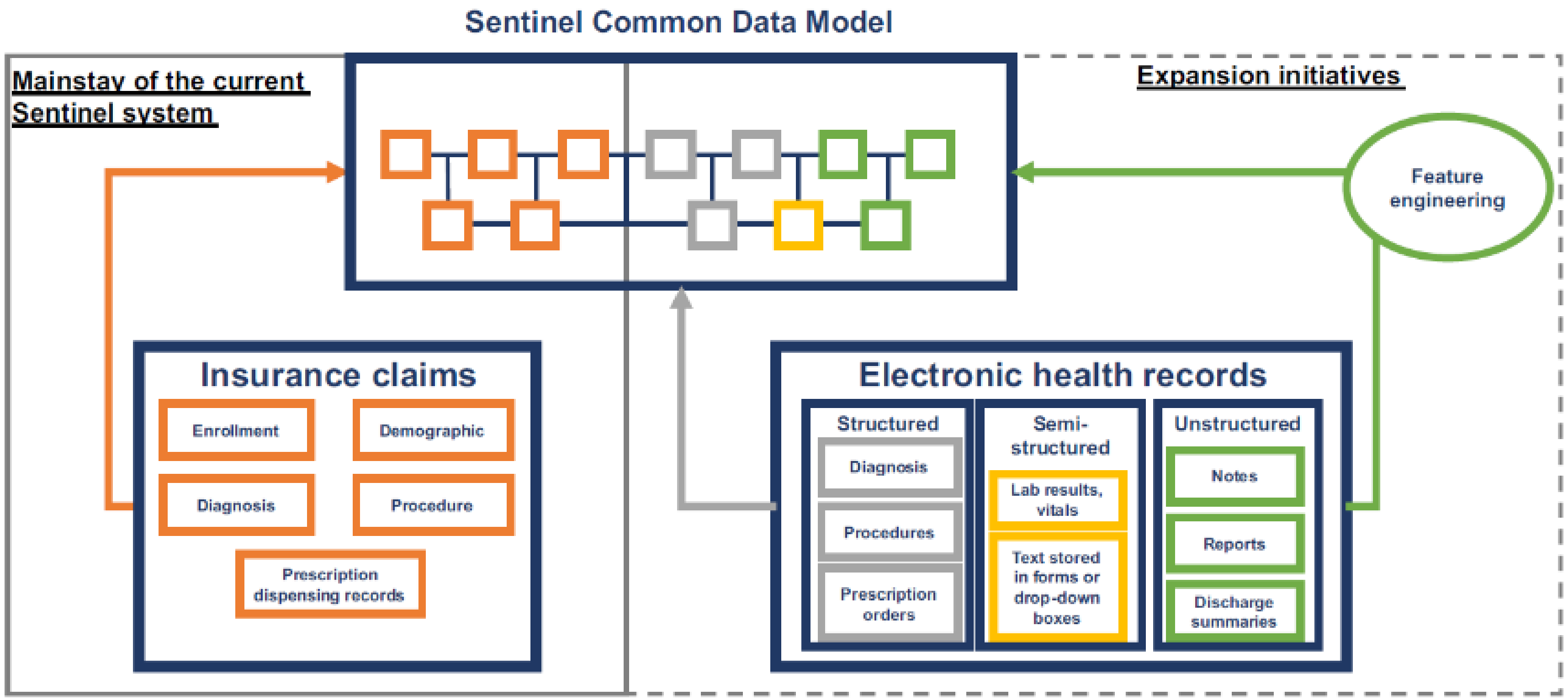


PERSPECTIVE OPEN

# Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework

Rishi J. Desai[1]✉, Michael E. Matheny[2], Kevin Johnson[2], Keith Marsolo[3], Lesley H. Curtis[3], Jennifer C. Nelson[4], Patrick J. Heagerty[5], Judith Maro[6], Jeffery Brown[6], Sengwee Toh[6], Michael Nguyen[7], Robert Ball[7], Gerald Dal Pan[7], Shirley V. Wang[1], Joshua J. Gagne[1,8] and Sebastian Schneeweiss[1]

npj Digital Medicine (2021) 170



## Current Sentinel System Limitations

Inability to identify certain study populations of interest from insurance claims

Inability to identify certain outcomes of interest from insurance claims

Other limitations (inadequate duration of follow-up, the need for additional signal identification tools)

## Sentinel Innovation Center Initiatives

### Data infrastructure (DI)

10+ million people
EHR
+
Claims

### Feature engineering (FE)

- Emerging methods including machine learning and scalable automated natural language processing (NLP) approaches to enable computable phenotyping from unstructured EHR data

### Causal inference (CI)

- Methodologic research to address specific challenges when using EHRs such as approaches to handle missing data, calibration methods for enhanced confounding adjustment

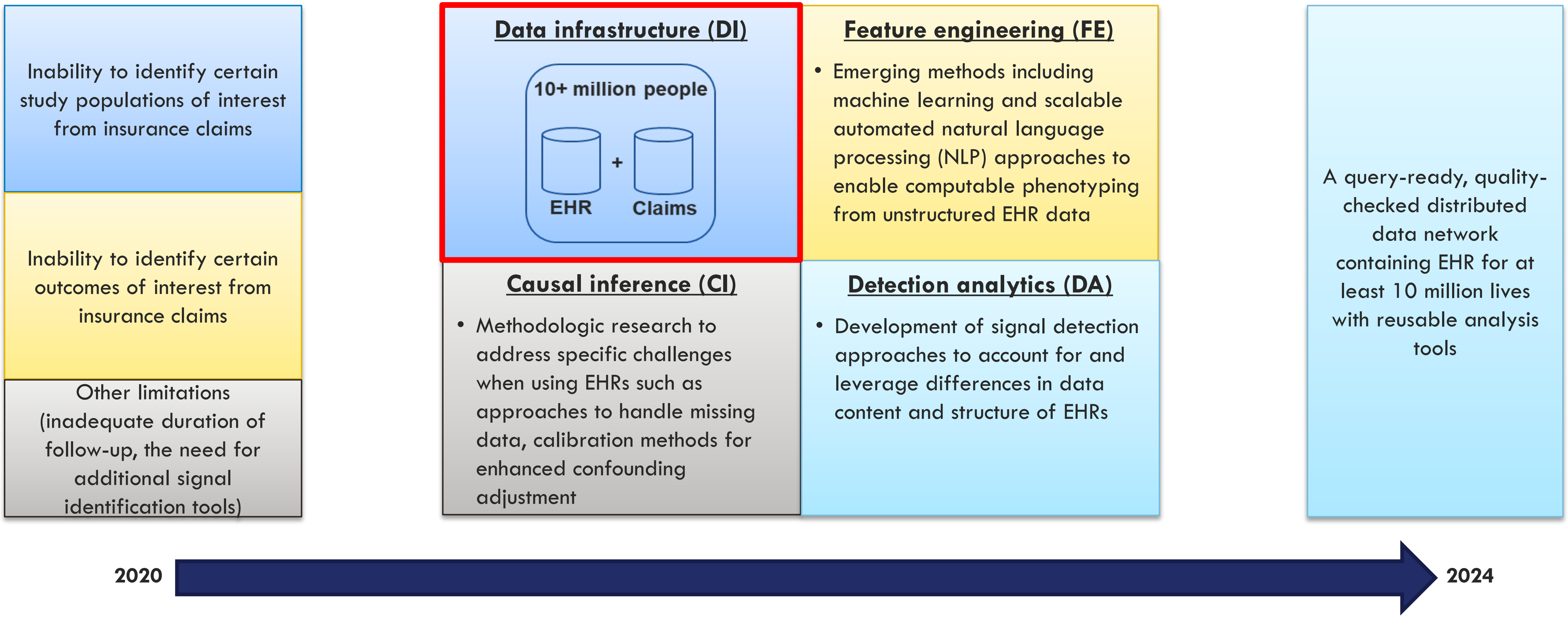
### Detection analytics (DA)

- Development of signal detection approaches to account for and leverage differences in data content and structure of EHRs

## Sentinel Innovation Center Vision

A query-ready, quality-checked distributed data network containing EHR for at least 10 million lives with reusable analysis tools

2020 → 2024



## Current Sentinel System Limitations

Inability to identify certain study populations of interest from insurance claims

Inability to identify certain outcomes of interest from insurance claims

Other limitations (inadequate duration of follow-up, the need for additional signal identification tools)

## Sentinel Innovation Center Initiatives

**Data infrastructure (DI)**

**Feature engineering (FE)**

- Emerging methods including machine learning and scalable automated natural language processing (NLP) approaches to enable computable phenotyping from unstructured EHR data

**Causal inference (CI)**

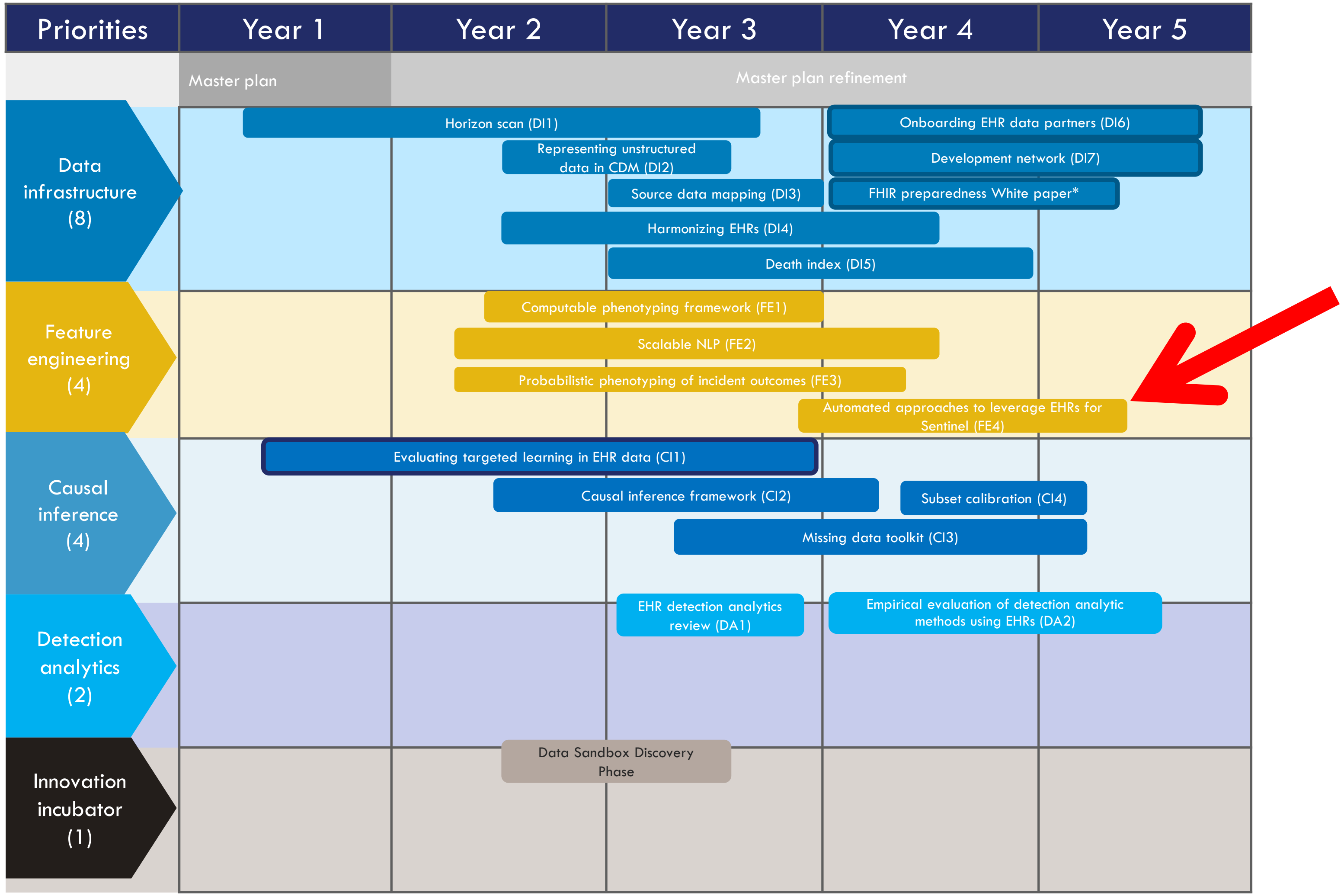
- Methodologic research to address specific challenges when using EHRs such as approaches to handle missing data, calibration methods for enhanced confounding adjustment

**Detection analytics (DA)**

- Development of signal detection approaches to account for and leverage differences in data content and structure of EHRs

## Sentinel Innovation Center Vision

A query-ready, quality-checked distributed data network containing EHR for at least 10 million lives with reusable analysis tools



Priorities | Year 1 | Year 2 | Year 3 | Year 4 | Year 5

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*

Feature engineering (4)

- Computable phenotyping framework (FE1)
- Scalable NLP (FE2)
- Probabilistic phenotyping of incident outcomes (FE3)
- Automated approaches to leverage EHRs for Sentinel (FE4)

Causal inference (4)

- Evaluating targeted learning in EHR data (CI1)
- Causal inference framework (CI2)
- Subset calibration (CI4)
- Missing data toolkit (CI3)

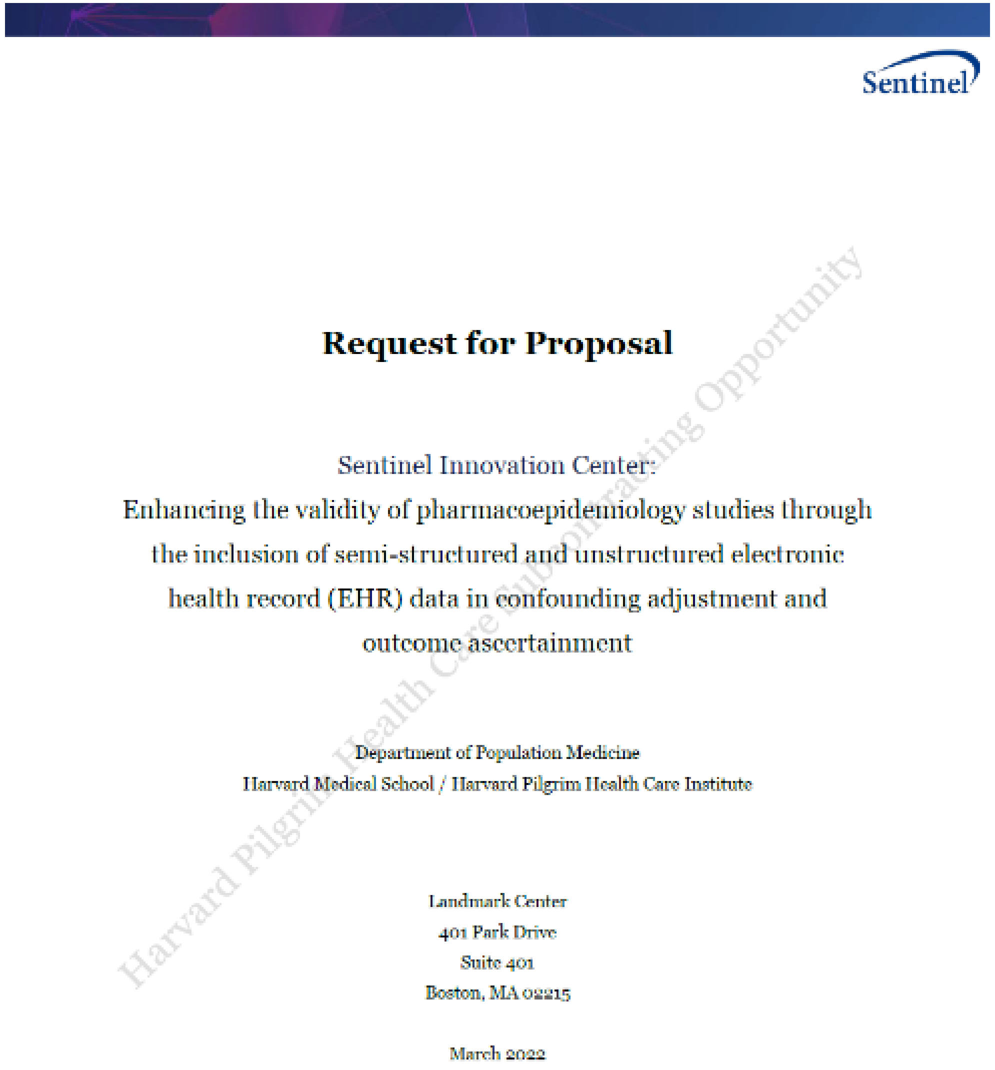
Detection analytics (2)

- EHR detection analytics review (DA1)
- Empirical evaluation of detection analytic methods using EHRs (DA2)

Innovation incubator (1)

- Data Sandbox Discovery Phase

*ASPE supported project



Sentinel

# Request for Proposal

Sentinel Innovation Center:
Enhancing the validity of pharmacoepidemiology studies through the inclusion of semi-structured and unstructured electronic health record (EHR) data in confounding adjustment and outcome ascertainment

Department of Population Medicine
Harvard Medical School / Harvard Pilgrim Health Care Institute

Landmark Center
401 Park Drive
Suite 401
Boston, MA 02215

March 2022

https://www.sentinelinitiative.org/methods-data-tools/methods/sentinel-innovation-center-request-proposal-feature-engineering

# MOSAIC-NLP
# Multi-source Observational Safety study for Advanced Information Classification using NLP

Dena Jaffe, PhD

Dena.Jaffe@oracle.com

# Project Team

## FDA

- **Sarah Dutcher**, Epidemiologist
- **Jummai Apata,** Epidemiologist
- **Robert Lim,** Medical Officer
- **Jie (Jenni) Li,** Epidemiologist
- **Jamal Jones,** Epidemiologist
- **Yong Ma,** Biostatistician
- **Tiffany Austin**, Project Manager

## Sentinel Operations Center/Harvard

- **Meighan Driscoll**, Program Manager
- **Kimberly Gegear**, Project Manager
- **Darren Toh**, Co-investigator, Pharmacoepidemiologist
- **Jenna Wong**, Pharmacoepidemiologist

## Mass General Brigham

- **Richard Wyss,** Co-Investigator, Epidemiologist
- **Jie Yang**, Principal Investigator
- **Rishi Desai**, Operations Chief
- **Josh Lin**, Epidemiologist

## Cerner Enviza, an Oracle Company

- **Elise Berliner**, Principal Investigator
- **Dena Jaffe**, Principal Investigator, Epidemiologist
- **Jenny Cai**, Project Manager
- **Sonam Lama**, Project Manager
- **Nathan Vavroch**, Data Strategist
- **Mike Jones**, Data Strategist
- **Vineela Kommuri**, Senior Data Engineer
- **Sravan Kumar Burla**, Software Engineer
- **Bridget Balkaran**, Lead Biostatistician
- **Austin Yue**, Biostatistician
- **Kyla Finlayson,** Biostatistician
- **Stacey Purinton**, Data Manager
- **Rob Taylor**, Data Manager
- **Eliza Celenti**, Medical Writer

## National Jewish Health

- **Michael Wechsler,** Pulmonologist
- **David Beuther,** Pulmonologist
- **Pearlanne Zelarney,** Research Informatics
- **Alicia Mitchell,** Developer
- **Sarah Rhoads**, Pulmonologist

## Children's Hospital of Orange County

- **Louis Ehwerhemuepha,** Clinical Data Scientist
- **Hoang Nguyen,** Psychiatrist
- **Michael Chu,** Psychiatrist
- **Heather Huszti,** Psychologist
- **Olga Guijon,** Pediatrician and Asthma specialist

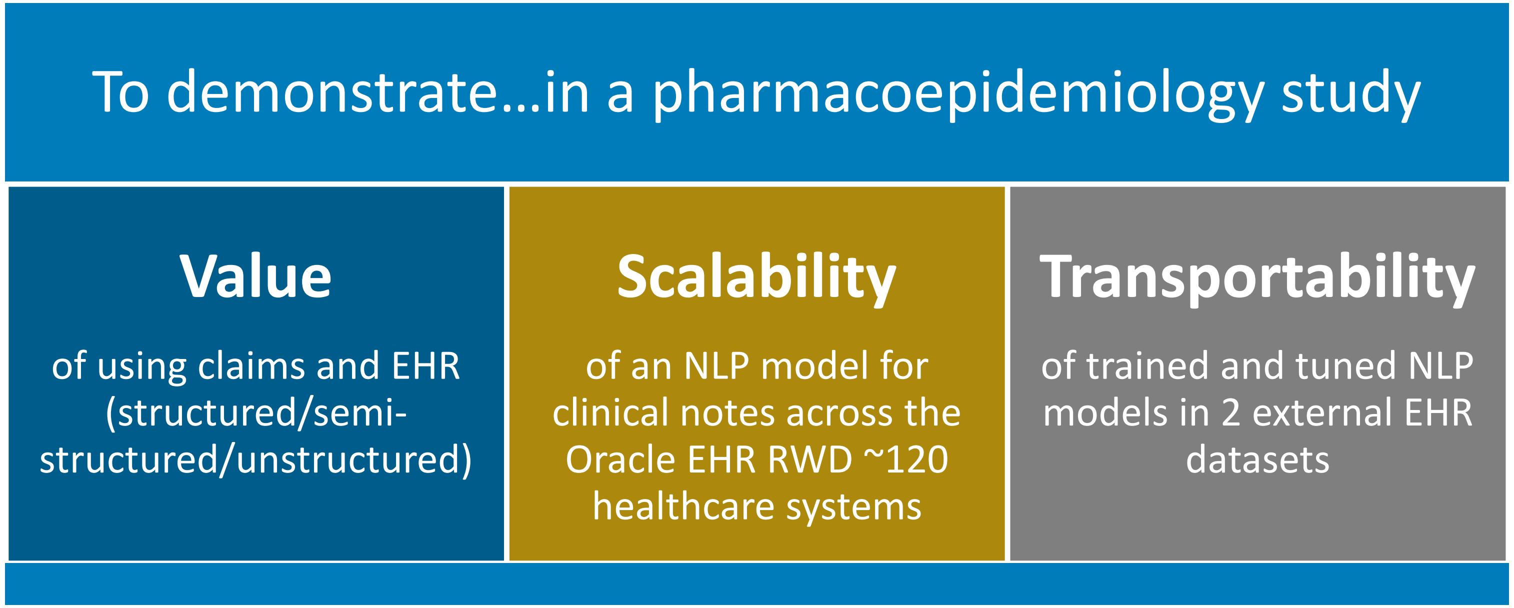
## John Snow Labs

- **David Talby**, CTO
- **Ace Vo**, Project Manager
- **Hasham Ul Haq**, Lead Senior NLP Data Scientist
- **Veysel Kocaman**, Data Scientist
- **Gursev Pirge**, Data Scientist
- **Ahmet Emin Tek**, Data Scientist
- **Andrei Marian Feier**, Clinical Annotation Lead
- **Denisa Popa**, Data Annotator
- **Aleksei Zhakarov**, Annotator
- **Jay Gil**, Annotator
- **Zhenya Nargizyan**, Annotator
- **Jiri Dobles**, Project Manager

## Kaiser Permanente Washington Health Research Institute

- **David Carrell,** NLP Expert Consultant

# Use of Natural Language Processing in a Pharmacoepidemiology Study: The Examination of Neuropsychiatric Events and Incident Use of Montelukast Among Patients with Asthma



To demonstrate…in a pharmacoepidemiology study

**Value**

of using claims and EHR (structured/semi-structured/unstructured)

**Scalability**

of an NLP model for clinical notes across the Oracle EHR RWD ~120 healthcare systems

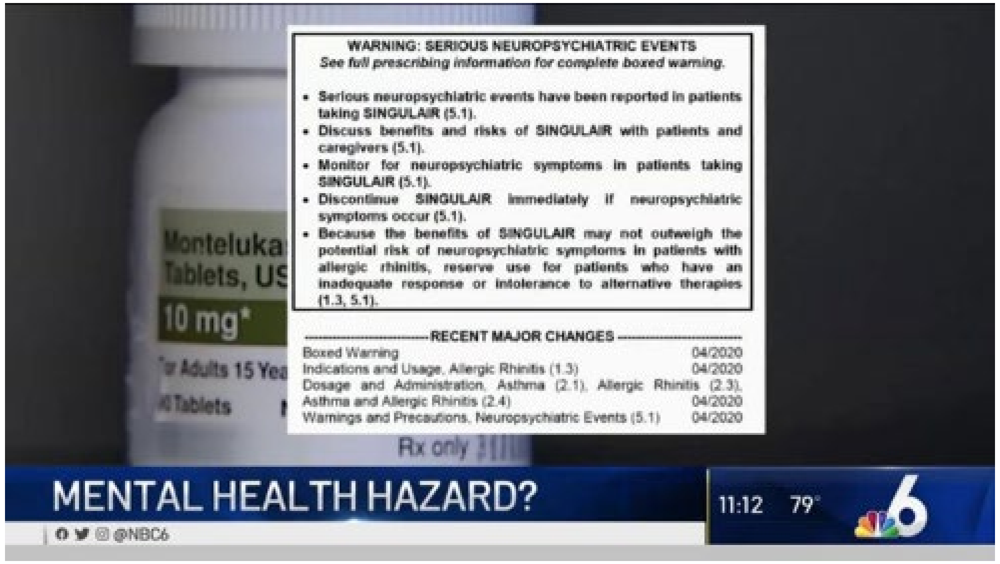
**Transportability**

of trained and tuned NLP models in 2 external EHR datasets

# Case for Action

**Montelukast**, a leukotriene-modifying agent (LTMA) is **US guideline recommended** for the treatment of asthma for all ages

- FDA approval in 1998
- In 2008 FDA warned of reports of suicidality and neuropsychiatric event associated with montelukast
- In **2020** FDA issues a *Boxed Warning* of neuropsychiatric adverse events based on expert panel determination as RWE was equivocal
- Sansing-Foster et al 2021 (Claims; Sentinel)
- Paljarvi et al 2022 (EHR)



**WARNING: SERIOUS NEUROPSYCHIATRIC EVENTS**

***See full prescribing information for complete boxed warning.***

- **Serious neuropsychiatric events have been reported in patients taking SINGULAIR (5.1).**
- **Discuss benefits and risks of SINGULAIR with patients and caregivers (5.1).**
- **Monitor for neuropsychiatric symptoms in patients taking SINGULAIR (5.1).**
- **Discontinue SINGULAIR immediately if neuropsychiatric symptoms occur (5.1).**
- **Because the benefits of SINGULAIR may not outweigh the potential risk of neuropsychiatric symptoms in patients with allergic rhinitis, reserve use for patients who have an inadequate response or intolerance to alternative therapies (1.3, 5.1).**

**RECENT MAJOR CHANGES**

Boxed Warning 04/2020
Indications and Usage, Allergic Rhinitis (1.3) 04/2020
Dosage and Administration, Asthma (2.1), Allergic Rhinitis (2.3), Asthma and Allergic Rhinitis (2.4) 04/2020
Warnings and Precautions, Neuropsychiatric Events (5.1) 04/2020

## Value

of using claims and EHR (structured/semi-structured/unstructured)

# ***MOSAIC-NLP***

*Study design:* Retrospective cohort study

*Study data:* EHR-claims linked data (2015-2022)

*Study cohort:* Patients with asthma newly initiating montelukast or inhaled corticosteroids

*Study outcomes:* Neuropsychiatric events

| Data source<br>Study stage | Cohort | Covariates | Outcomes |
|---|---|---|---|
| Study 1 | EHR-s/us + claims | Claims | Claims |
| Study 2 | EHR-s/us + claims | EHR-s + claims | EHR-s + claims |
| Study 3 | EHR-s/us + claims | EHR-s/us + claims | EHR-s/us + claims |

## Scalability

of an NLP model for clinical notes across the Oracle EHR RWD 100+ healthcare systems

# *MOSAIC-NLP*

*Study cohort:* 109,076 patients

*Healthcare systems:* 119

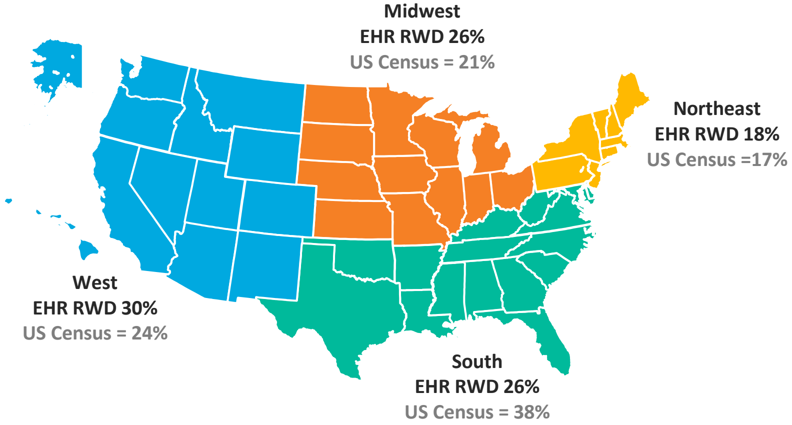
*Clinical notes:* 17+ million

## EHR Oracle RWD

**105 million patients**

**LNH** member healthcare systems

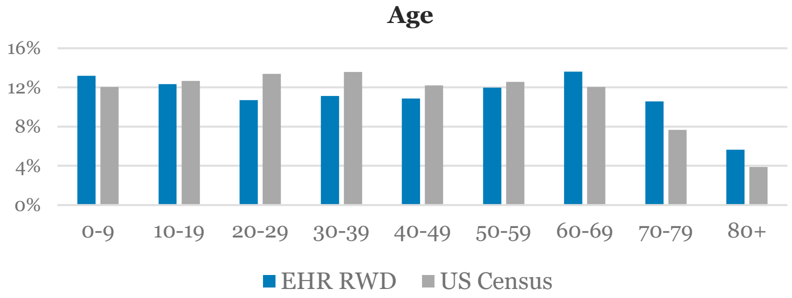
- ➢ Pediatric hospitals
- ➢ Critical access hospitals
- ➢ IDN
- ➢ Acute care hospitals
- ➢ Physician groups



**125M**
emergency encounters

**56M**
inpatient encounters

**972M**
outpatient encounters



## Claims

- **200 million patients**
- **Closed** medical and pharmacy claims
  - ➢ Commercial
  - ➢ Medicare Advantage
  - ➢ Medicaid Managed Care
- **National** representation

# Considerations for NLP Entity Extraction at Scale

## De-identification of notes

- Acceptable level of de-identification
- Separate workspace

## Training set

- Sampling frame – healthcare system, age, note type

## Entity identification

- Outcomes – boxed warning
- Covariates
- Rare entities/events
- Questionnaires (semi-structured data)

# Considerations When Creating Training Dataset for Annotation

## Healthcare system

Cannot assume EHR features are similar across healthcare systems or facilities

- Copy-pasting in notes
- Templates
- ‘Required’ fields
- Use of EHR platform for note taking
- Use of decimal points

## Age group

Treatment and care differ for children and adults

- Diagnoses
- Symptoms
- Concerns
- Treatment

## Note type

Variability between note type content and value

- Facility (ER vs clinic)
- Physician type (psychiatrist vs GP)
- Discharge note vs progress note…

# Neuropsychiatric Events

## *FDA's Boxed Warning*

- Agitation, including aggressive behavior or hostility
- Attention problems
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

## Structured Data

**Hospitalization/ER**

*OR*

**Diagnosis *AND* Treatment**

□ Depression
□ Self harm
□ Psychotic disorder
□ Mood disorder
□ Anxiety disorder
□ OCD
□ Manic or bipolar disorder
□ Personality disorder
□ Hyperactivity or aggressive behavior or harm

**Treatment for sleep disorder diagnosis**

□ Insomnia
□ Hypersomnia
□ Circadian rhythm disorder
□ Parasomnia
□ Movement disorder
□ Other undefined sleep disorder

## Unstructured Data

□ Aggressive behavior or hostility
□ Agitation
□ Attention problems
□ Bad or vivid dreams
□ Depression
□ Disorientation or confusion
□ Dream abnormalities
□ Feeling anxious
□ Hallucinations
□ Irritability
□ Memory problems
□ Obsessive-compulsive symptoms
□ Restlessness
□ Sleepwalking
□ Stuttering
□ Suicidal thoughts and actions
□ Tremor or shakiness
□ Trouble sleeping
□ Uncontrolled muscle movements

# Taxonomy: 54 Named Entities

### Drugs

- MON (montelukast)
- ICS (inhaled corticosteroids)
- Antidepressant
- Oral corticosteroid
- Short-acting beta-agonists (SABA)
- other drugs

### Sleep Disorders

- Insomnia
- Hypersomnia
- Circadian rhythm disorder
- Parasomnia
- Movement disorder
- Othersleep disorder

### SDOH

- Education
- Employment
- Exercise
- Lives alone

### Neuropsychiatric Symptoms and Disorders

- Aggressive behavior or hostility
- Agitation
- Attention problems
- Bad or vivid dreams
- Confusion/ Disorientation
- Depression
- Dream abnormalities
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive disorder or symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Completed suicide
- Suicide attempt
- Suicidal ideation
- Self-harm
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements
- Other psychiatric disorder

### Current Health Status

- Psychotherapy
- Alcohol use
- Marijuana use
- Substance abuse

### Test Results

- FEV1 %
- FEV1 % result

### Utilization

- Emergency visit
- Hospitalization
- Psychotherapy

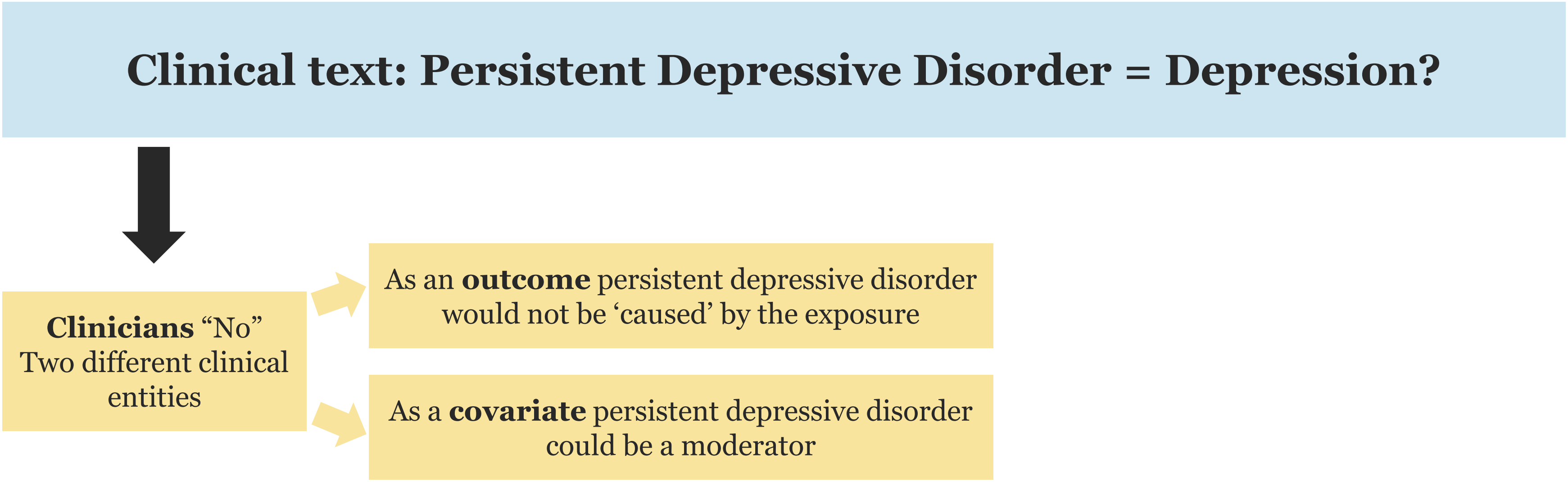
### Respiratory Symptoms/Disease

- Cough
- Snoring
- Asthma
- COPD
- Allergic rhinitis

### Disease Severity and Control

- Intermittent
- Mild
- Moderate
- Severe
- Alleviated
- Worsened
- Other Modifier

# Entity – Example of Decisions



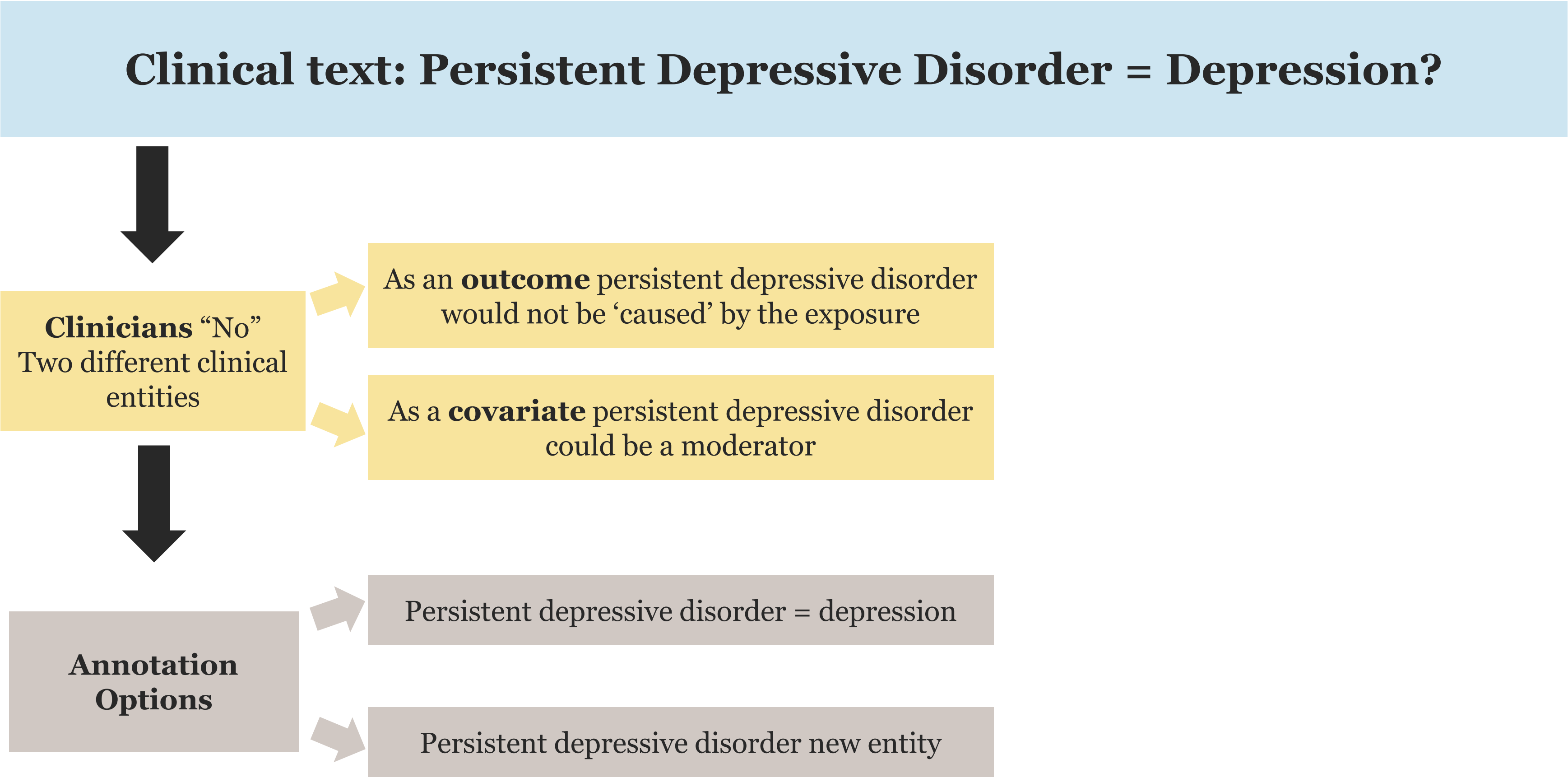
**Clinical text: Persistent Depressive Disorder = Depression?**

**Clinicians** “No”
Two different clinical entities

As an **outcome** persistent depressive disorder would not be ‘caused’ by the exposure

As a **covariate** persistent depressive disorder could be a moderator

# Entity – Example of Decisions



**Clinical text: Persistent Depressive Disorder = Depression?**

**Clinicians** "No"
Two different clinical entities

As an **outcome** persistent depressive disorder would not be 'caused' by the exposure

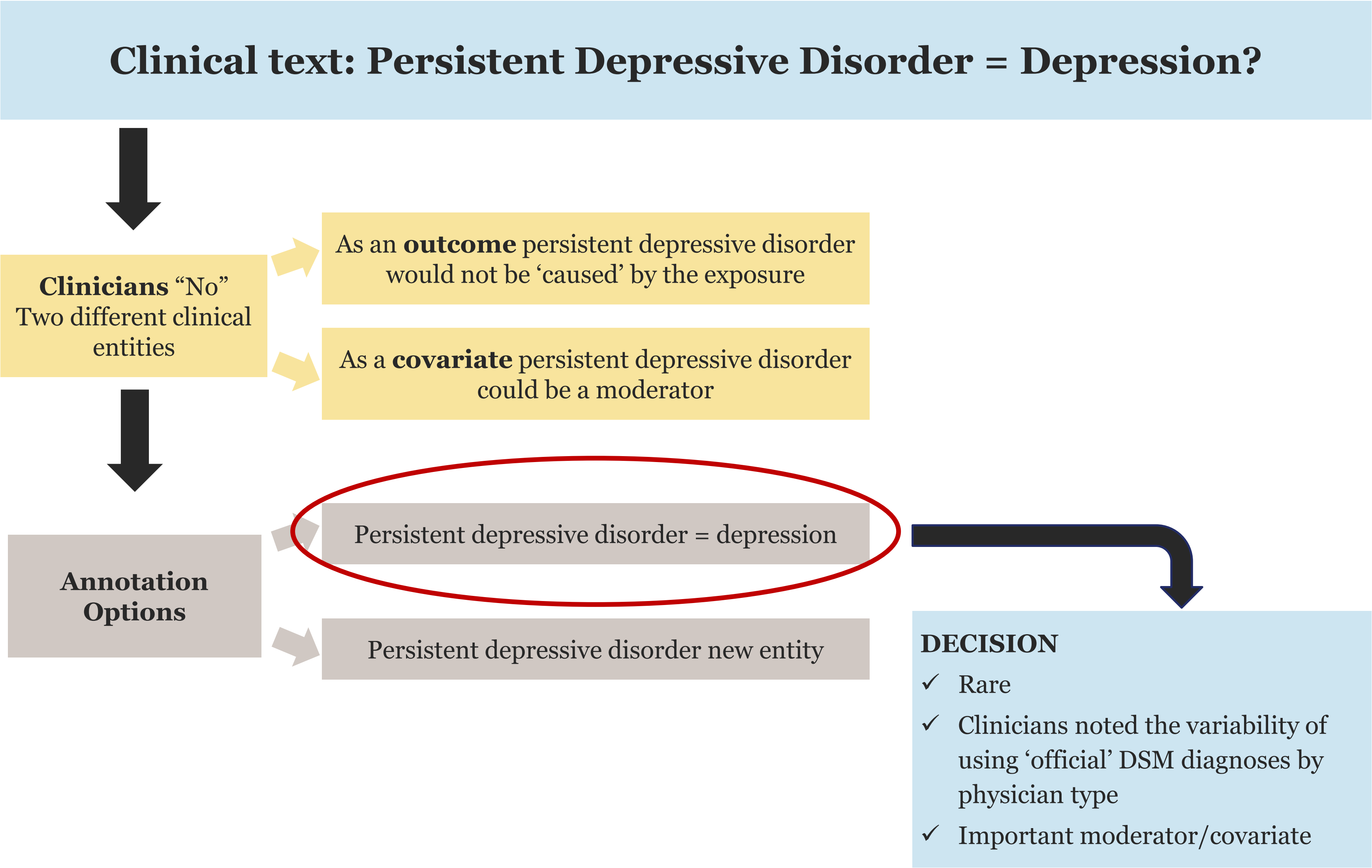
As a **covariate** persistent depressive disorder could be a moderator

**Annotation Options**

Persistent depressive disorder = depression

Persistent depressive disorder new entity

# Entity – Example of Decisions



**Clinical text: Persistent Depressive Disorder = Depression?**

**Clinicians** “No”
Two different clinical entities

As an **outcome** persistent depressive disorder would not be ‘caused’ by the exposure

As a **covariate** persistent depressive disorder could be a moderator

**Annotation Options**

Persistent depressive disorder = depression

Persistent depressive disorder new entity

**DECISION**

- ✓ Rare
- ✓ Clinicians noted the variability of using ‘official’ DSM diagnoses by physician type
- ✓ Important moderator/covariate

# Summary

- To our knowledge this is the first pharmacoepidemiology study to use linked EHR-claims data and extract semi/unstructured data at scale
- Methodology requires considerations related to the high degree of heterogeneity in the clinical notes
- Gather and use experts to build the NLP model:
  - ✓ NLP experts
  - ✓ Biostatisticians
  - ✓ Clinicians (subject matter experts)
  - ✓ Epidemiologists

# MOSAIC-NLP

## Technical Aspects and Lessons Learned

Hasham Ul Haq

John Snow Labs

hasham@johnsnowlabs.com

# Technology Stack & Rationale

Multiple methods of dealing with the problem

- Self/Unsupervised models
  - LLMs (ChatGPT/Llama)
    - Q&A approach
    - Prompt Engineering
    - Few-Shot Approach

- Less / no training required – high generalization
- Easy to setup & use

- May not work as well for specific use-cases.
- Much more costly/difficult to train – if required

- DL – supervised approach
  - NER models - BiLSTM

- Easy to train and adapt to use-cases.
- Comparable performance on specified use-cases.
- Computationally efficient.

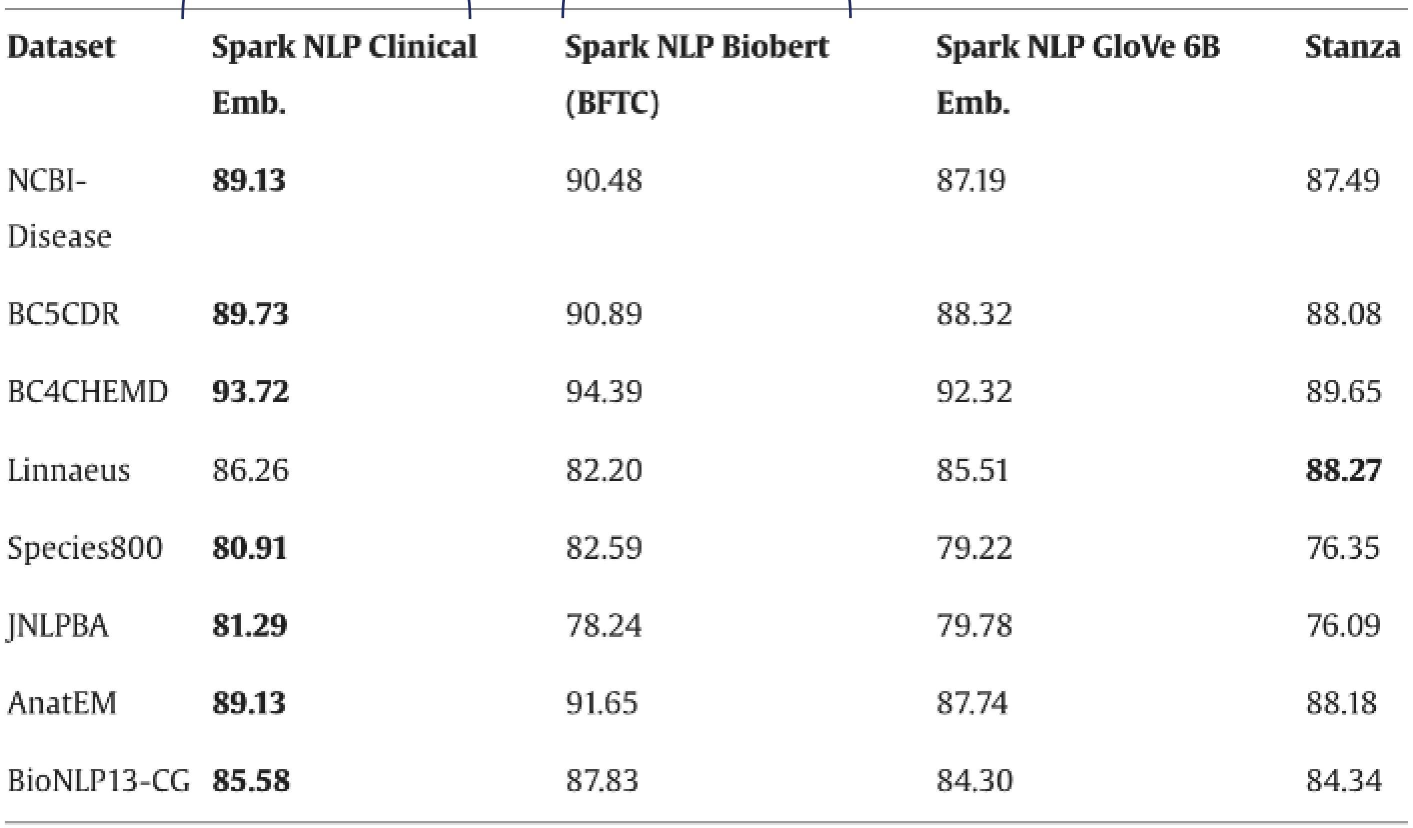
- Training is required – low generalization
- Bigger models may outperform

# Technology Stack & Rationale

- Named Entity Recognition Models
- Transformers based models – latest

Bi-LSTM

Transformers



| Dataset | Spark NLP Clinical Emb. | Spark NLP Biobert (BFTC) | Spark NLP GloVe 6B Emb. | Stanza |
|---|---|---|---|---|
| NCBI-Disease | **89.13** | 90.48 | 87.19 | 87.49 |
| BC5CDR | **89.73** | 90.89 | 88.32 | 88.08 |
| BC4CHEMD | **93.72** | 94.39 | 92.32 | 89.65 |
| Linnaeus | 86.26 | 82.20 | 85.51 | **88.27** |
| Species800 | **80.91** | 82.59 | 79.22 | 76.35 |
| JNLPBA | **81.29** | 78.24 | 79.78 | 76.09 |
| AnatEM | **89.13** | 91.65 | 87.74 | 88.18 |
| BioNLP13-CG | **85.58** | 87.83 | 84.30 | 84.34 |

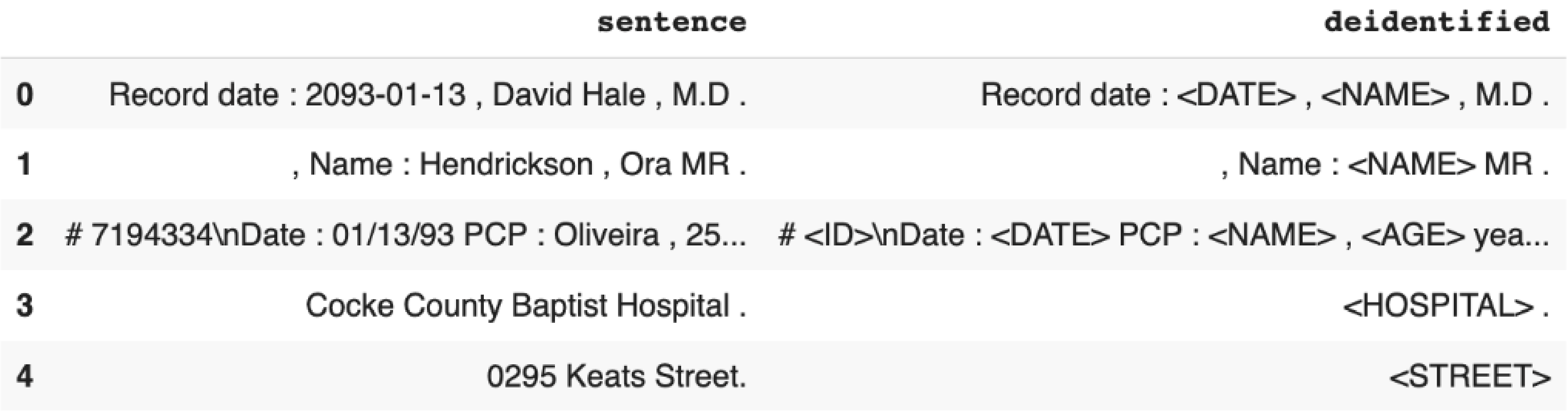
# Factors to Consider While Choosing ML Model Architecure

- How many documents to process?
- What type of hardware resources are available?
- What is a feasible total runtime?

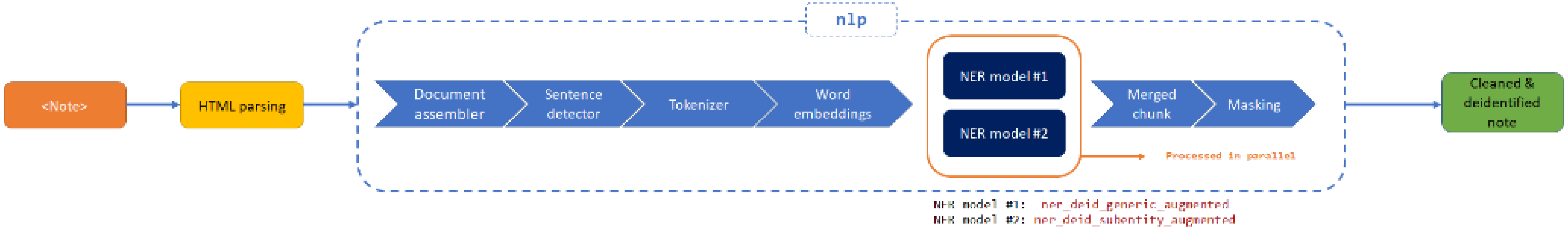
- The end goal is to process Tens of Millions of records.
- Avoid high costs of GPUs
  - Expensive to scale compared to CPUs.

- Develop efficient models that are performant in terms of memory and CPU utilization, while delivering comparable performance.

# First Step: De-Identification of Documents



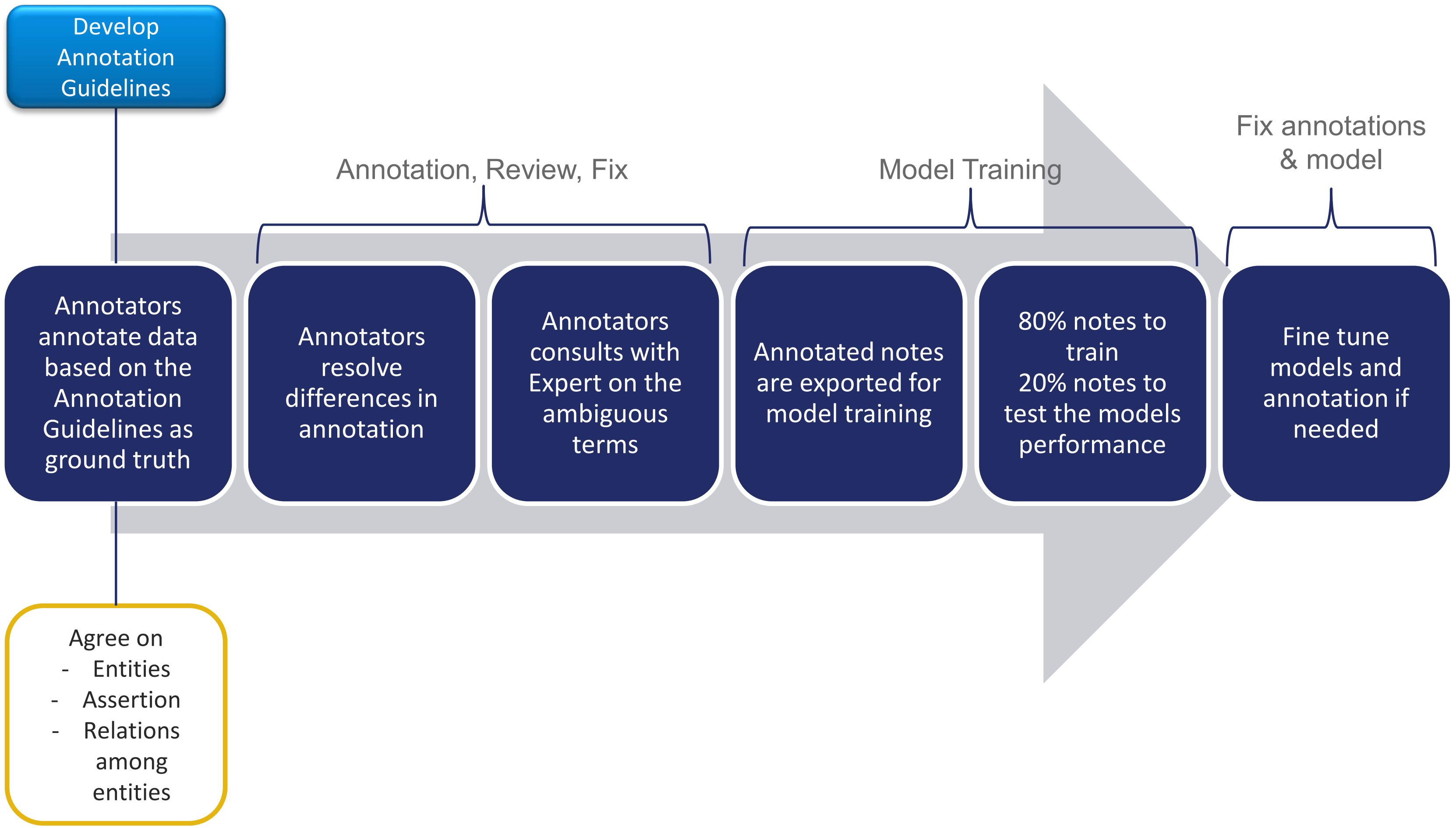
| | sentence | deidentified |
|---|---|---|
| **0** | Record date : 2093-01-13 , David Hale , M.D . | Record date : <DATE> , <NAME> , M.D . |
| **1** | , Name : Hendrickson , Ora MR . | , Name : <NAME> MR . |
| **2** | # 7194334\nDate : 01/13/93 PCP : Oliveira , 25... | # <ID>\nDate : <DATE> PCP : <NAME> , <AGE> yea... |
| **3** | Cocke County Baptist Hospital . | <HOSPITAL> . |
| **4** | 0295 Keats Street. | <STREET> |



# De-Identification - Evaluation

- Total Notes: 100 – randomly selected
- Occurrences of sensitive information: 1967
  - Name, Address, Date etc..
- Recall (sensitivity) = 93.54%

# NLP Process Overview



Develop Annotation Guidelines

Annotation, Review, Fix

Model Training

Fix annotations & model

Annotators annotate data based on the Annotation Guidelines as ground truth

Annotators resolve differences in annotation

Annotators consults with Expert on the ambiguous terms

Annotated notes are exported for model training

80% notes to train
20% notes to test the models performance

Fine tune models and annotation if needed

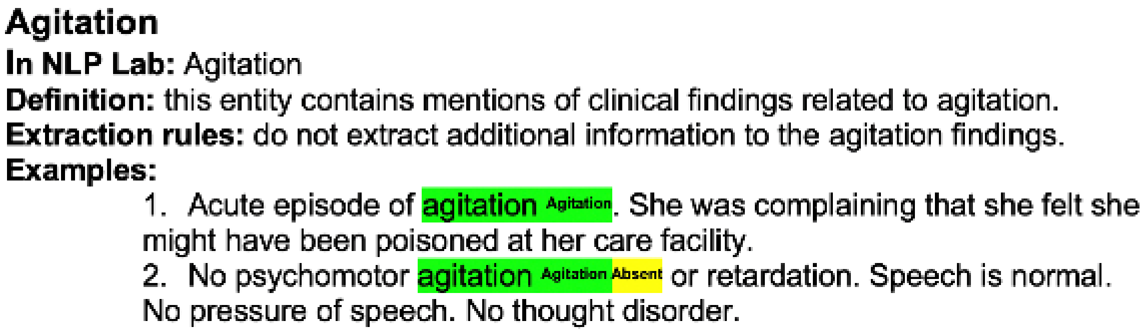
Agree on
- Entities
- Assertion
- Relations among entities

# Annotation Guidelines

54 **Named Entities**: Word or series of words that refer to a specific concept

8 **Assertions**: indicates an attribute of an entity

- **Present**, Past, Absent, Family_history, someone_else, possible, planned, hypothetical



**Agitation**

**In NLP Lab:** Agitation

**Definition:** this entity contains mentions of clinical findings related to agitation.

**Extraction rules:** do not extract additional information to the agitation findings.

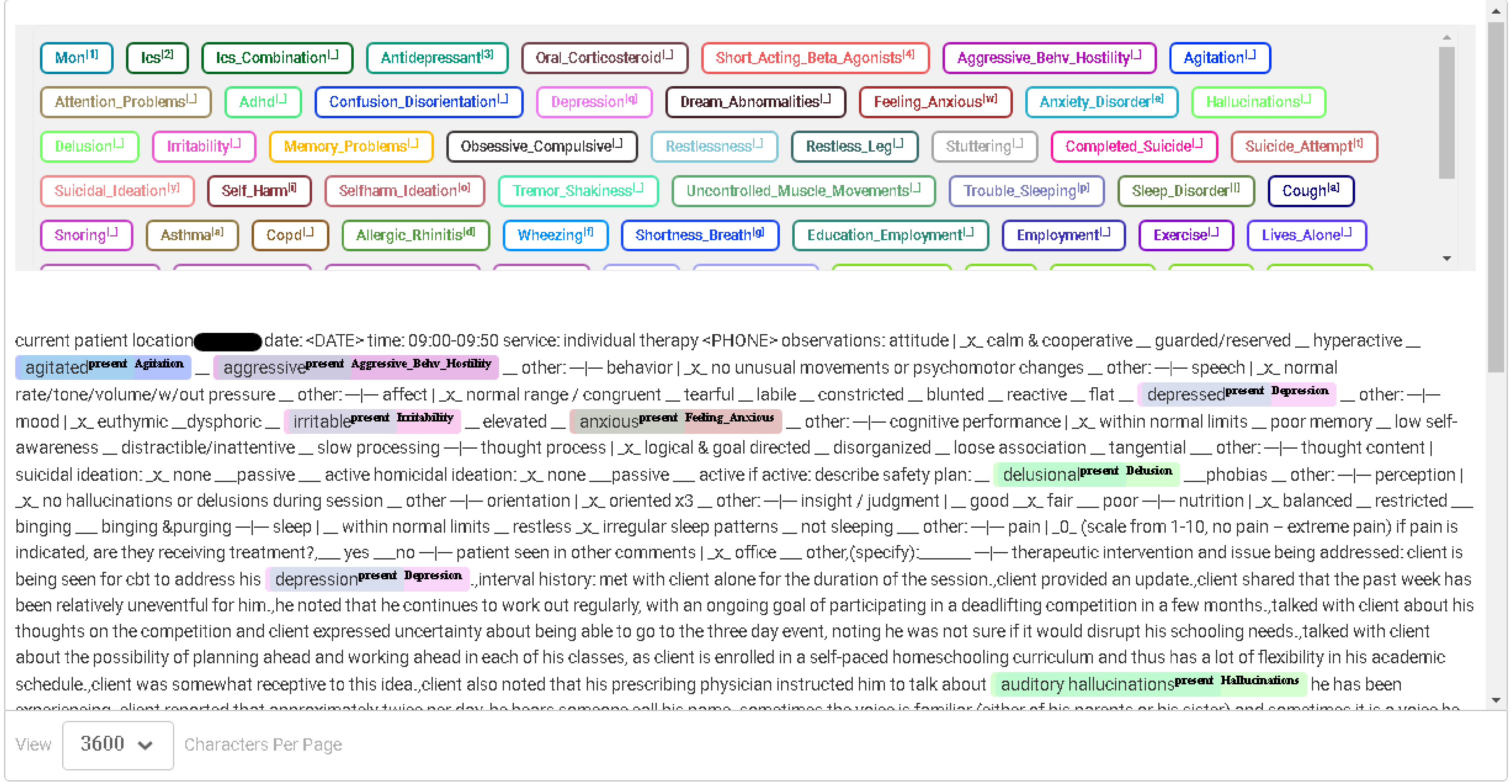
**Examples:**

1. Acute episode of agitation Agitation. She was complaining that she felt she might have been poisoned at her care facility.
2. No psychomotor agitation Agitation Absent or retardation. Speech is normal. No pressure of speech. No thought disorder.

An Entity desrcibed in the Annotation Guideline

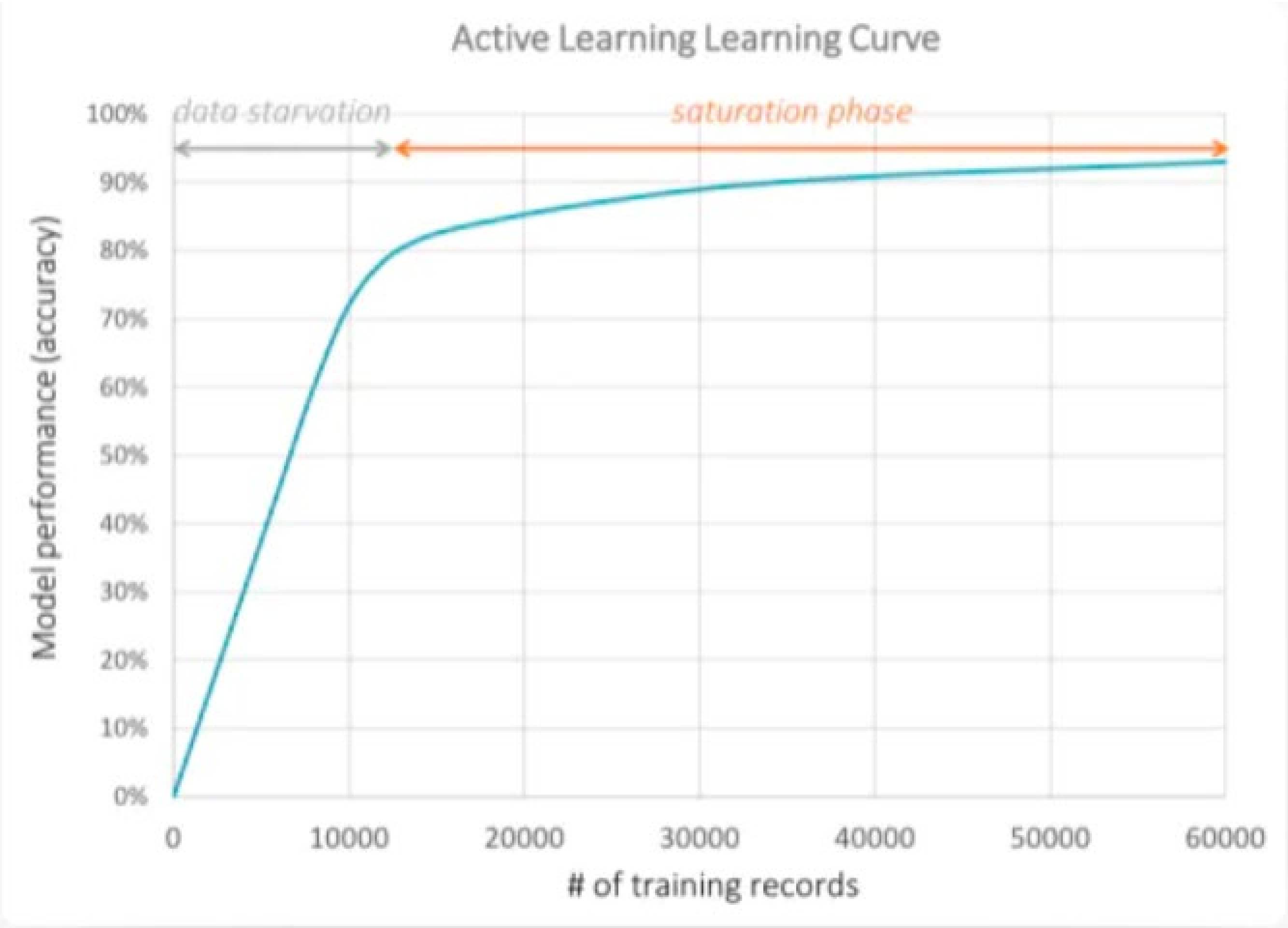
# Annotations

- Annotation in NLP Lab

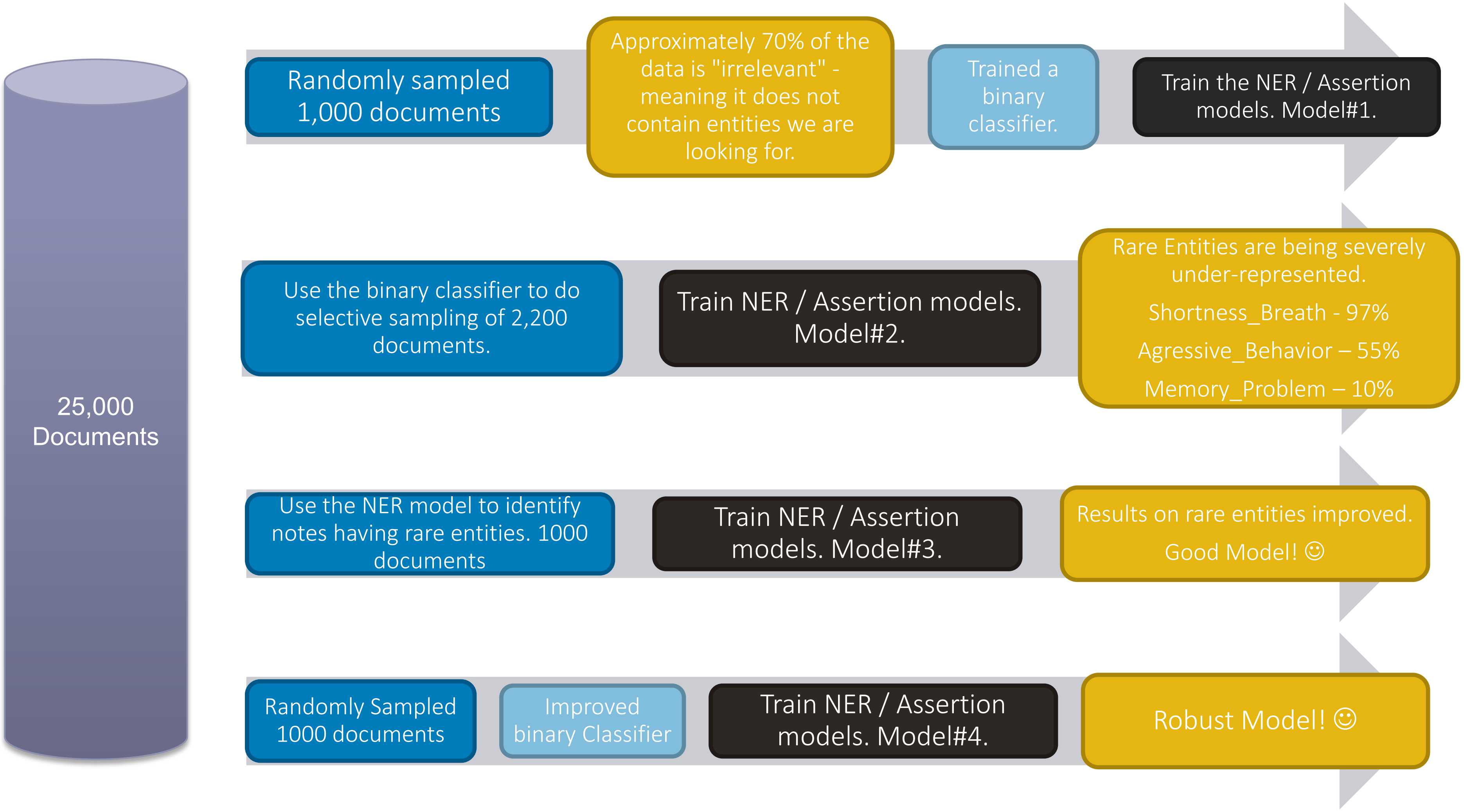


current patient location[redacted]date: <DATE> time: 09:00-09:50 service: individual therapy <PHONE> observations: attitude | _x_ calm & cooperative __ guarded/reserved __ hyperactive __ agitated present Agitation __ aggressive present Aggressive_Behv_Hostility __ other: —|— behavior | _x_ no unusual movements or psychomotor changes __ other: —|— speech | _x_ normal rate/tone/volume/w/out pressure __ other: —|— affect | _x_ normal range / congruent __ tearful __ labile __ constricted __ blunted __ reactive __ flat __ depressed present Depression __ other: —|— mood | _x_ euthymic __dysphoric __ irritable present Irritability __ elevated __ anxious present Feeling_Anxious __ other: —|— cognitive performance | _x_ within normal limits __ poor memory __ low self-awareness __ distractible/inattentive __ slow processing —|— thought process | _x_ logical & goal directed __ disorganized __ loose association __ tangential ___ other: —|— thought content | suicidal ideation: _x_ none ___passive ___ active homicidal ideation: _x_ none ___passive ___ active if active: describe safety plan: __ delusional present Delusion ___phobias __ other: —|— perception | _x_ no hallucinations or delusions during session __ other —|— orientation | _x_ oriented x3 __ other: —|— insight / judgment | __ good __x_ fair ___ poor —|— nutrition | _x_ balanced __ restricted ___ binging ___ binging &purging —|— sleep | __ within normal limits __ restless _x_ irregular sleep patterns __ not sleeping ___ other: —|— pain | _0_ (scale from 1-10, no pain – extreme pain) if pain is indicated, are they receiving treatment?,___ yes ___no —|— patient seen in other comments | _x_ office ___ other,(specify):_______ —|— therapeutic intervention and issue being addressed: client is being seen for cbt to address his depression present Depression .,interval history: met with client alone for the duration of the session.,client provided an update.,client shared that the past week has been relatively uneventful for him.,he noted that he continues to work out regularly, with an ongoing goal of participating in a deadlifting competition in a few months.,talked with client about his thoughts on the competition and client expressed uncertainty about being able to go to the three day event, noting he was not sure if it would disrupt his schooling needs.,talked with client about the possibility of planning ahead and working ahead in each of his classes, as client is enrolled in a self-paced homeschooling curriculum and thus has a lot of flexibility in his academic schedule.,client was somewhat receptive to this idea.,client also noted that his prescribing physician instructed him to talk about auditory hallucinations present Hallucinations he has been experiencing.,client reported that approximately twice per day he hears someone call his name, sometimes the voice is familiar (either of his parents or his sister) and sometimes it is a voice he

# How Much Data to Annotate?



# NLP Training and Evaluation Process



25,000 Documents

Randomly sampled 1,000 documents

Approximately 70% of the data is "irrelevant" - meaning it does not contain entities we are looking for.

Trained a binary classifier.

Train the NER / Assertion models. Model#1.

Use the binary classifier to do selective sampling of 2,200 documents.

Train NER / Assertion models. Model#2.

Rare Entities are being severely under-represented.

Shortness_Breath - 97%

Agressive_Behavior – 55%

Memory_Problem – 10%

Use the NER model to identify notes having rare entities. 1000 documents

Train NER / Assertion models. Model#3.

Results on rare entities improved.

Good Model! ☺

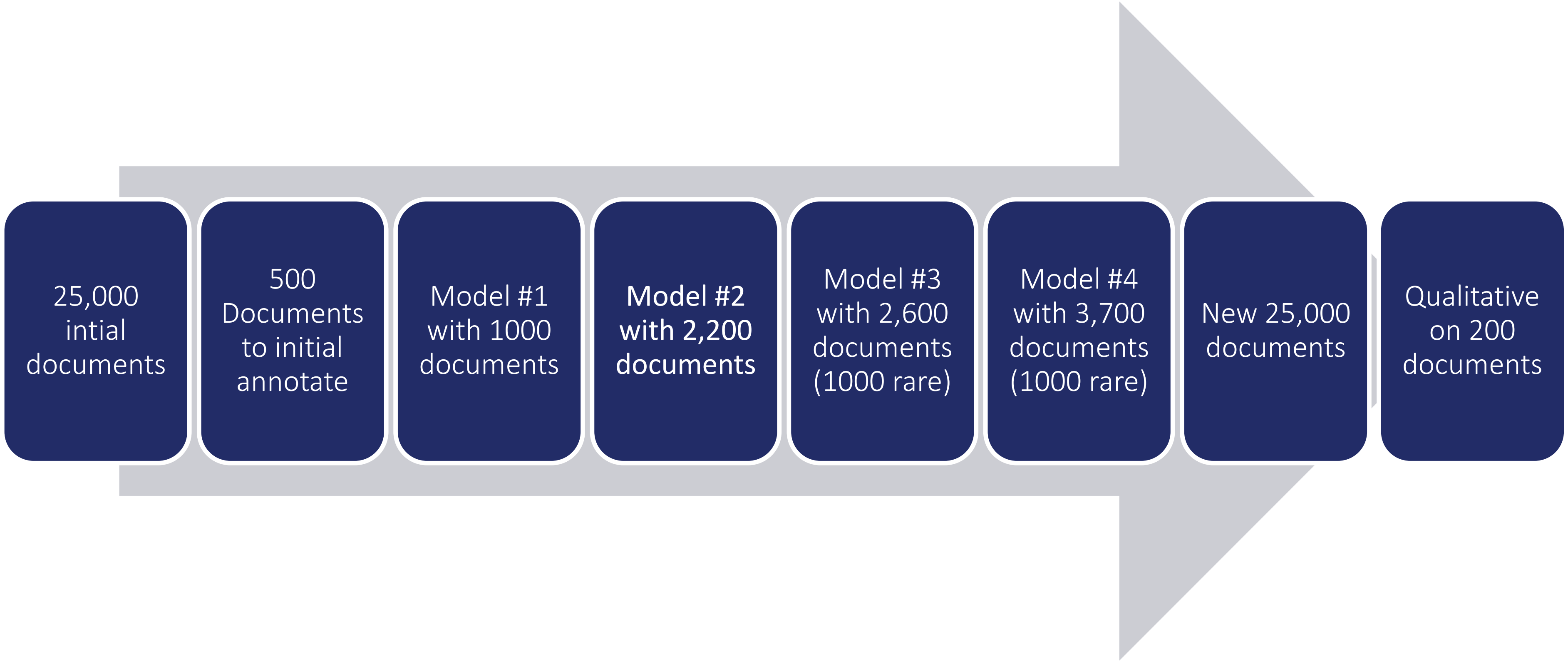
Randomly Sampled 1000 documents

Improved binary Classifier

Train NER / Assertion models. Model#4.

Robust Model! ☺

# NLP Training and Evaluation Process



25,000 intial documents

500 Documents to initial annotate

Model #1 with 1000 documents

**Model #2 with 2,200 documents**

Model #3 with 2,600 documents (1000 rare)

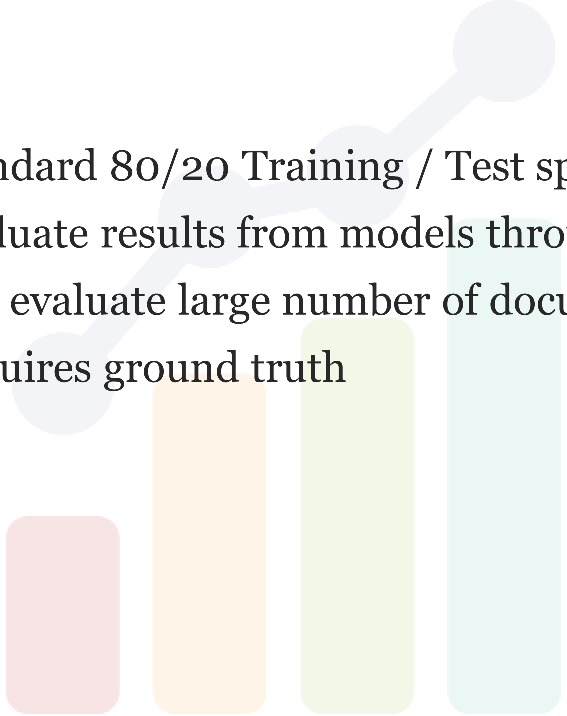
Model #4 with 3,700 documents (1000 rare)

New 25,000 documents

Qualitative on 200 documents

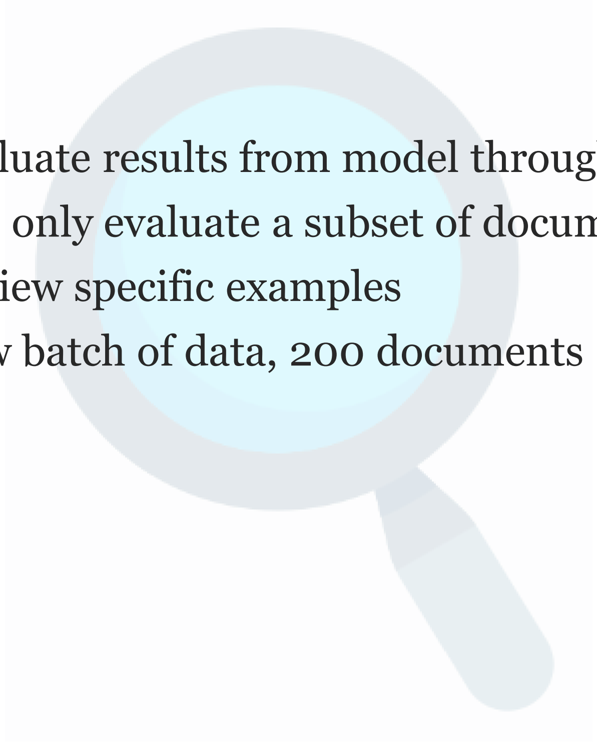
# Quantitative vs Qualitative Evaluation

## Quantitative



- Standard 80/20 Training / Test split.
- Evaluate results from models through metrics
- Can evaluate large number of documents
- Requires ground truth

## Qualitative



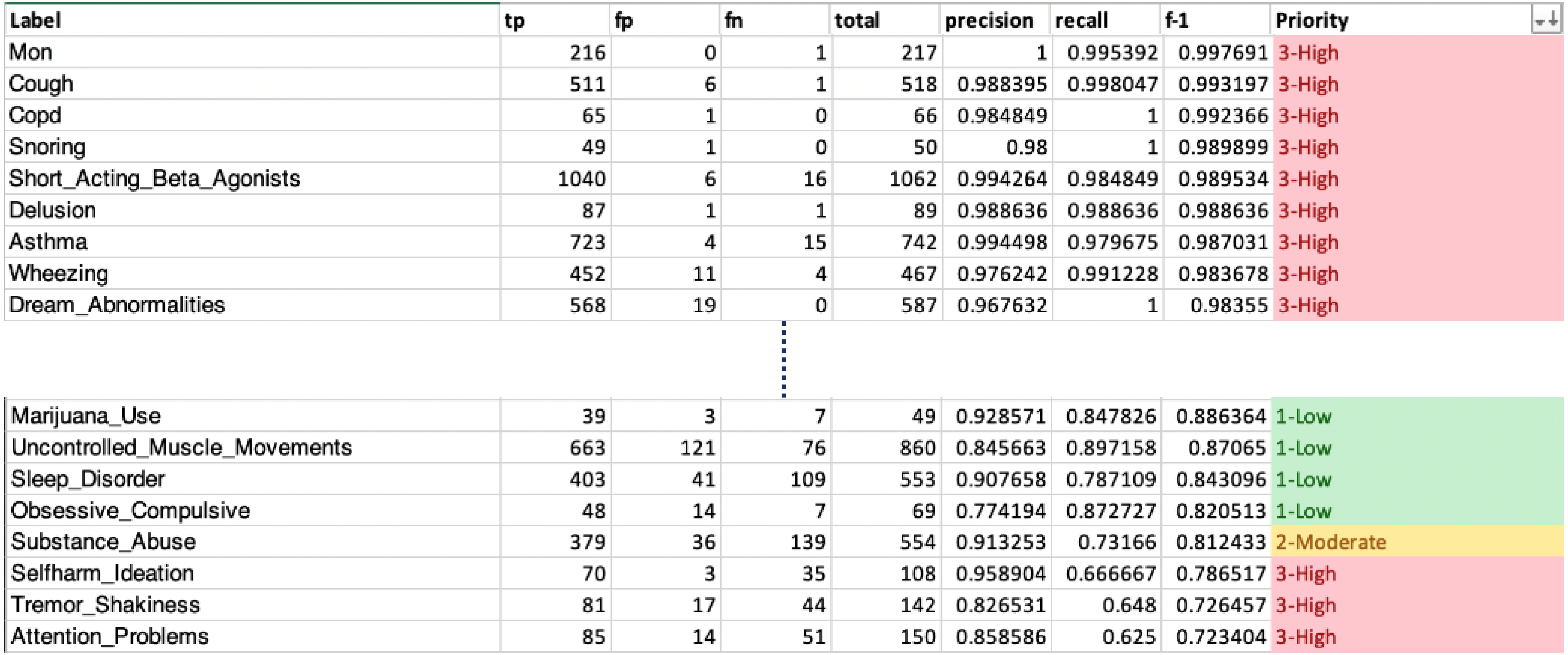
- Evaluate results from model through SME
- Can only evaluate a subset of documents
- Review specific examples
- New batch of data, 200 documents

# NER Quantitative Model Results

| Date | Micro f-1 | Macro f-1 | NER Label under f-1 0.80 |
|---|---|---|---|
| 15-May | 0.832 | 0.559 | 32 |
| 22-Jun | 0.912 | 0.698 | 21 |
| 3-Jul | 0.932 | 0.802 | 10 |
| 24-Jul | 0.935 | 0.828 | 7 |

# Examples of NER Label Accuracies

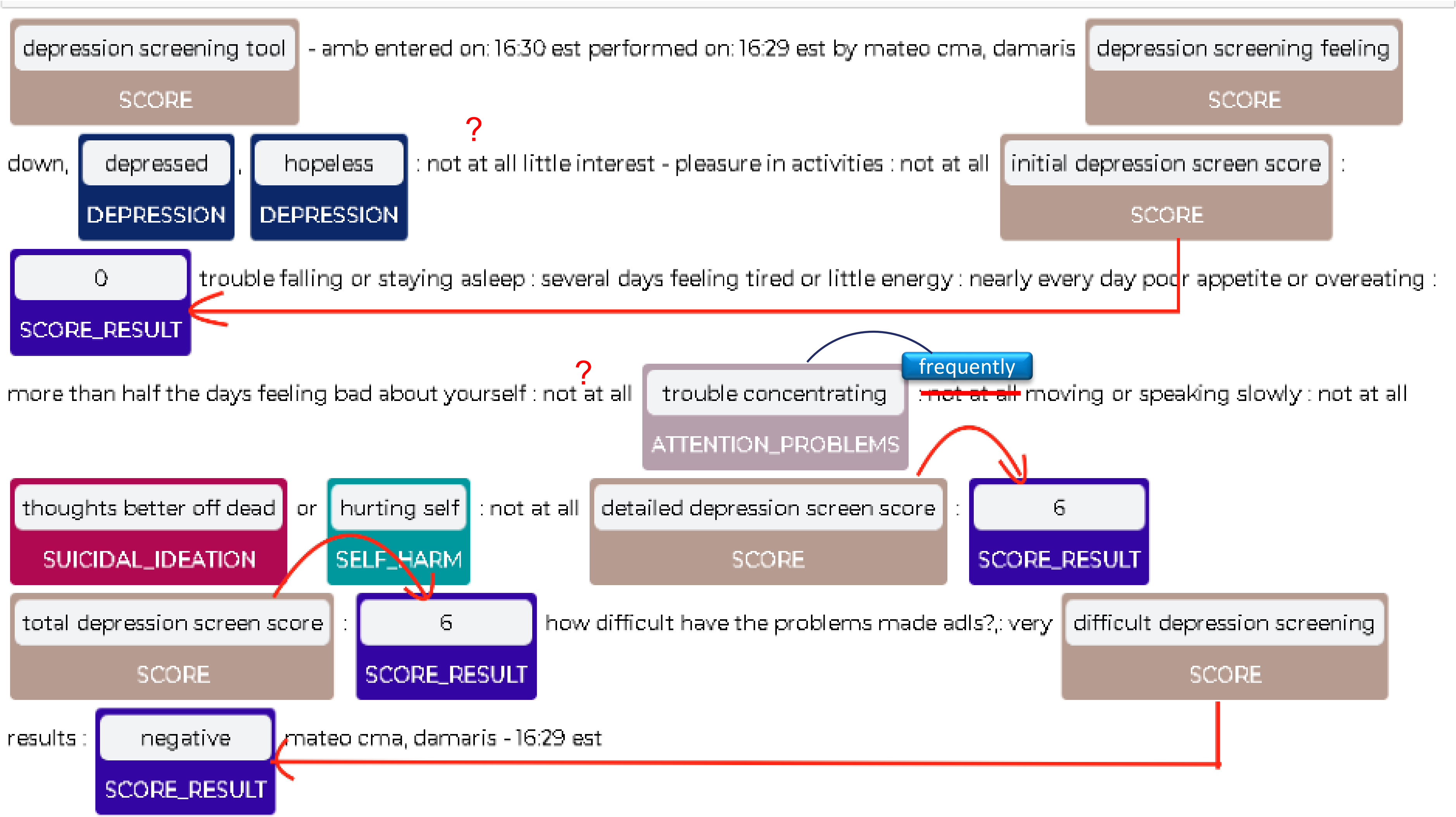
Original taxonomy included stuttering, but we had too few mentions in the notes (8 mentions in 25K notes)



| Label | tp | fp | fn | total | precision | recall | f-1 | Priority |
|---|---|---|---|---|---|---|---|---|
| Mon | 216 | 0 | 1 | 217 | 1 | 0.995392 | 0.997691 | 3-High |
| Cough | 511 | 6 | 1 | 518 | 0.988395 | 0.998047 | 0.993197 | 3-High |
| Copd | 65 | 1 | 0 | 66 | 0.984849 | 1 | 0.992366 | 3-High |
| Snoring | 49 | 1 | 0 | 50 | 0.98 | 1 | 0.989899 | 3-High |
| Short_Acting_Beta_Agonists | 1040 | 6 | 16 | 1062 | 0.994264 | 0.984849 | 0.989534 | 3-High |
| Delusion | 87 | 1 | 1 | 89 | 0.988636 | 0.988636 | 0.988636 | 3-High |
| Asthma | 723 | 4 | 15 | 742 | 0.994498 | 0.979675 | 0.987031 | 3-High |
| Wheezing | 452 | 11 | 4 | 467 | 0.976242 | 0.991228 | 0.983678 | 3-High |
| Dream_Abnormalities | 568 | 19 | 0 | 587 | 0.967632 | 1 | 0.98355 | 3-High |
| | | | ⋮ | | | | | |
| Marijuana_Use | 39 | 3 | 7 | 49 | 0.928571 | 0.847826 | 0.886364 | 1-Low |
| Uncontrolled_Muscle_Movements | 663 | 121 | 76 | 860 | 0.845663 | 0.897158 | 0.87065 | 1-Low |
| Sleep_Disorder | 403 | 41 | 109 | 553 | 0.907658 | 0.787109 | 0.843096 | 1-Low |
| Obsessive_Compulsive | 48 | 14 | 7 | 69 | 0.774194 | 0.872727 | 0.820513 | 1-Low |
| Substance_Abuse | 379 | 36 | 139 | 554 | 0.913253 | 0.73166 | 0.812433 | 2-Moderate |
| Selfharm_Ideation | 70 | 3 | 35 | 108 | 0.958904 | 0.666667 | 0.786517 | 3-High |
| Tremor_Shakiness | 81 | 17 | 44 | 142 | 0.826531 | 0.648 | 0.726457 | 3-High |
| Attention_Problems | 85 | 14 | 51 | 150 | 0.858586 | 0.625 | 0.723404 | 3-High |

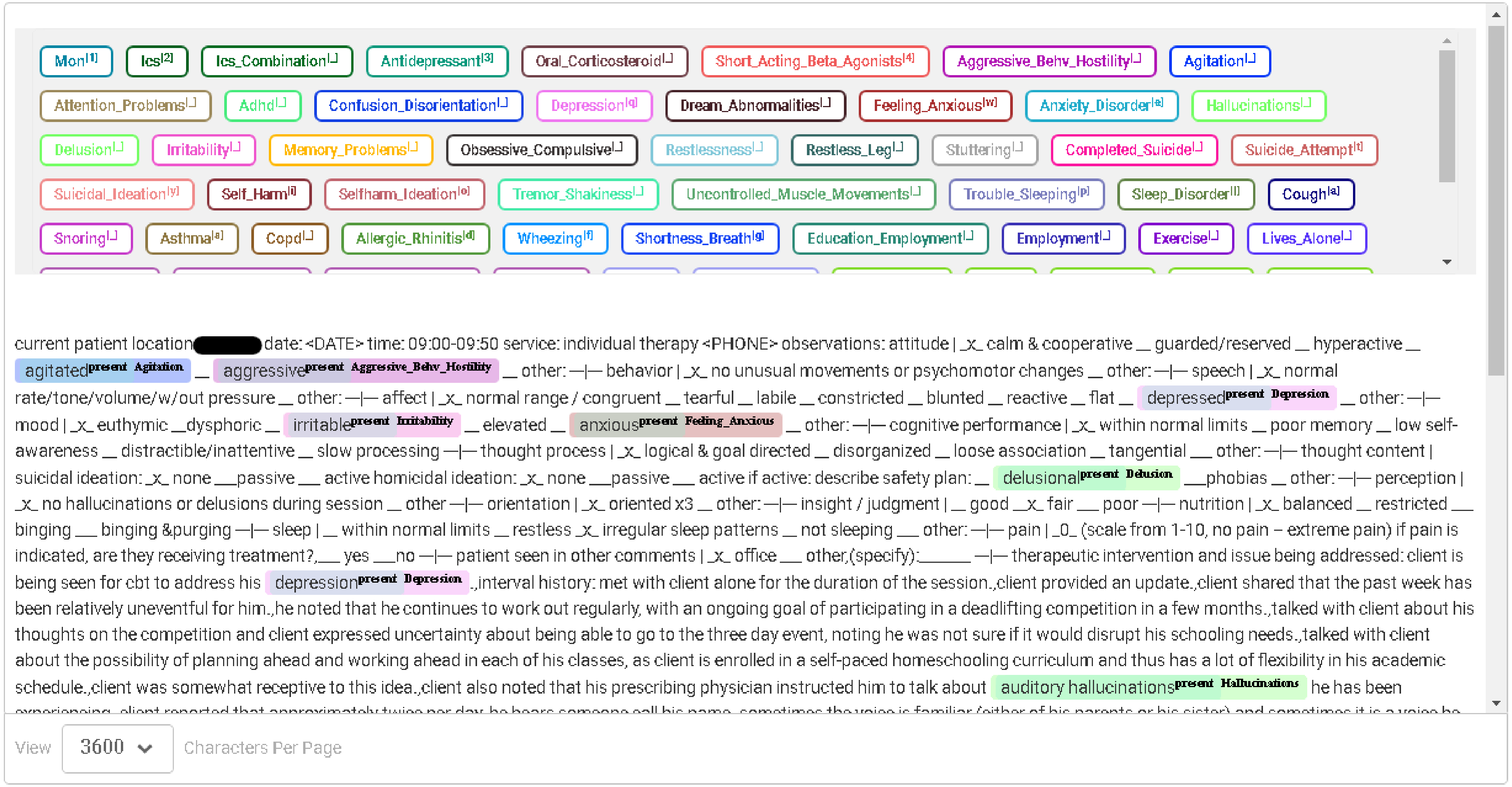
# The Data – A Single Example Says it All

## Semi-structured questionnaires in notes



depression screening tool - amb entered on: 16:30 est performed on: 16:29 est by mateo cma, damaris depression screening feeling down, depressed , hopeless : not at all little interest - pleasure in activities : not at all initial depression screen score : 0 trouble falling or staying asleep : several days feeling tired or little energy : nearly every day poor appetite or overeating : more than half the days feeling bad about yourself : not at all trouble concentrating ~~: not at all~~ moving or speaking slowly : not at all thoughts better off dead or hurting self : not at all detailed depression screen score : 6 total depression screen score : 6 how difficult have the problems made adls?,: very difficult depression screening results : negative mateo cma, damaris - 16:29 est

# Variation in How Questionnaires Show up in the Notes



current patient location [redacted] date: <DATE> time: 09:00-09:50 service: individual therapy <PHONE> observations: attitude | _x_ calm & cooperative __ guarded/reserved __ hyperactive __ agitated present Agitation __ aggressive present Aggressive_Behv_Hostility __ other: —|— behavior | _x_ no unusual movements or psychomotor changes __ other: —|— speech | _x_ normal rate/tone/volume/w/out pressure __ other: —|— affect | _x_ normal range / congruent __ tearful __ labile __ constricted __ blunted __ reactive __ flat __ depressed present Depression __ other: —|— mood | _x_ euthymic __dysphoric __ irritable present Irritability __ elevated __ anxious present Feeling_Anxious __ other: —|— cognitive performance | _x_ within normal limits __ poor memory __ low self-awareness __ distractible/inattentive __ slow processing —|— thought process | _x_ logical & goal directed __ disorganized __ loose association __ tangential ___ other: —|— thought content | suicidal ideation: _x_ none ___passive ___ active homicidal ideation: _x_ none ___passive ___ active if active: describe safety plan: __ delusional present Delusion ___phobias __ other: —|— perception | _x_ no hallucinations or delusions during session __ other —|— orientation | _x_ oriented x3 __ other: —|— insight / judgment | __ good __x_ fair ___ poor —|— nutrition | _x_ balanced __ restricted ___ binging ___ binging &purging —|— sleep | __ within normal limits __ restless _x_ irregular sleep patterns __ not sleeping ___ other: —|— pain | _0_ (scale from 1-10, no pain – extreme pain) if pain is indicated, are they receiving treatment?,___ yes ___no —|— patient seen in other comments | _x_ office ___ other,(specify):_______ —|— therapeutic intervention and issue being addressed: client is being seen for cbt to address his depression present Depression .,interval history: met with client alone for the duration of the session.,client provided an update.,client shared that the past week has been relatively uneventful for him.,he noted that he continues to work out regularly, with an ongoing goal of participating in a deadlifting competition in a few months.,talked with client about his thoughts on the competition and client expressed uncertainty about being able to go to the three day event, noting he was not sure if it would disrupt his schooling needs.,talked with client about the possibility of planning ahead and working ahead in each of his classes, as client is enrolled in a self-paced homeschooling curriculum and thus has a lot of flexibility in his academic schedule.,client was somewhat receptive to this idea.,client also noted that his prescribing physician instructed him to talk about auditory hallucinations present Hallucinations he has been

View 3600 Characters Per Page

# Lessons Learned

- Wide variety between EHR sites
- Structured forms being transferred to unstructured free-text makes NLP more difficult – unless done right!
- Annotation Guideline needs to be adaptive to new examples
- Constant communication between the annotators, the Subject Matter Experts, and the Data Scientist is necessary for building a good model

- For safety signals we are looking for rare events, but the fewer mentions of those events make them more challenging to capture; we need more notes to train models using those rare events.

# Questions?