Sentinel

Advances in Drug Safety Surveillance Infrastructure in the US FDA Sentinel

Rishi J. Desai, MS, PhD
Associate Professor of Medicine
Division of Pharmacoepidemiology and Pharmacoeconomics
Department of Medicine
Brigham and Women’s Hospital, Harvard Medical School, Boston

✉ rdesai@bwh.harvard.edu   @Rishidesai11

05/19/2025

Disclaimer

This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA). The views expressed are those of the author and not necessarily those of the US FDA.

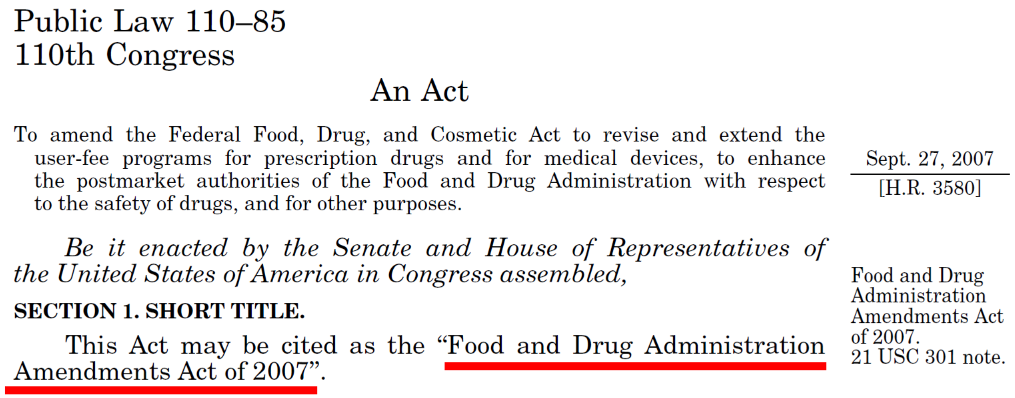
Agenda

01 **What is Sentinel?**

02 **Data Infrastructure: RWE –DE**

03 **Methodological Initiatives**

What is Sentinel?



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.

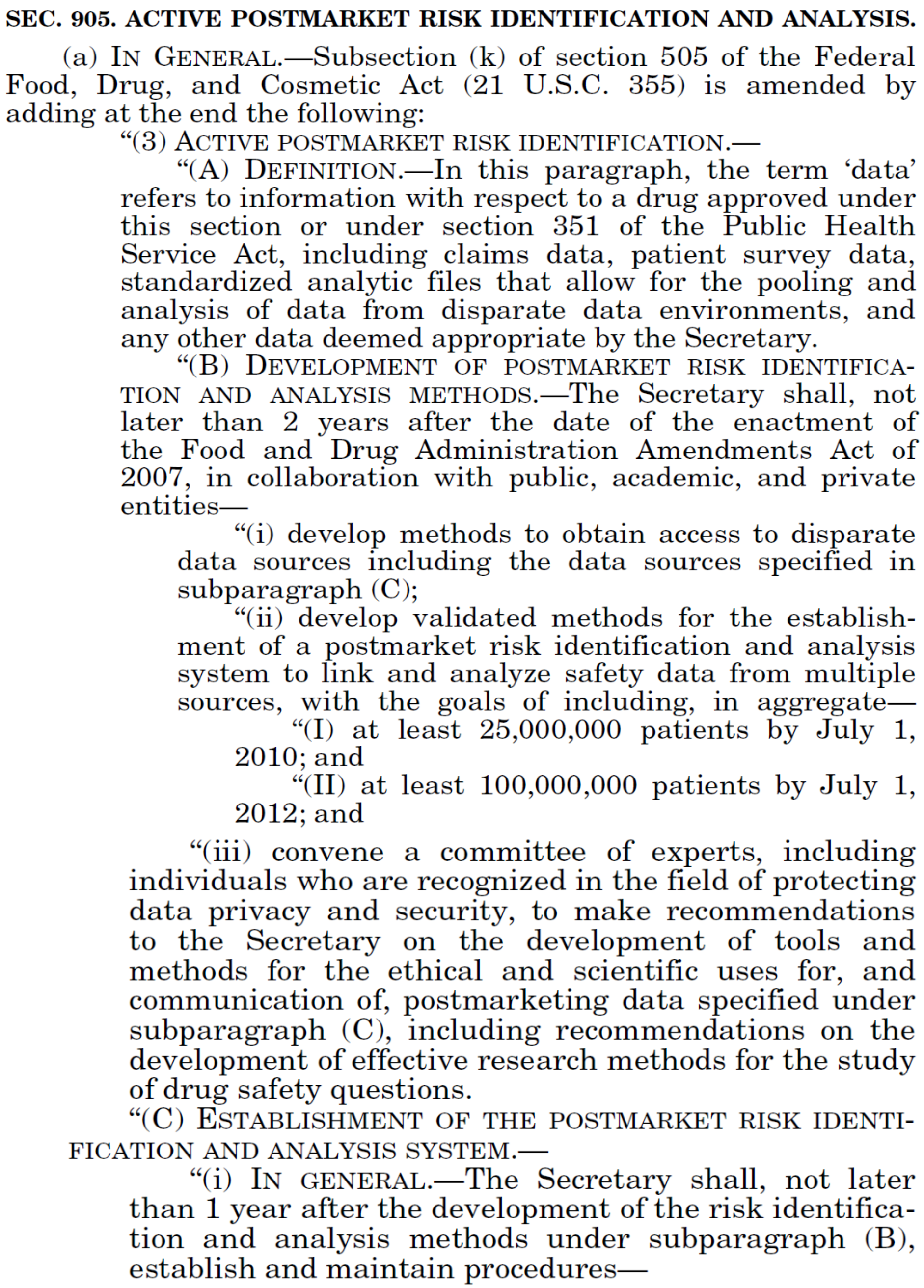
Sept. 27, 2007
[H.R. 3580]

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

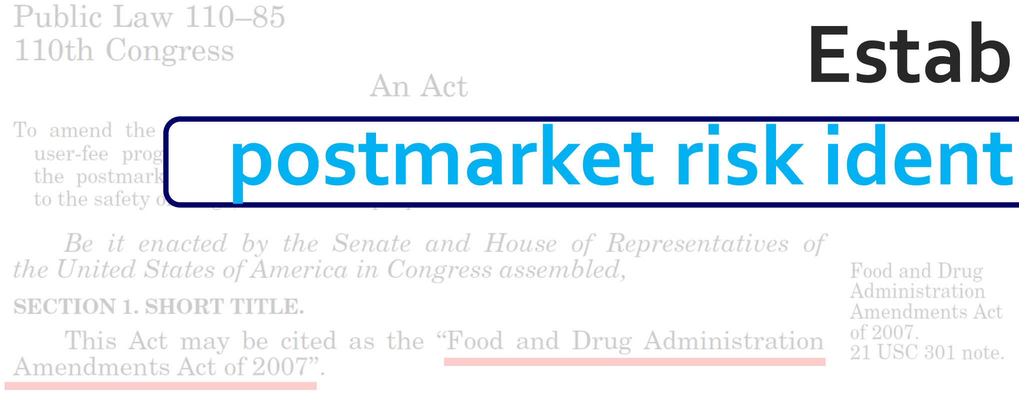
"(iii) convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data specified 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

Establishment of a postmarket risk identification and analysis system



Public Law 110–85
110th Congress

An Act

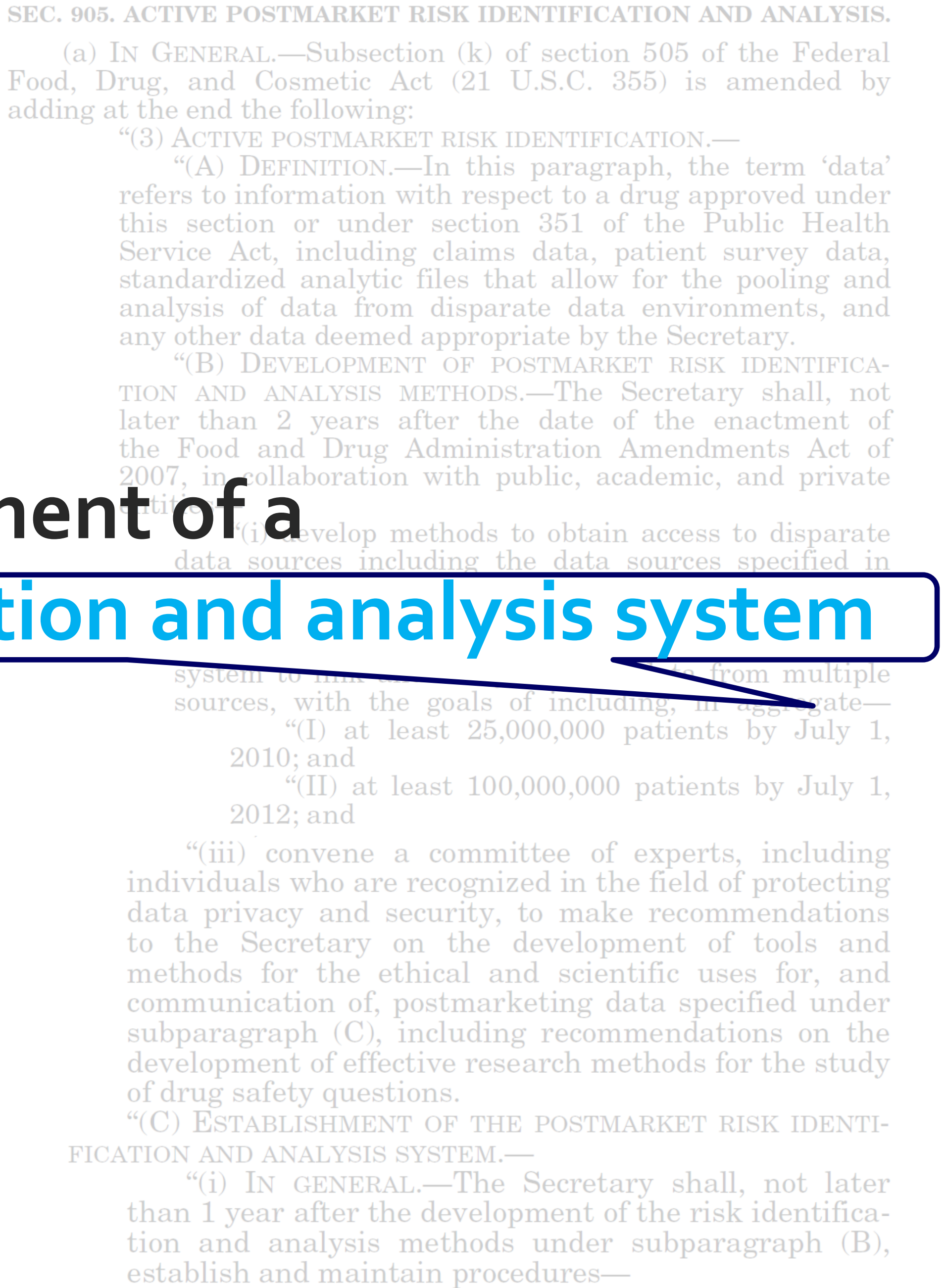
To amend the
user-fee prog
the postmark
to the safety o

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

system to link 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.—

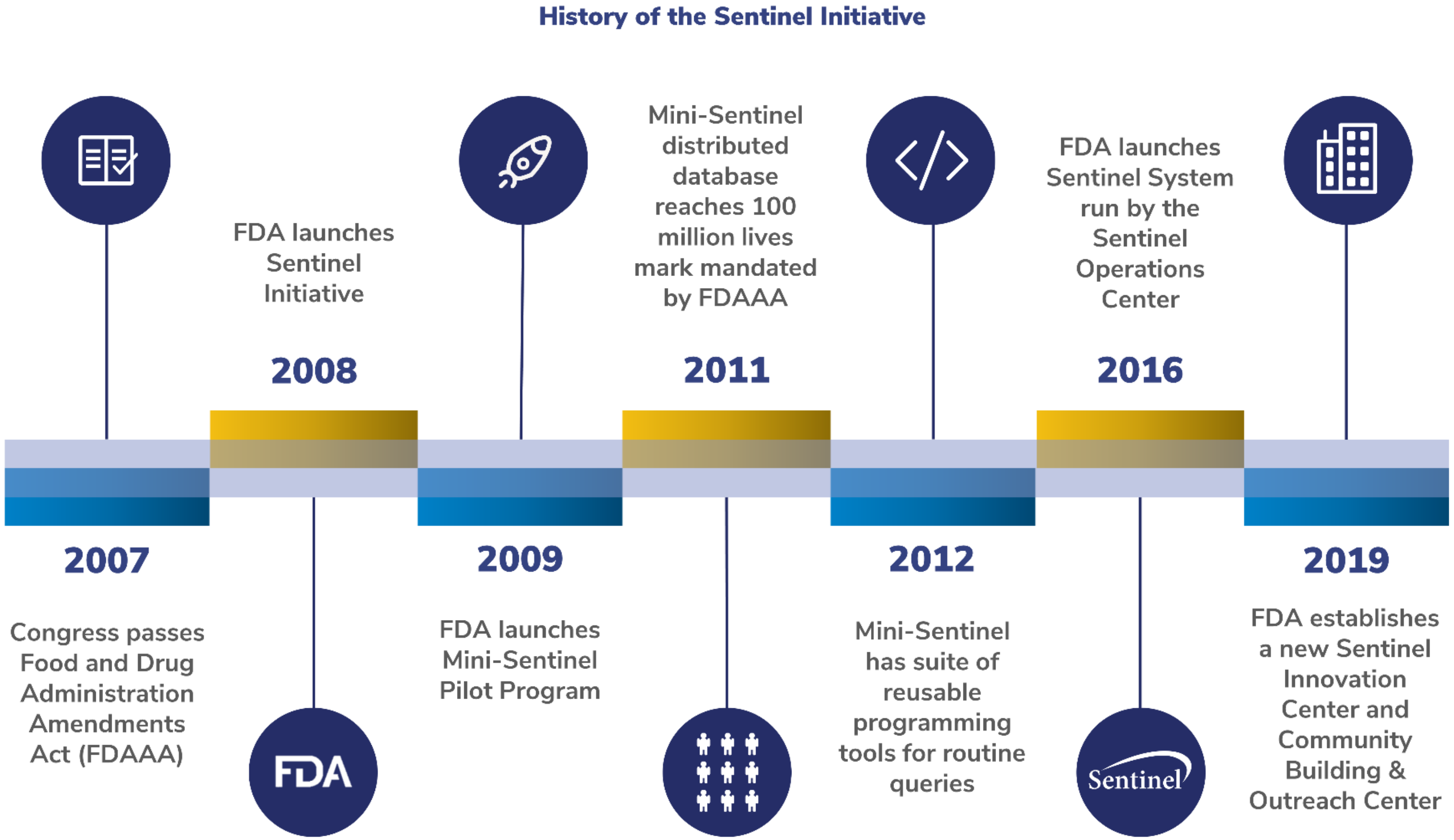
"(i) IN GENERAL.—The Secretary shall, not later than 1 year after the development of the risk identification and analysis methods under subparagraph (B), establish and maintain procedures—

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

FDA’s Sentinel System

2007 FDA Amendments Act mandates FDA to establish ***active surveillance system*** for monitoring safety of drugs using electronic healthcare data

Through the Sentinel Initiative, FDA aims to assess the post-marketing safety of approved medical products



History of the Sentinel Initiative

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

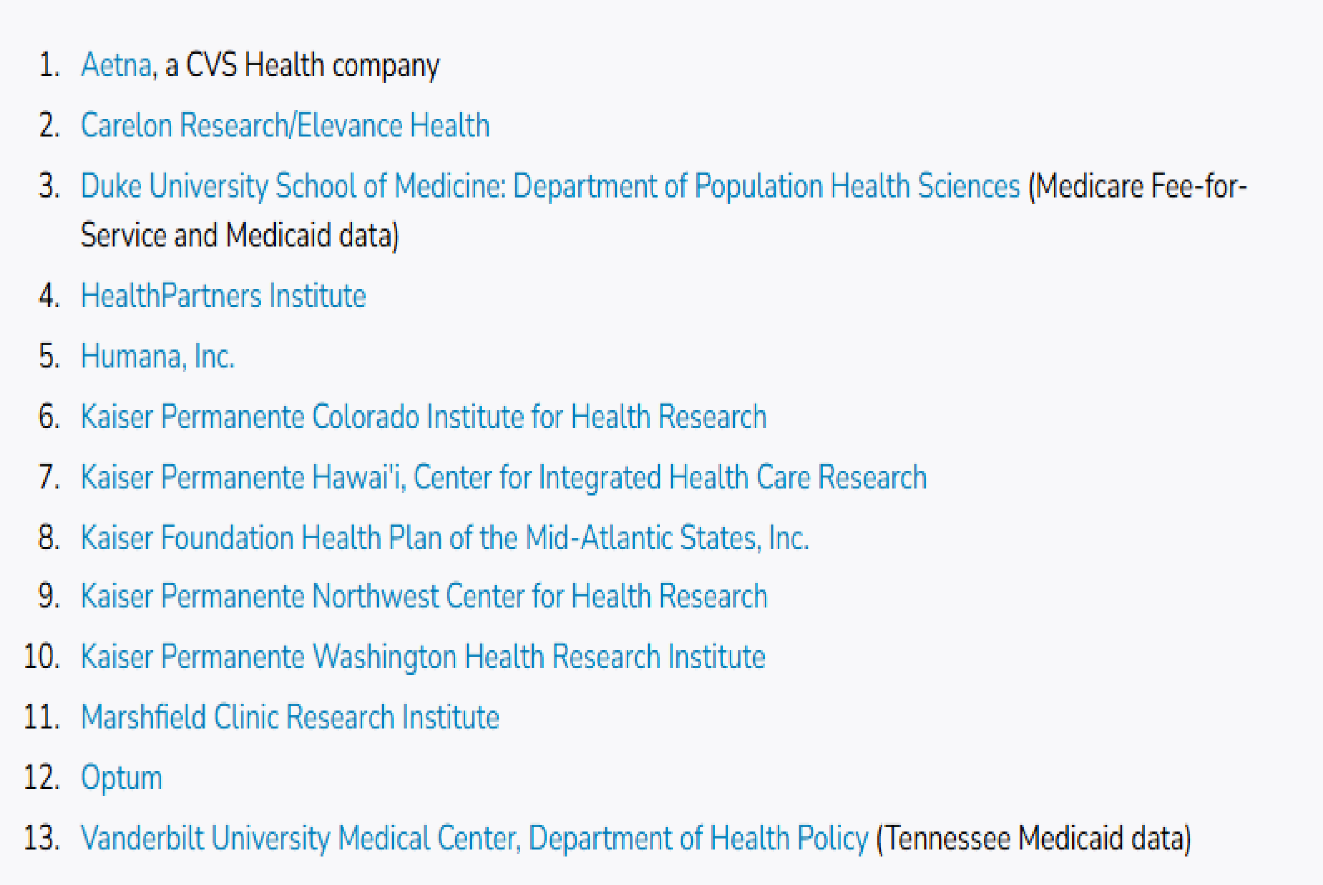
2016

FDA launches Sentinel System run by the Sentinel Operations Center

2019

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

Sentinel Distributed Database (SDD)



1. Aetna, a CVS Health company
2. Carelon Research/Elevance Health
3. Duke University School of Medicine: Department of Population Health Sciences (Medicare Fee-for-Service and Medicaid data)
4. HealthPartners Institute
5. Humana, Inc.
6. Kaiser Permanente Colorado Institute for Health Research
7. Kaiser Permanente Hawai'i, Center for Integrated Health Care Research
8. Kaiser Foundation Health Plan of the Mid-Atlantic States, Inc.
9. Kaiser Permanente Northwest Center for Health Research
10. Kaiser Permanente Washington Health Research Institute
11. Marshfield Clinic Research Institute
12. Optum
13. Vanderbilt University Medical Center, Department of Health Policy (Tennessee Medicaid data)

500.1 million unique patient identifiers (2000-2024)*

128.7 million members currently accruing new data

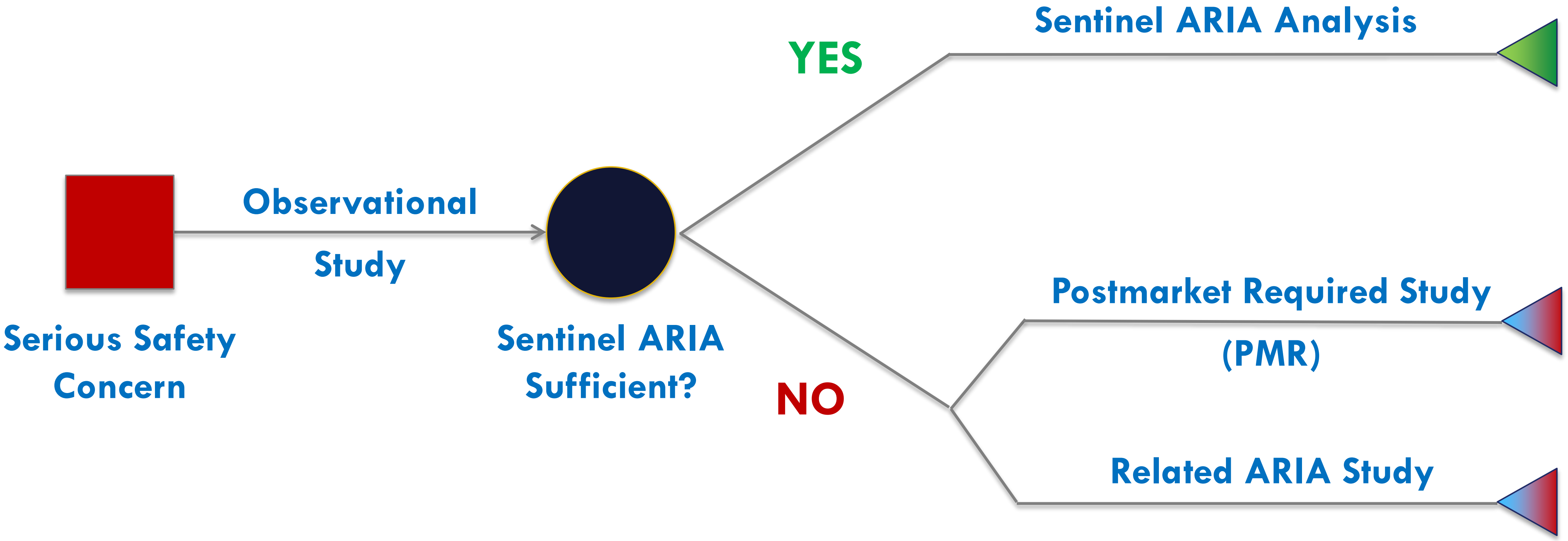
22.3 billion pharmacy dispensings

24 billion unique medical encounters

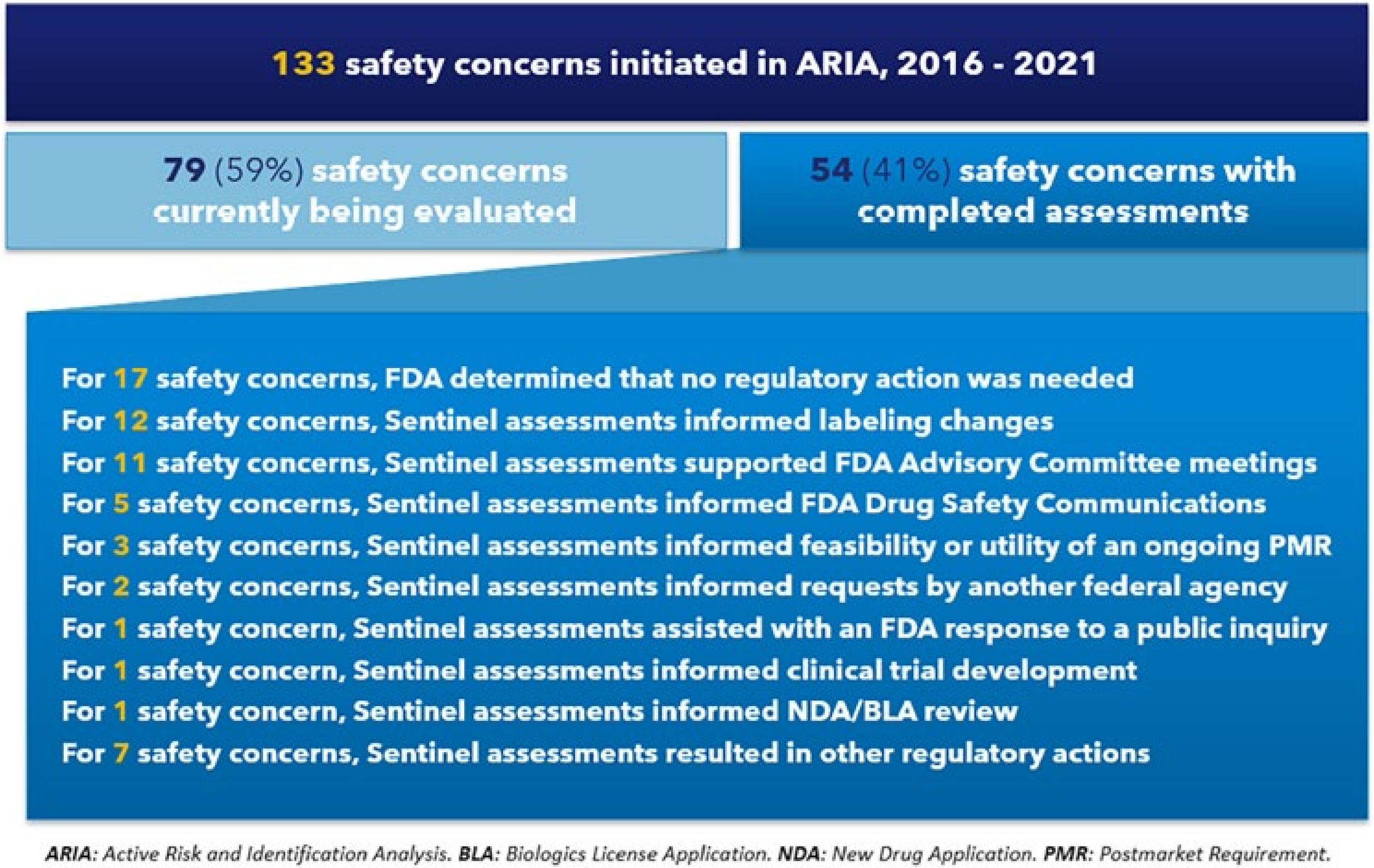
*Potential for double-counting if individuals moved between Data Partner health plans.

https://www.sentinelinitiative.org/

ARIA (Active Risk Identification and Analysis)



Impact of ARIA



133 safety concerns initiated in ARIA, 2016 - 2021

79 (59%) **safety concerns currently being evaluated**

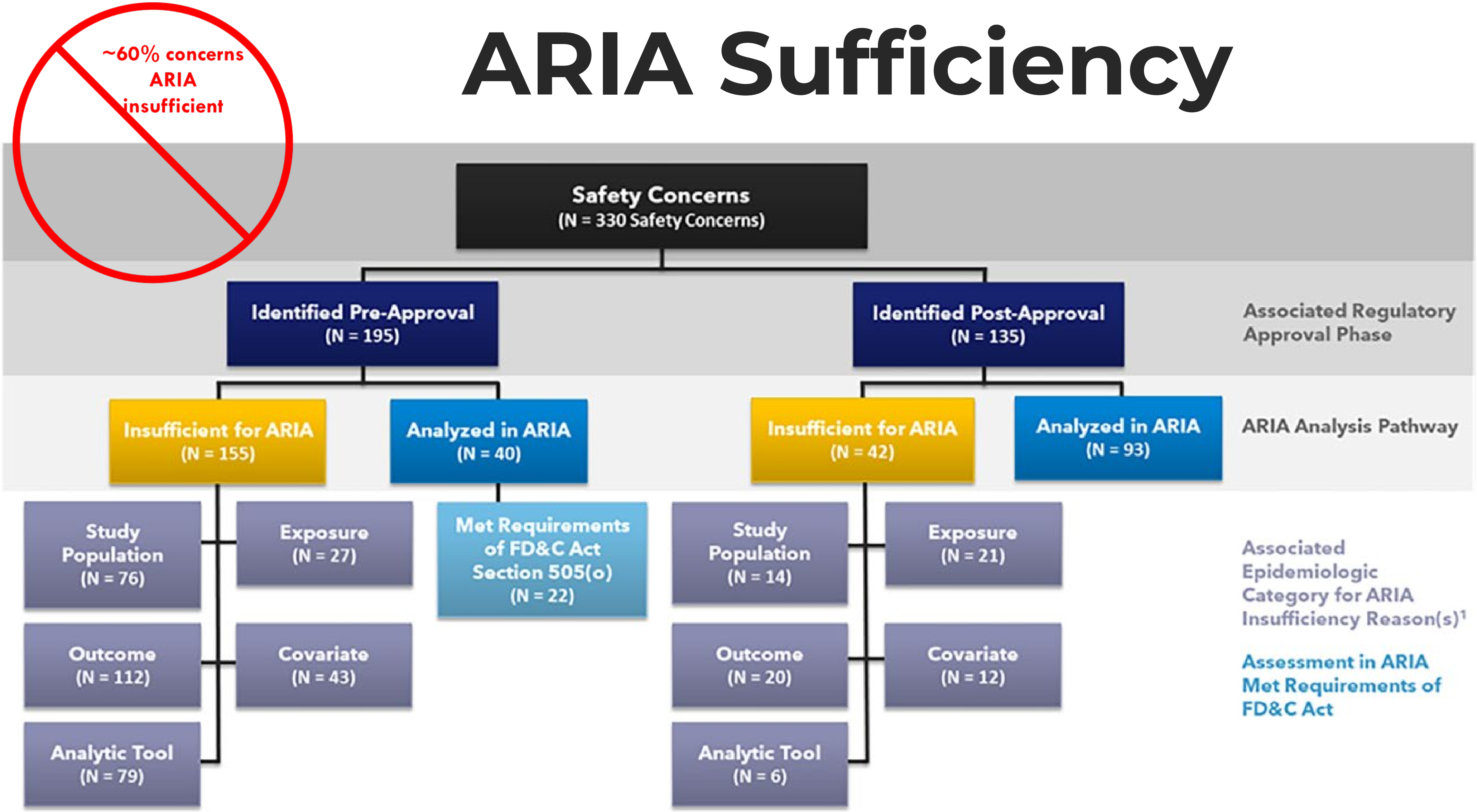
54 (41%) **safety concerns with completed assessments**

- For 17 safety concerns, FDA determined that no regulatory action was needed
- For 12 safety concerns, Sentinel assessments informed labeling changes
- For 11 safety concerns, Sentinel assessments supported FDA Advisory Committee meetings
- For 5 safety concerns, Sentinel assessments informed FDA Drug Safety Communications
- For 3 safety concerns, Sentinel assessments informed feasibility or utility of an ongoing PMR
- For 2 safety concerns, Sentinel assessments informed requests by another federal agency
- For 1 safety concern, Sentinel assessments assisted with an FDA response to a public inquiry
- For 1 safety concern, Sentinel assessments informed clinical trial development
- For 1 safety concern, Sentinel assessments informed NDA/BLA review
- For 7 safety concerns, Sentinel assessments resulted in other regulatory actions

***ARIA**: Active Risk and Identification Analysis. **BLA**: Biologics License Application. **NDA**: New Drug Application. **PMR**: Postmarket Requirement.*

Maro et al. *CPT. 2023*

ARIA Sufficiency



~60% concerns ARIA insufficient

Safety Concerns
(N = 330 Safety Concerns)

Identified Pre-Approval
(N = 195)

Identified Post-Approval
(N = 135)

Associated Regulatory Approval Phase

Insufficient for ARIA
(N = 155)

Analyzed in ARIA
(N = 40)

Insufficient for ARIA
(N = 42)

Analyzed in ARIA
(N = 93)

ARIA Analysis Pathway

Study Population
(N = 76)

Exposure
(N = 27)

Outcome
(N = 112)

Covariate
(N = 43)

Analytic Tool
(N = 79)

Met Requirements of FD&C Act Section 505(o)
(N = 22)

Study Population
(N = 14)

Exposure
(N = 21)

Outcome
(N = 20)

Covariate
(N = 12)

Analytic Tool
(N = 6)

Associated Epidemiologic Category for ARIA Insufficiency Reason(s)[1]

Assessment in ARIA Met Requirements of FD&C Act

[1]A single safety concern may be insufficient for analysis in ARIA for several reasons; thus, a single safety concern may be counted in multiple epidemiologic categories.

***ARIA:** Active Risk Identification & Analysis. **FD&C Act:** Federal Food, Drug, and Cosmetic Act.*

Maro et al. *CPT. 2023*

ARIA Insufficiency Reasons

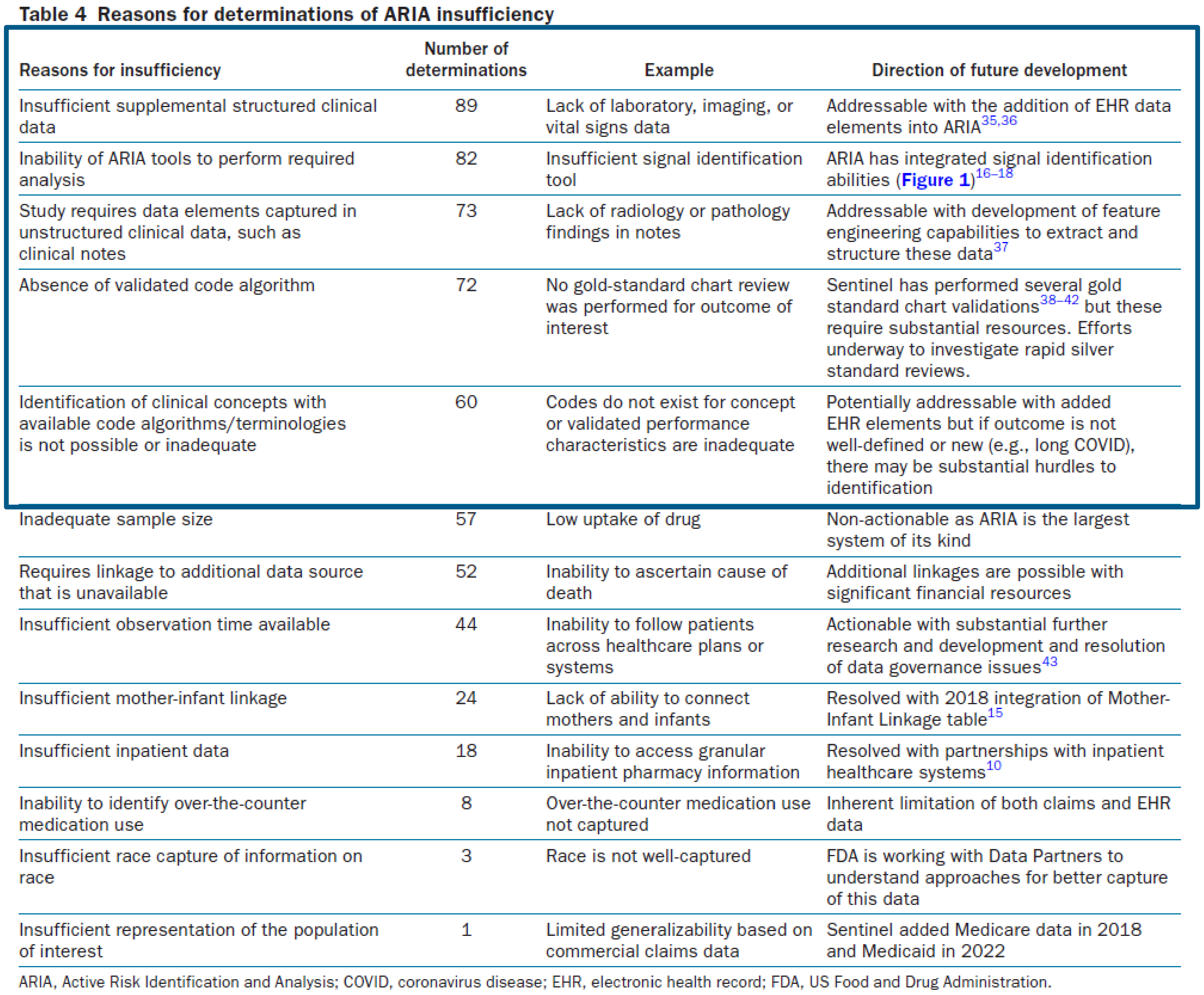


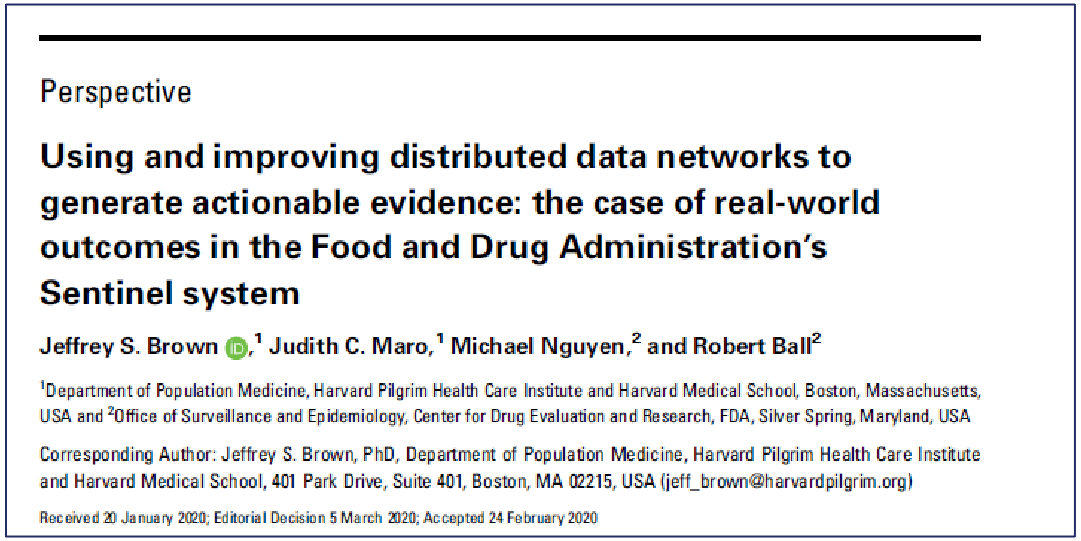
Table 4 Reasons for determinations of ARIA insufficiency

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.

Maro et al. *CPT. 2023*

Recognizing the need to harness alternative data sources and methods



Perspective

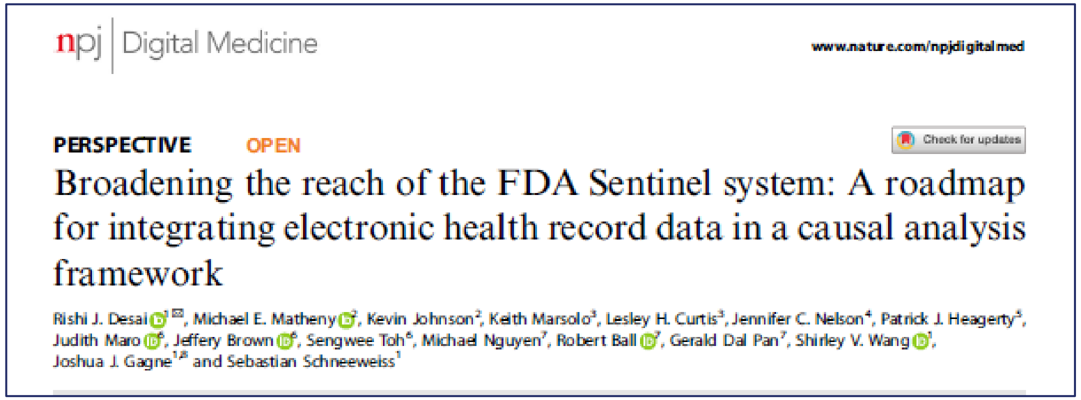
Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system

Jeffrey S. Brown,[1] Judith C. Maro,[1] Michael Nguyen,[2] and Robert Ball[2]

[1]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, USA and [2]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, Silver Spring, Maryland, USA

Corresponding Author: Jeffrey S. Brown, PhD, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, 401 Park Drive, Suite 401, Boston, MA 02215, USA (jeff_brown@harvardpilgrim.org)

Received 20 January 2020; Editorial Decision 5 March 2020; Accepted 24 February 2020



npj | Digital Medicine

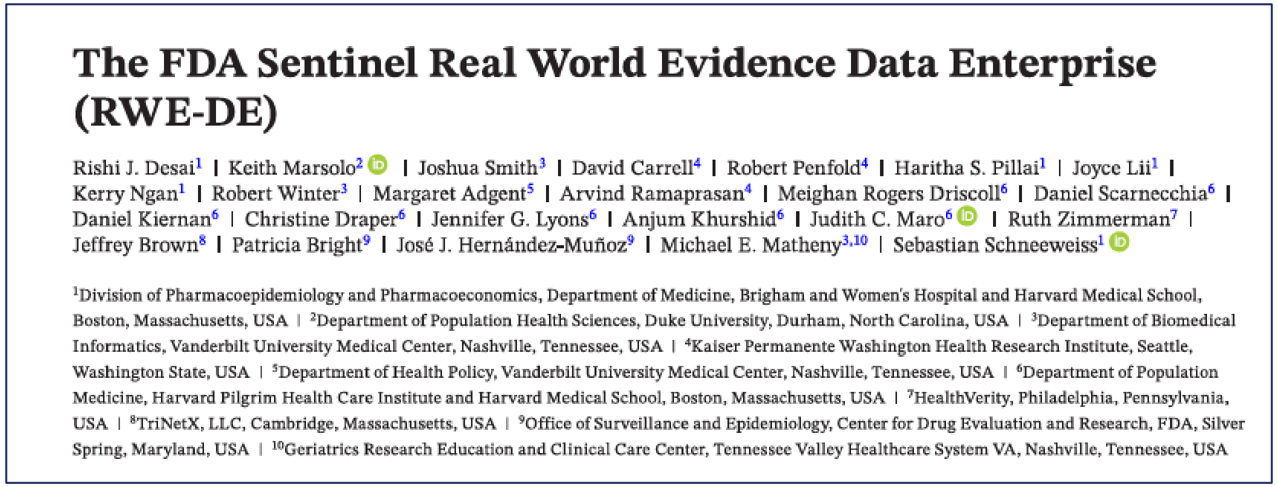
www.nature.com/npjdigitalmed

PERSPECTIVE OPEN

Check for updates

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

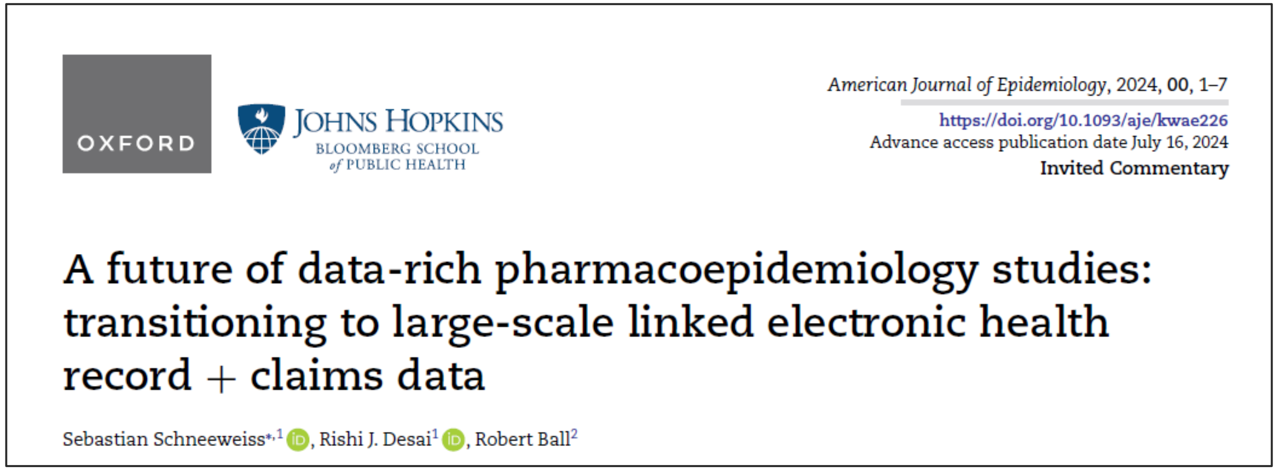
Rishi J. Desai[1] ✉, Michael E. Matheny[2], Kevin Johnson[2], Keith Marsolo[3], Lesley H. Curtis[3], Jennifer C. Nelson[4], Patrick J. Heagerty[5], Judith Maro[6], Jeffery Brown[6], Sengwee Toh[6], Michael Nguyen[7], Robert Ball[7], Gerald Dal Pan[7], Shirley V. Wang[1], Joshua J. Gagne[1,8] and Sebastian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



The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE)

Rishi J. Desai[1] | Keith Marsolo[2] | Joshua Smith[3] | David Carrell[4] | Robert Penfold[4] | Haritha S. Pillai[1] | Joyce Lii[1] | Kerry Ngan[1] | Robert Winter[3] | Margaret Adgent[5] | Arvind Ramaprasan[4] | Meighan Rogers Driscoll[6] | Daniel Scarnecchia[6] | Daniel Kiernan[6] | Christine Draper[6] | Jennifer G. Lyons[6] | Anjum Khurshid[6] | Judith C. Maro[6] | Ruth Zimmerman[7] | Jeffrey Brown[8] | Patricia Bright[9] | José J. Hernández-Muñoz[9] | Michael E. Matheny[3,10] | Sebastian Schneeweiss[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA | [2]Department of Population Health Sciences, Duke University, Durham, North Carolina, USA | [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA | [4]Kaiser Permanente Washington Health Research Institute, Seattle, Washington State, USA | [5]Department of Health Policy, Vanderbilt University Medical Center, Nashville, Tennessee, USA | [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, USA | [7]HealthVerity, Philadelphia, Pennsylvania, USA | [8]TriNetX, LLC, Cambridge, Massachusetts, USA | [9]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, Silver Spring, Maryland, USA | [10]Geriatrics Research Education and Clinical Care Center, Tennessee Valley Healthcare System VA, Nashville, Tennessee, USA



OXFORD

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

American Journal of Epidemiology, 2024, **00**, 1–7
https://doi.org/10.1093/aje/kwae226
Advance access publication date July 16, 2024

Invited Commentary

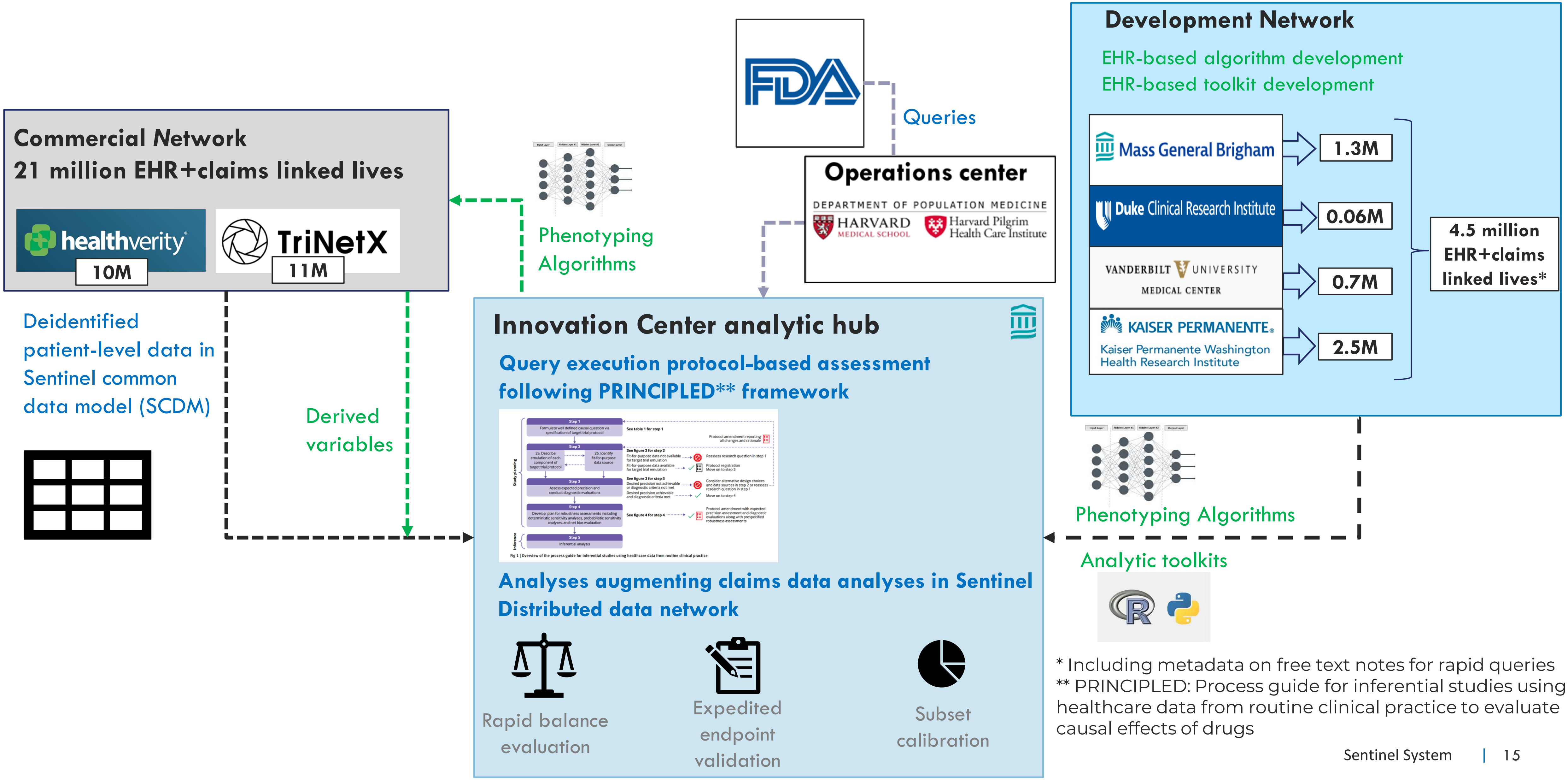
A future of data-rich pharmacoepidemiology studies: transitioning to large-scale linked electronic health record + claims data

Sebastian Schneeweiss*,[1], Rishi J. Desai[1], Robert Ball[2]

Brown et al. JAMIA 2020
Desai et al. npj Digital Medicine 2021
Schneeweiss et al. AJE 2024
Desai et al. PDS 2024

Real World Evidence Data Enterprise (RWE-DE)

The Sentinel RWE-DE based on EHR+claims data today



Commercial Network
21 million EHR+claims linked lives

healthverity 10M

TriNetX 11M

Deidentified patient-level data in Sentinel common data model (SCDM)

Derived variables

Phenotyping Algorithms

FDA

Queries

Operations center

DEPARTMENT OF POPULATION MEDICINE
HARVARD MEDICAL SCHOOL
Harvard Pilgrim Health Care Institute

Development Network

EHR-based algorithm development
EHR-based toolkit development

Mass General Brigham	1.3M
Duke Clinical Research Institute	0.06M
VANDERBILT UNIVERSITY MEDICAL CENTER	0.7M
KAISER PERMANENTE. Kaiser Permanente Washington Health Research Institute	2.5M

4.5 million EHR+claims linked lives*

Innovation Center analytic hub

Query execution protocol-based assessment following PRINCIPLED** framework

Study planning

Step 1
Formulate well defined causal question via specification of target trial protocol

See table 1 for step 1

Protocol amendment reporting all changes and rationale

Step 2
2a. Describe emulation of each component of target trial protocol
2b. Identify fit-for-purpose data source

See figure 2 for step 2
Fit-for-purpose data not available for target trial emulation → Reassess research question in step 1
Fit-for-purpose data available for target trial emulation → ✓ Protocol registration Move on to step 3

Step 3
Assess expected precision and conduct diagnostic evaluations

See figure 3 for step 3
Desired precision not achievable or diagnostic criteria not met → Consider alternative design choices and data sources in step 2 or reassess research question in step 1
Desired precision achievable and diagnostic criteria met → ✓ Move on to step 4

Step 4
Develop plan for robustness assessments including deterministic sensitivity analyses, probabilistic sensitivity analyses, and net bias evaluation

See figure 4 for step 4 → ✓ Protocol amendment with expected precision assessment and diagnostic evaluations along with prespecified robustness assessments

Inference

Step 5
Inferential analysis

Fig 1 | Overview of the process guide for inferential studies using healthcare data from routine clinical practice

Analyses augmenting claims data analyses in Sentinel Distributed data network

Rapid balance evaluation

Expedited endpoint validation

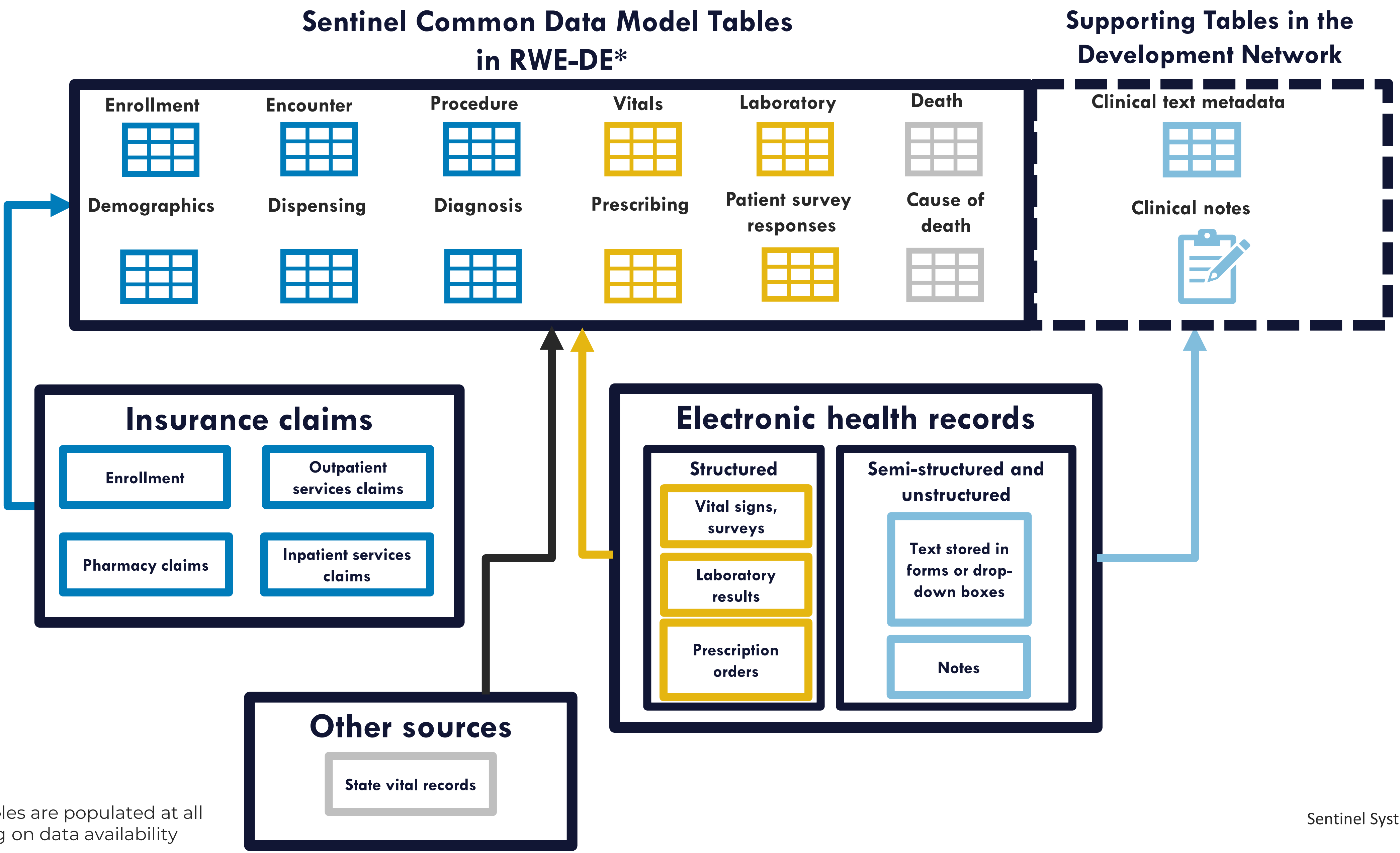
Subset calibration

Phenotyping Algorithms

Analytic toolkits

* Including metadata on free text notes for rapid queries
** PRINCIPLED: Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs

Data Sources and Availability in the RWE-DE



Sentinel Common Data Model Tables in RWE-DE*

Enrollment

Encounter

Procedure

Vitals

Laboratory

Death

Demographics

Dispensing

Diagnosis

Prescribing

Patient survey responses

Cause of death

Supporting Tables in the Development Network

Clinical text metadata

Clinical notes

Insurance claims

Enrollment

Outpatient services claims

Pharmacy claims

Inpatient services claims

Electronic health records

Structured

Vital signs, surveys

Laboratory results

Prescription orders

Semi-structured and unstructured

Text stored in forms or drop-down boxes

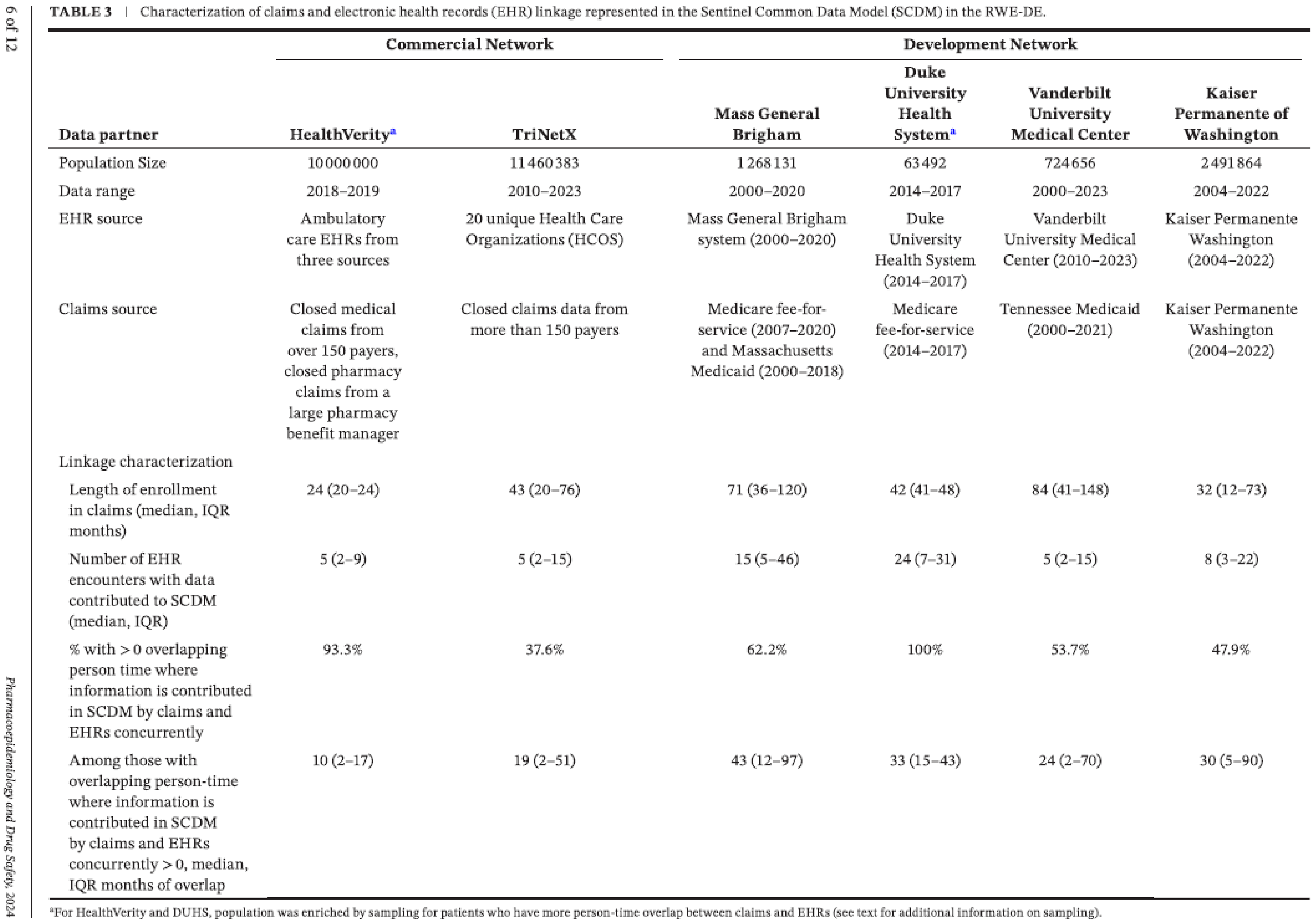
Notes

Other sources

State vital records

* Not all the tables are populated at all sites depending on data availability

Overview of the Data Sources at RWE-DE Sites



6 of 12

TABLE 3 | Characterization of claims and electronic health records (EHR) linkage represented in the Sentinel Common Data Model (SCDM) in the RWE-DE.

	Commercial Network		Development Network			
Data partner	**HealthVerity[a]**	**TriNetX**	**Mass General Brigham**	**Duke University Health System[a]**	**Vanderbilt University Medical Center**	**Kaiser Permanente of Washington**
Population Size	10000000	11460383	1268131	63492	724656	2491864
Data range	2018–2019	2010–2023	2000–2020	2014–2017	2000–2023	2004–2022
EHR source	Ambulatory care EHRs from three sources	20 unique Health Care Organizations (HCOS)	Mass General Brigham system (2000–2020)	Duke University Health System (2014–2017)	Vanderbilt University Medical Center (2010–2023)	Kaiser Permanente Washington (2004–2022)
Claims source	Closed medical claims from over 150 payers, closed pharmacy claims from a large pharmacy benefit manager	Closed claims data from more than 150 payers	Medicare fee-for-service (2007–2020) and Massachusetts Medicaid (2000–2018)	Medicare fee-for-service (2014–2017)	Tennessee Medicaid (2000–2021)	Kaiser Permanente Washington (2004–2022)
Linkage characterization						
Length of enrollment in claims (median, IQR months)	24 (20–24)	43 (20–76)	71 (36–120)	42 (41–48)	84 (41–148)	32 (12–73)
Number of EHR encounters with data contributed to SCDM (median, IQR)	5 (2–9)	5 (2–15)	15 (5–46)	24 (7–31)	5 (2–15)	8 (3–22)
% with > 0 overlapping person time where information is contributed in SCDM by claims and EHRs concurrently	93.3%	37.6%	62.2%	100%	53.7%	47.9%
Among those with overlapping person-time where information is contributed in SCDM by claims and EHRs concurrently > 0, median, IQR months of overlap	10 (2–17)	19 (2–51)	43 (12–97)	33 (15–43)	24 (2–70)	30 (5–90)

[a]For HealthVerity and DUHS, population was enriched by sampling for patients who have more person-time overlap between claims and EHRs (see text for additional information on sampling).

Pharmacoepidemiology and Drug Safety, 2024

Desai RJ, Marsolo K, Smith J, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). Pharmacoepidemiol Drug Saf. 2024;33(10):e70028. doi:10.1002/pds.70028

Overview of the Populations Covered in RWE-DE

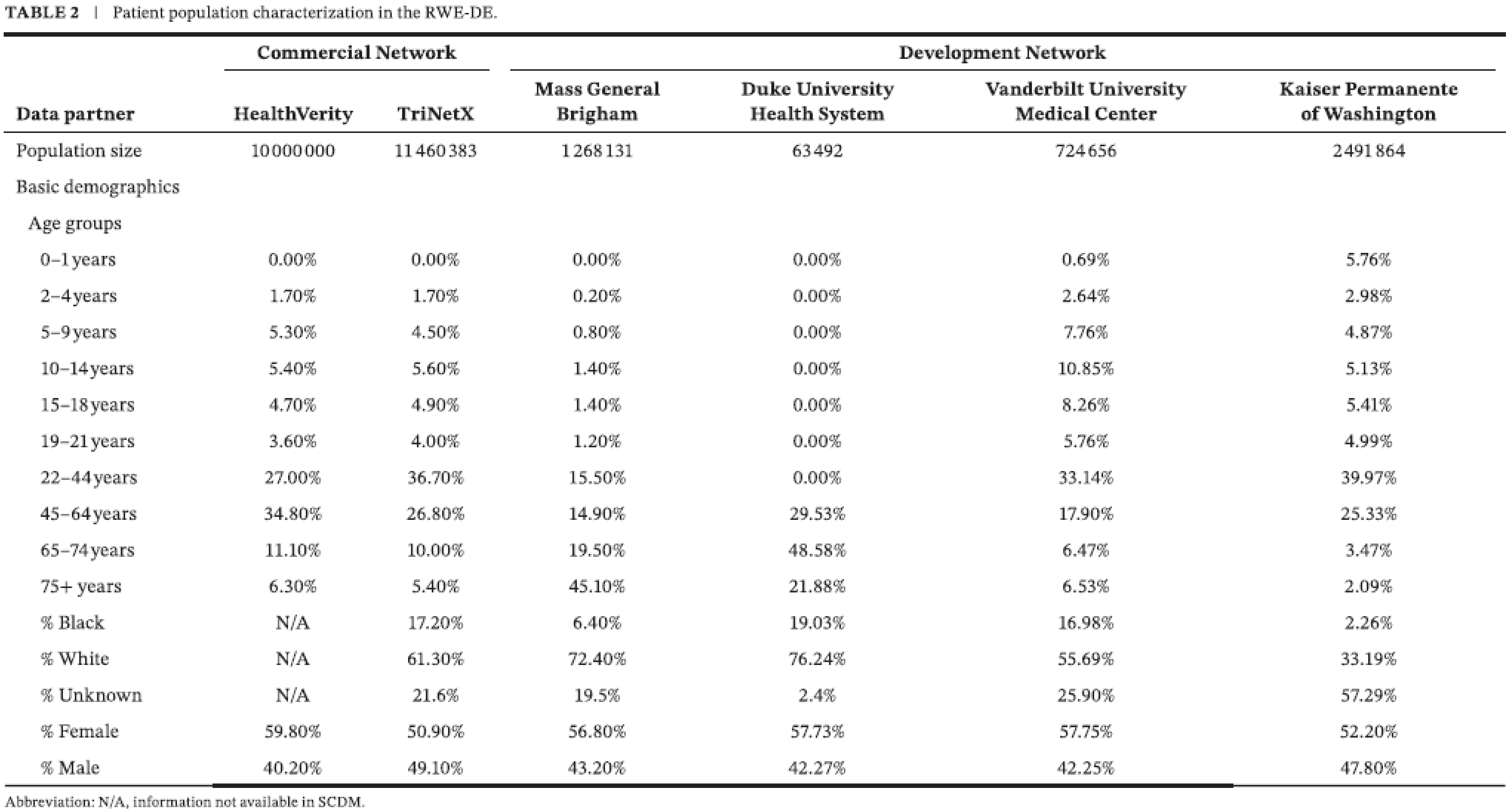


TABLE 2 | Patient population characterization in the RWE-DE.

	Commercial Network		Development Network			
Data partner	**HealthVerity**	**TriNetX**	**Mass General Brigham**	**Duke University Health System**	**Vanderbilt University Medical Center**	**Kaiser Permanente of Washington**
Population size	10 000 000	11 460 383	1 268 131	63 492	724 656	2 491 864
Basic demographics						
Age groups						
0–1 years	0.00%	0.00%	0.00%	0.00%	0.69%	5.76%
2–4 years	1.70%	1.70%	0.20%	0.00%	2.64%	2.98%
5–9 years	5.30%	4.50%	0.80%	0.00%	7.76%	4.87%
10–14 years	5.40%	5.60%	1.40%	0.00%	10.85%	5.13%
15–18 years	4.70%	4.90%	1.40%	0.00%	8.26%	5.41%
19–21 years	3.60%	4.00%	1.20%	0.00%	5.76%	4.99%
22–44 years	27.00%	36.70%	15.50%	0.00%	33.14%	39.97%
45–64 years	34.80%	26.80%	14.90%	29.53%	17.90%	25.33%
65–74 years	11.10%	10.00%	19.50%	48.58%	6.47%	3.47%
75+ years	6.30%	5.40%	45.10%	21.88%	6.53%	2.09%
% Black	N/A	17.20%	6.40%	19.03%	16.98%	2.26%
% White	N/A	61.30%	72.40%	76.24%	55.69%	33.19%
% Unknown	N/A	21.6%	19.5%	2.4%	25.90%	57.29%
% Female	59.80%	50.90%	56.80%	57.73%	57.75%	52.20%
% Male	40.20%	49.10%	43.20%	42.27%	42.25%	47.80%

Abbreviation: N/A, information not available in SCDM.

Desai RJ, Marsolo K, Smith J, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). Pharmacoepidemiol Drug Saf. 2024;33(10):e70028. doi:10.1002/pds.70028

Methodological Initiatives

Causal Inference Requirements

Design Layer

Achieve causal study design

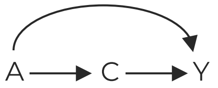
Considering:
- Study question
- Exposure variation
- Measurement quality

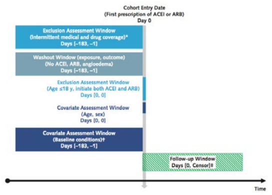
DESIGN CHOICE

1) Controlled 2) self-controlled 3) scanning
- Medically-informed target population
- Patient-informed outcomes
- Biologically-informed effect window

BIAS REDUCTION

- New users, active comparators
- Causal temporality
  Exposure before outcome
  Confounder before exposure




Cohort Entry Date
(First prescription of ACEI or ARB)
Day 0
Exclusion Assessment Window
(Intermittent medical and drug coverage)*
Days [−183, −1]
Washout Window (exposure, outcome)
(No ACEI, ARB, angioedema)
Days [−183, −1]
Exclusion Assessment Window
(Age ≤18 y, initiate both ACEI and ARB)
Days [0, 0]
Covariate Assessment Window
(Age, sex)
Days [0, 0]
Covariate Assessment Window
(Baseline conditions)†
Days [−183, −1]
Follow-up Window
Days [0, Censor]‡
Time

Measures Layer

Achieve fit-for-purpose measurement

Considering:
- sensitivity
- specificity,
- completeness
- mean sqr diff

Filling Rx
Prescribing Rx, self-report, infusers, pill caps, UDI from OR notes

EXPOSURE

Dx, Px codes
Labs, imaging, digital health dev, physician notes, patient reports

OUTCOME

Dx, Px, Rx codes
Labs, stage, imaging, BMI, genomics, physician notes, services use intensity

CONFOUNDERS

Dx, Px, Rx codes
Monitors, physician notes, biomarker, omics, behavior, socio-econ

TARGET POP^N

Analytics Layer

Achieve causal analysis

Considering:
- Confounders
- Follow-up model
- Measurement quality

BALANCE

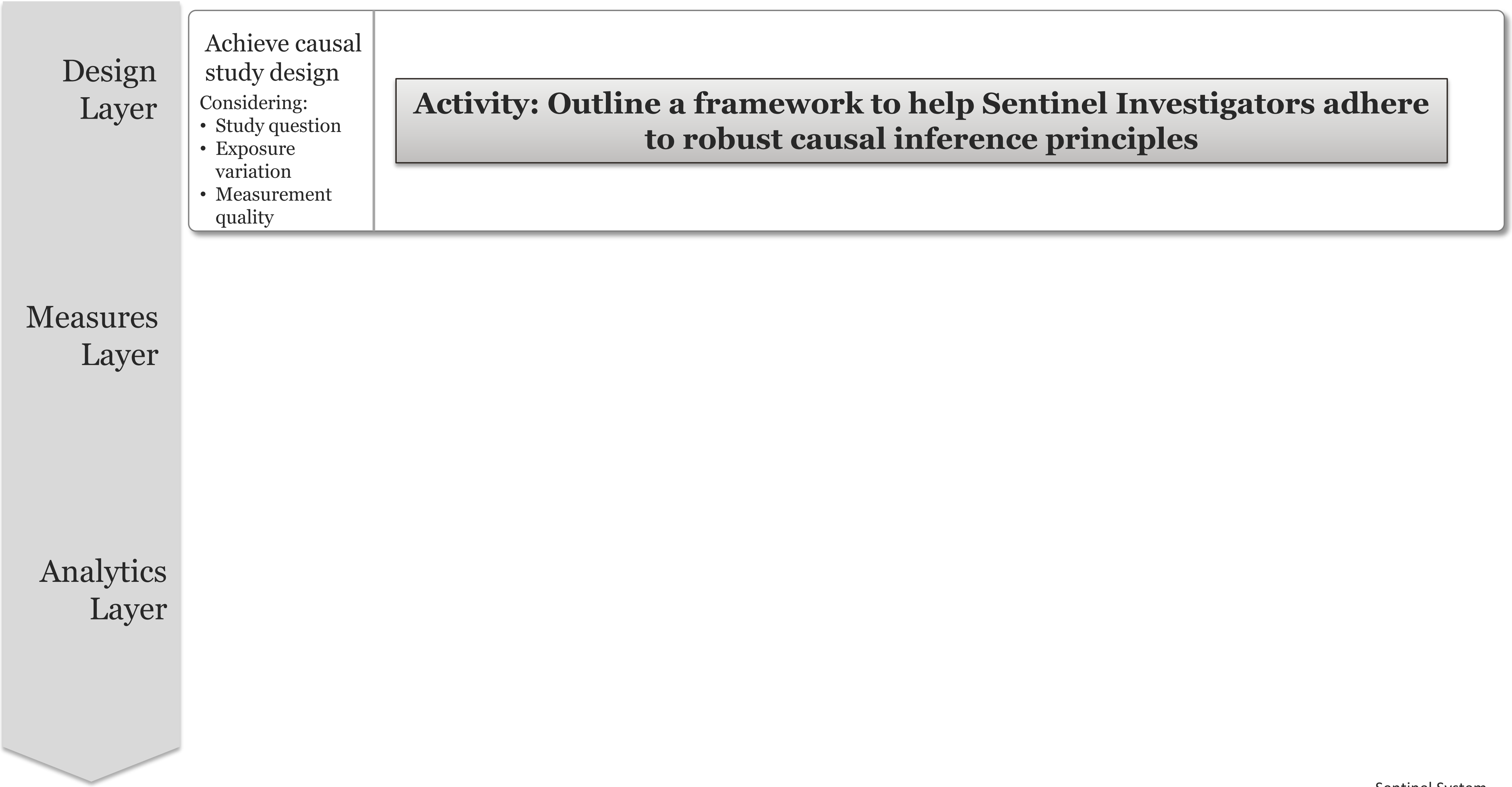
- Achieve balance:
  Regression, PS analysis
  Proxy adjustment: HDPS, CTMLE
  Time-varying exposure: MSM
- Check balance:
  SD, residuals, c-stat

ROBUSTNESS

- Sensitivity analyses of design
- Quantitative bias analysis
- Neg./pos. control endpoints
- Balance in unmeasured confounders
- Multiple comparisons

Schneeweiss & Patorno *Endocrine Reviews, 2021, Vol. 42, No. 5, 658–690*

Causal Inference Requirements



Design Layer

Measures Layer

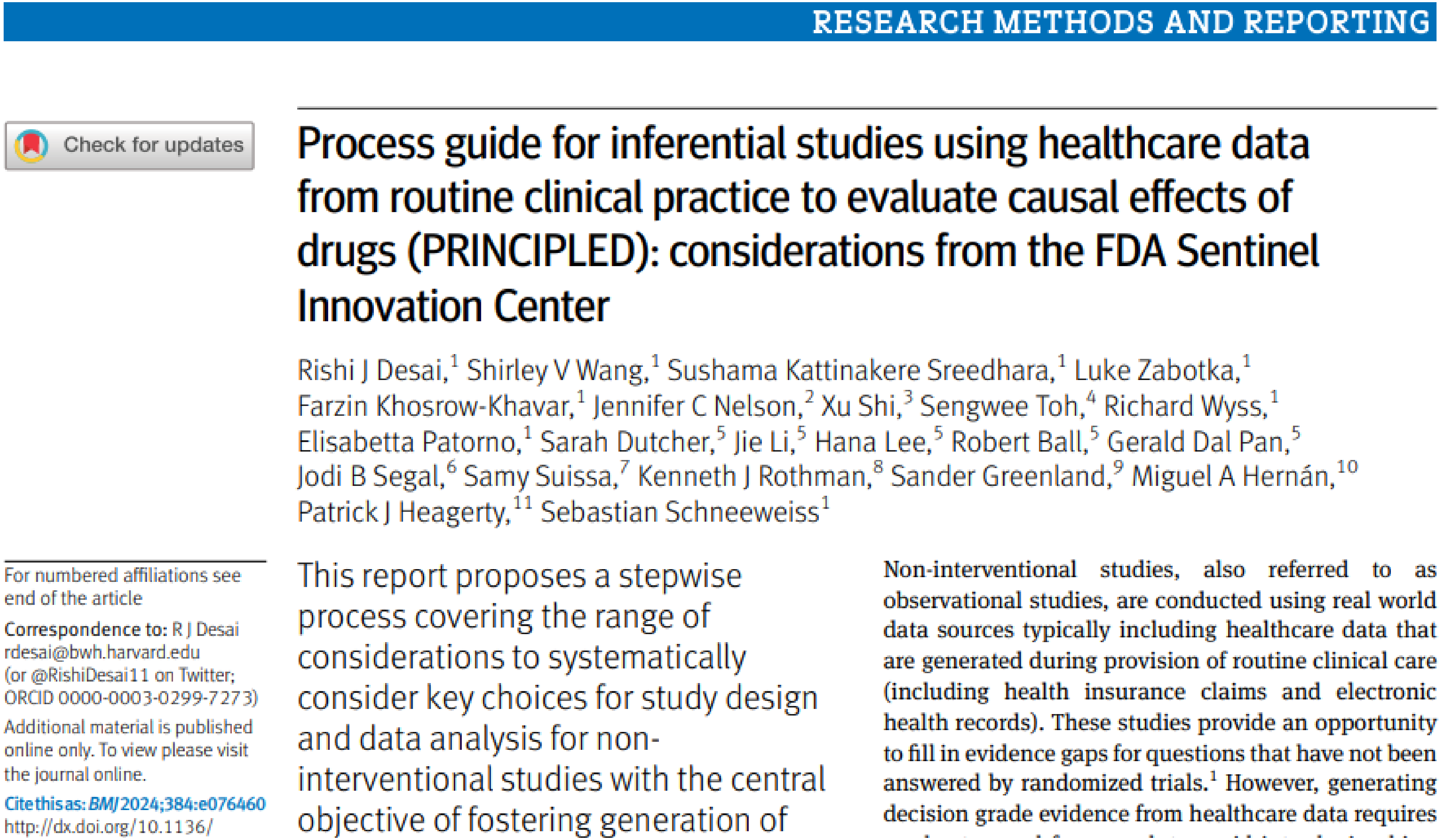
Analytics Layer

Achieve causal study design

Considering:

- Study question
- Exposure variation
- Measurement quality

Activity: Outline a framework to help Sentinel Investigators adhere to robust causal inference principles



RESEARCH METHODS AND REPORTING

Check for updates

Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center

Rishi J Desai,[1] Shirley V Wang,[1] Sushama Kattinakere Sreedhara,[1] Luke Zabotka,[1] Farzin Khosrow-Khavar,[1] Jennifer C Nelson,[2] Xu Shi,[3] Sengwee Toh,[4] Richard Wyss,[1] Elisabetta Patorno,[1] Sarah Dutcher,[5] Jie Li,[5] Hana Lee,[5] Robert Ball,[5] Gerald Dal Pan,[5] Jodi B Segal,[6] Samy Suissa,[7] Kenneth J Rothman,[8] Sander Greenland,[9] Miguel A Hernán,[10] Patrick J Heagerty,[11] Sebastian Schneeweiss[1]

For numbered affiliations see end of the article

Correspondence to: R J Desai rdesai@bwh.harvard.edu (or @RishiDesai11 on Twitter; ORCID 0000-0003-0299-7273)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;384:e076460 http://dx.doi.org/10.1136/

This report proposes a stepwise process covering the range of considerations to systematically consider key choices for study design and data analysis for non-interventional studies with the central objective of fostering generation of

Non-interventional studies, also referred to as observational studies, are conducted using real world data sources typically including healthcare data that are generated during provision of routine clinical care (including health insurance claims and electronic health records). These studies provide an opportunity to fill in evidence gaps for questions that have not been answered by randomized trials.[1] However, generating decision grade evidence from healthcare data requires

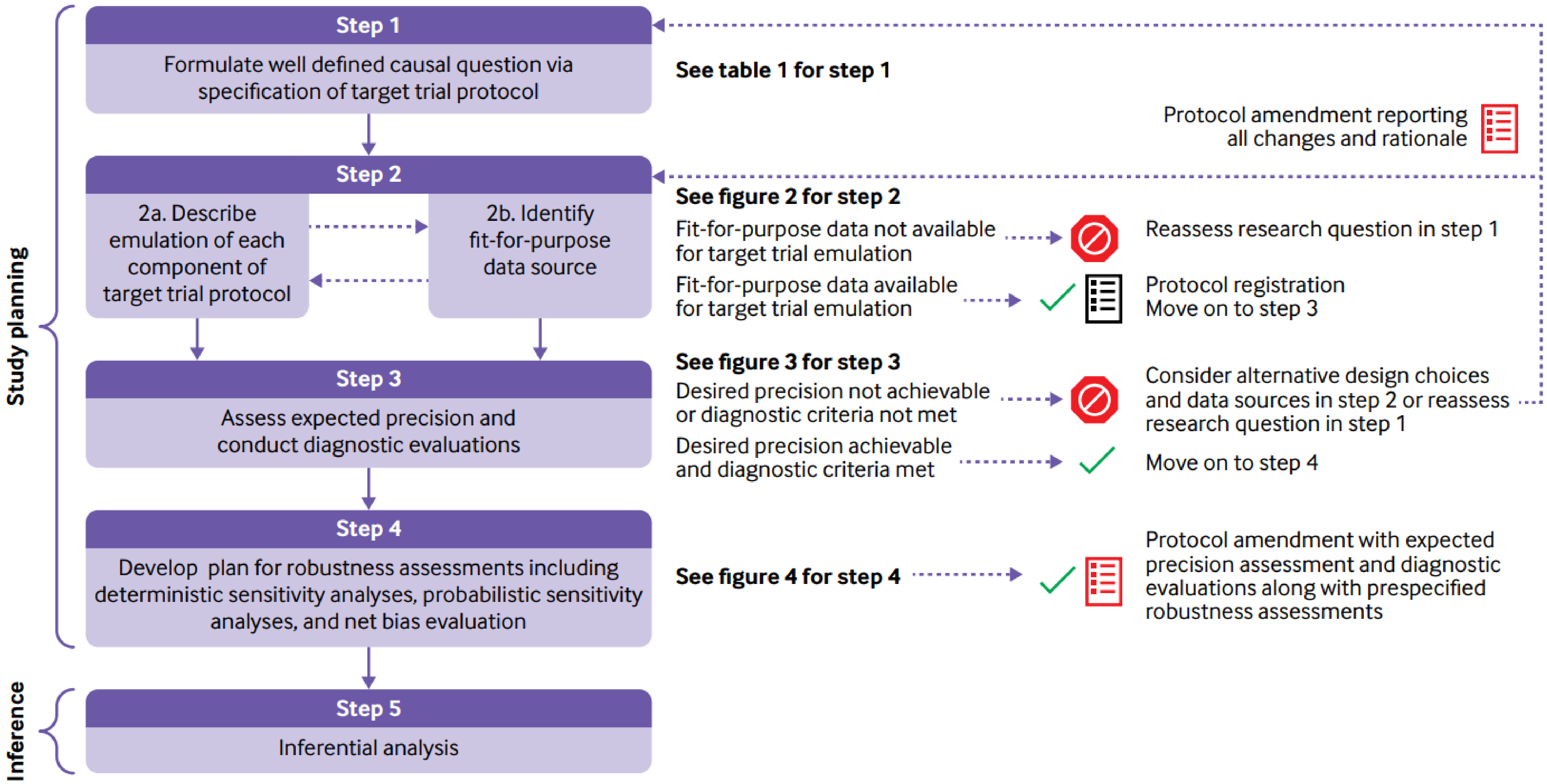
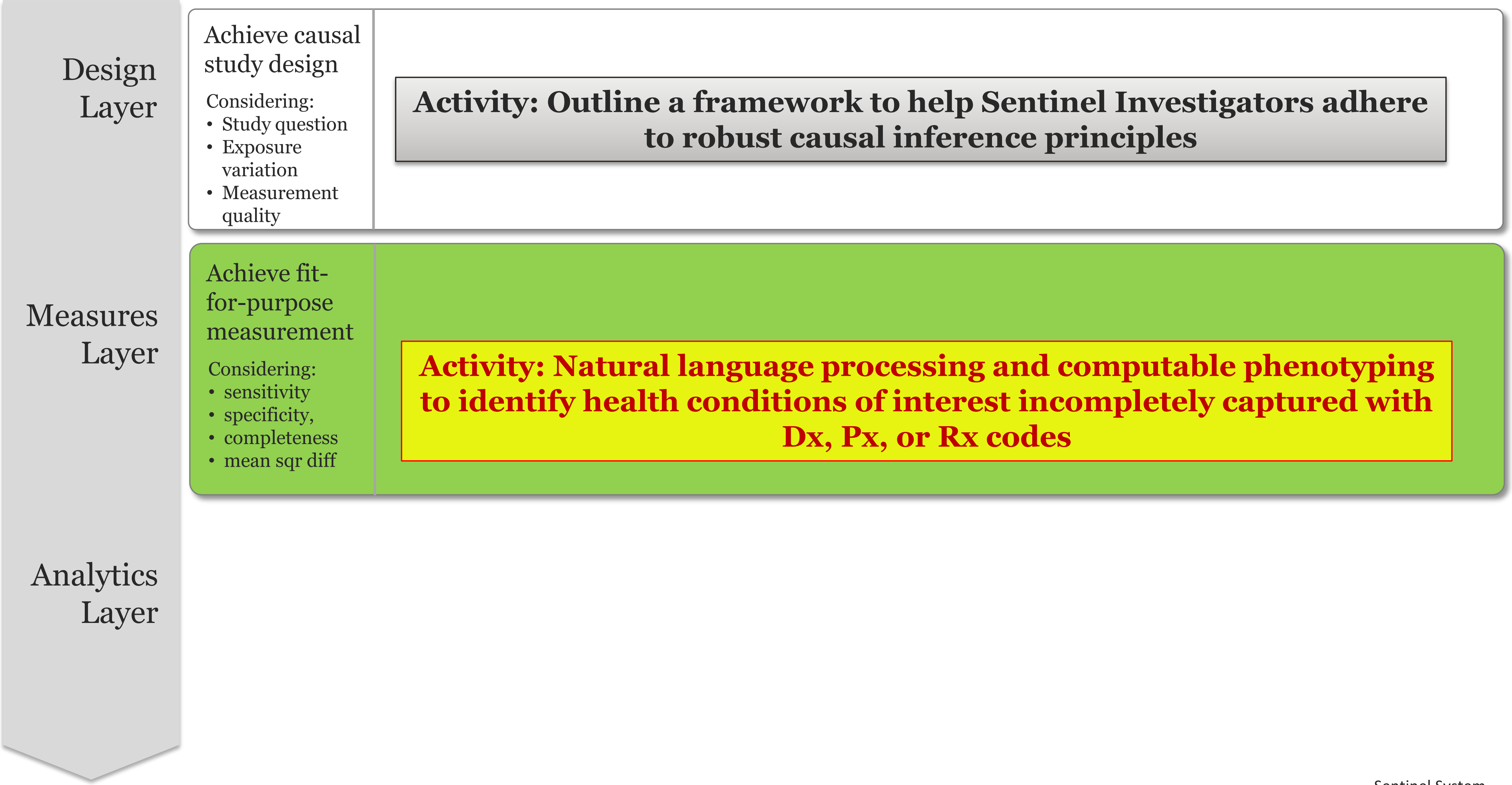

Study planning
Inference
Step 1
Formulate well defined causal question via specification of target trial protocol
Step 2
2a. Describe emulation of each component of target trial protocol
2b. Identify fit-for-purpose data source
Step 3
Assess expected precision and conduct diagnostic evaluations
Step 4
Develop plan for robustness assessments including deterministic sensitivity analyses, probabilistic sensitivity analyses, and net bias evaluation
Step 5
Inferential analysis
See table 1 for step 1
Protocol amendment reporting all changes and rationale
See figure 2 for step 2
Fit-for-purpose data not available for target trial emulation
Reassess research question in step 1
Fit-for-purpose data available for target trial emulation
Protocol registration
Move on to step 3
See figure 3 for step 3
Desired precision not achievable or diagnostic criteria not met
Consider alternative design choices and data sources in step 2 or reassess research question in step 1
Desired precision achievable and diagnostic criteria met
Move on to step 4
See figure 4 for step 4
Protocol amendment with expected precision assessment and diagnostic evaluations along with prespecified robustness assessments

Fig 1 | Overview of the process guide for inferential studies using healthcare data from routine clinical practice

Causal Inference Requirements: Role of Advanced Methods



Design Layer

Achieve causal study design

Considering:
- Study question
- Exposure variation
- Measurement quality

Activity: Outline a framework to help Sentinel Investigators adhere to robust causal inference principles

Measures Layer

Achieve fit-for-purpose measurement

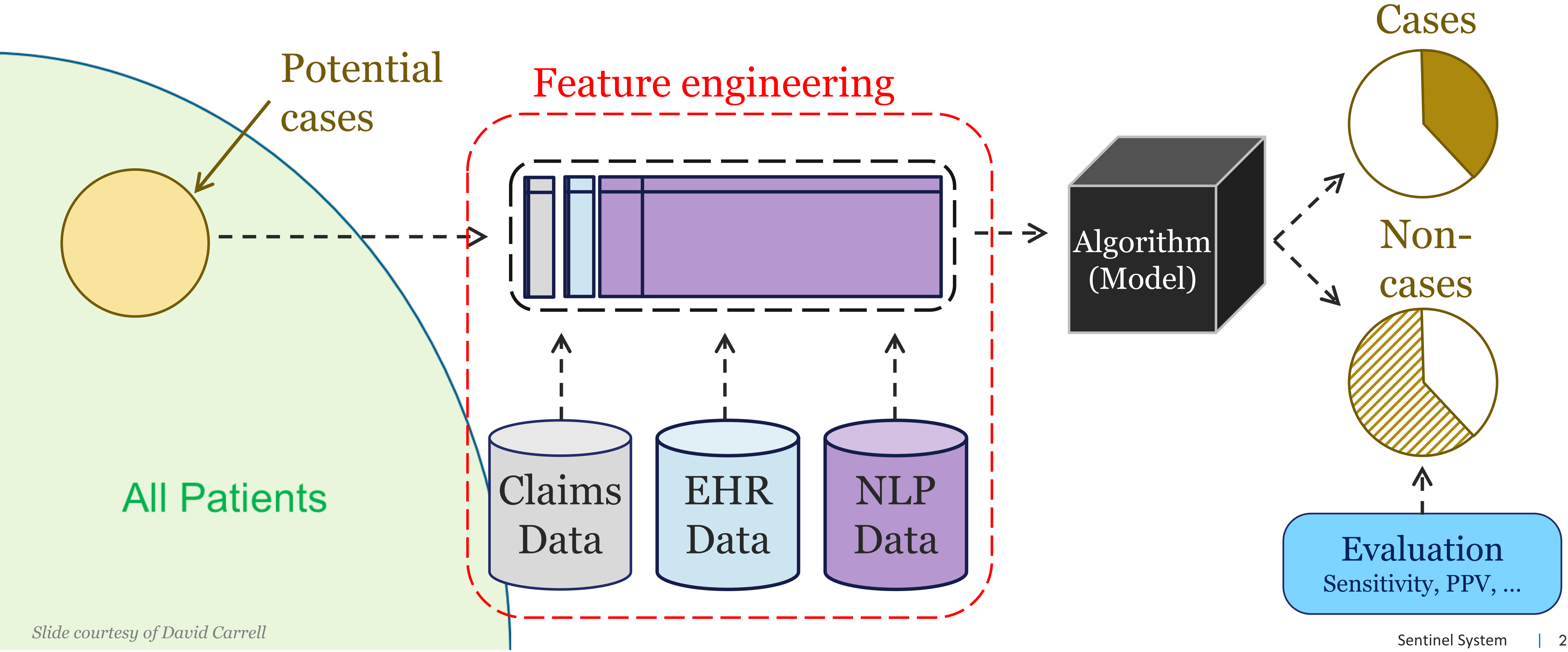
Considering:
- sensitivity
- specificity,
- completeness
- mean sqr diff

Activity: Natural language processing and computable phenotyping to identify health conditions of interest incompletely captured with Dx, Px, or Rx codes

Analytics Layer

What is computable phenotyping?

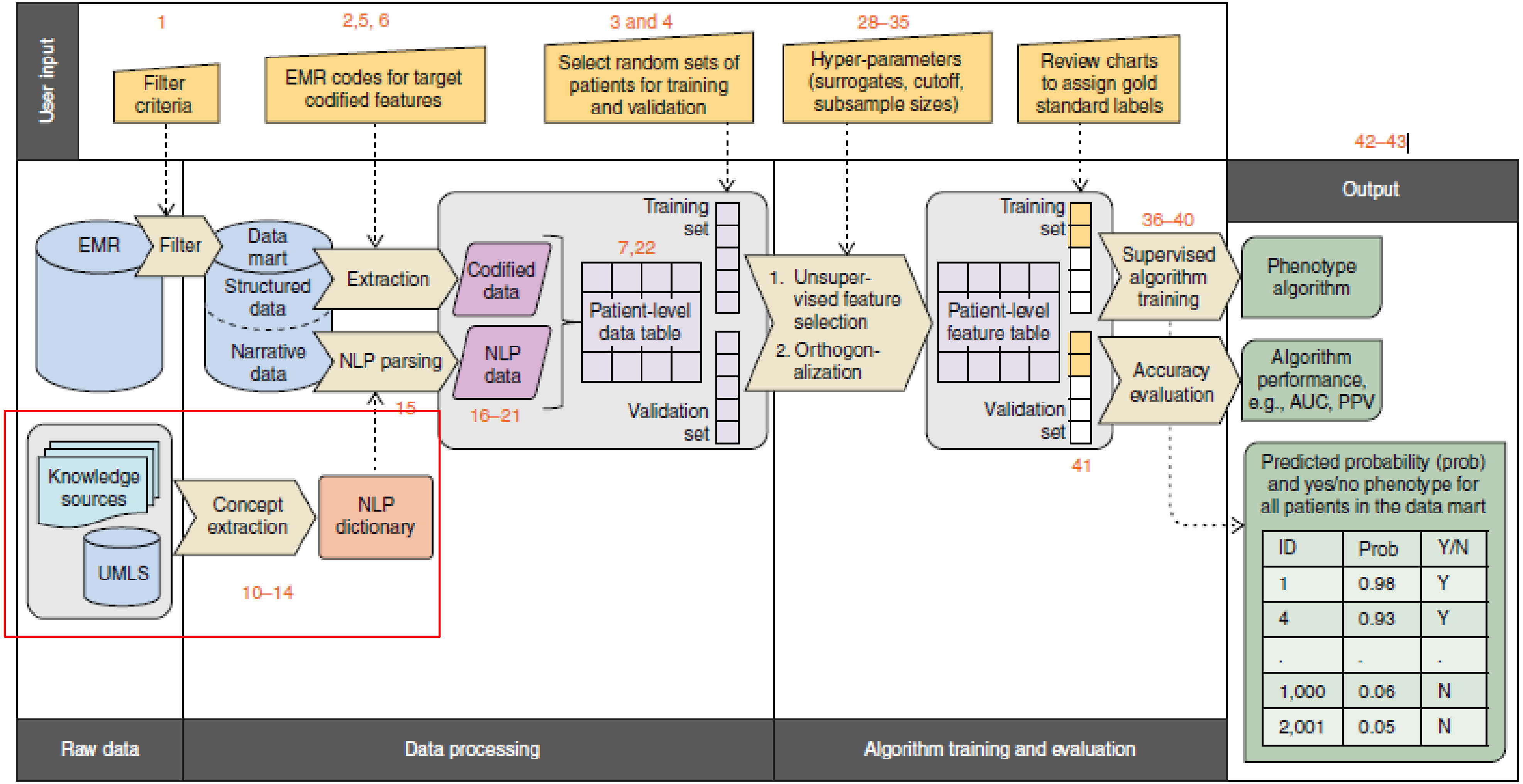
Use of algorithms (or models) to determine which patients have a particular clinical condition (AKA phenotype, health outcome of interest, "is a case")



Slide courtesy of David Carrell

High throughput phenotyping - steps

Feature engineering

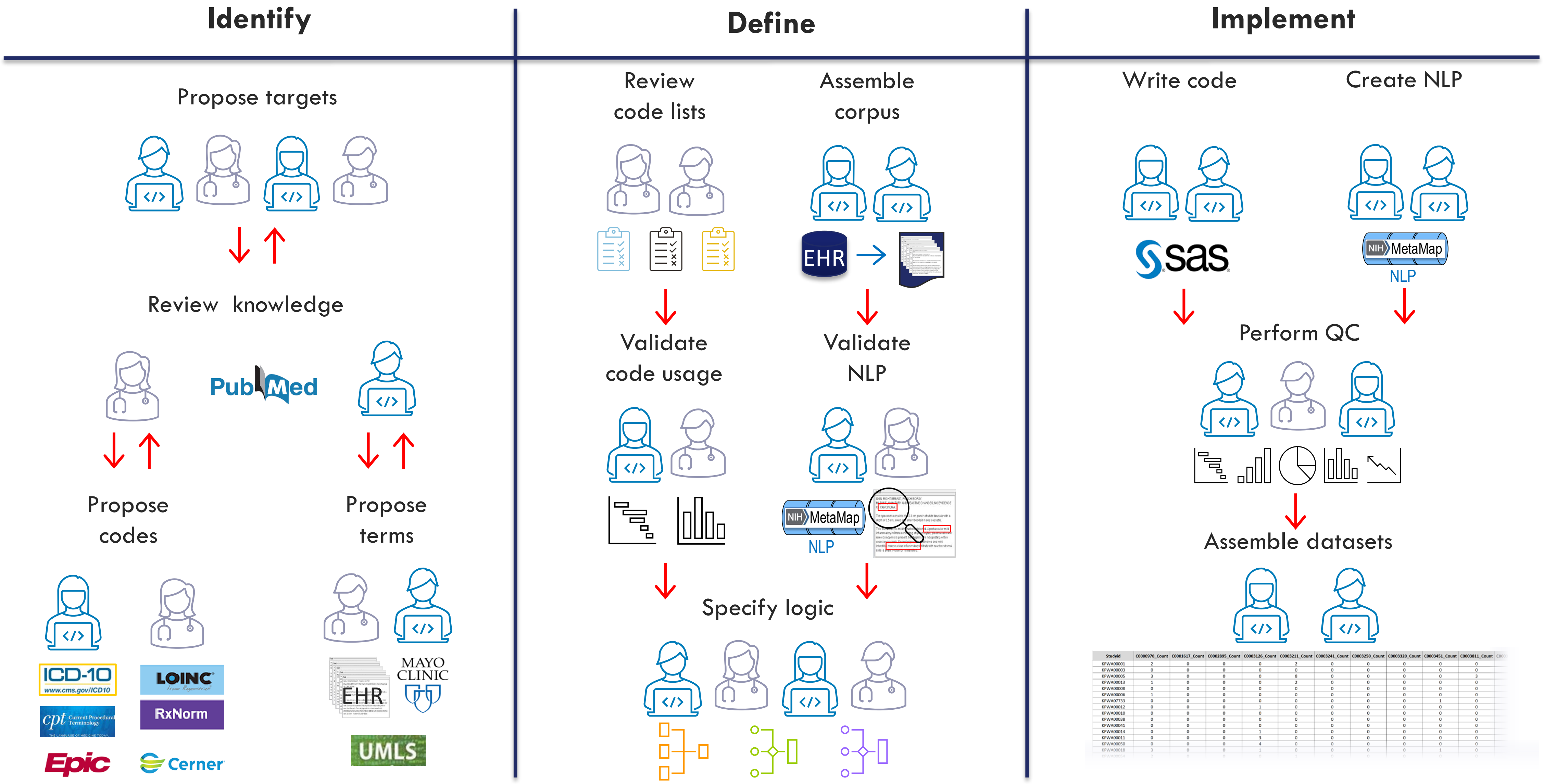

User input
1
Filter criteria
2,5, 6
EMR codes for target codified features
3 and 4
Select random sets of patients for training and validation
28–35
Hyper-parameters (surrogates, cutoff, subsample sizes)
Review charts to assign gold standard labels
42–43
Output
EMR
Filter
Data mart
Structured data
Narrative data
Extraction
NLP parsing
Codified data
NLP data
7,22
16–21
15
Training set
Patient-level data table
Validation set
1. Unsupervised feature selection
2. Orthogonalization
Training set
Patient-level feature table
Validation set
41
36–40
Supervised algorithm training
Accuracy evaluation
Phenotype algorithm
Algorithm performance, e.g., AUC, PPV
Predicted probability (prob) and yes/no phenotype for all patients in the data mart
Knowledge sources
UMLS
Concept extraction
NLP dictionary
10–14
Raw data
Data processing
Algorithm training and evaluation

ID	Prob	Y/N
1	0.98	Y
4	0.93	Y
.	.	.
1,000	0.06	N
2,001	0.05	N

Zhang et al. *Nat protocols*. 2019

Feature Engineering: *Manual*

= Clinicians = Informaticists



Identify

Propose targets

Review knowledge

Propose codes

Propose terms

Define

Review code lists

Assemble corpus

Validate code usage

Validate NLP

Specify logic

Implement

Write code

Create NLP

Perform QC

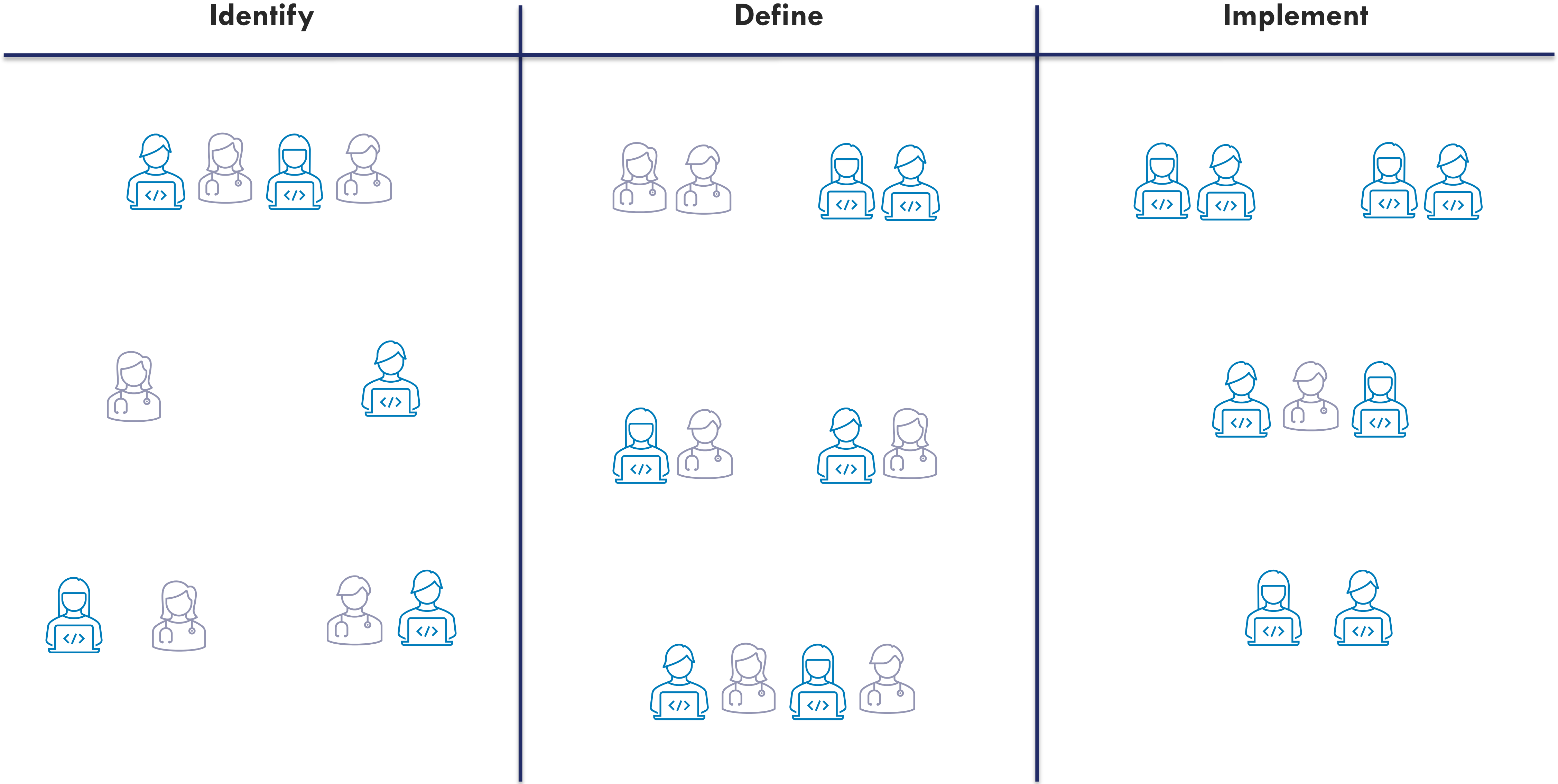
Assemble datasets

Studyid	C0000970_Count	C0001617_Count	C0002895_Count	C0003126_Count	C0003211_Count	C0003241_Count	C0003250_Count	C0003320_Count	C0003451_Count	C0003811_Count
KPWA00001	2	0	0	0	2	0	0	0	0	0
KPWA00003	0	0	0	0	0	0	0	0	0	0
KPWA00005	3	0	0	0	8	0	0	0	0	3
KPWA00013	1	0	0	0	2	0	0	0	0	0
KPWA00008	0	0	0	0	0	0	0	0	0	0
KPWA00006	1	0	0	0	0	0	0	0	0	0
KPWA07733	0	0	0	0	0	0	0	0	1	0
KPWA00012	0	0	0	1	0	0	0	0	0	0
KPWA00010	0	0	0	0	0	0	0	0	0	0
KPWA00038	0	0	0	0	0	0	0	0	0	0
KPWA00041	0	0	0	0	0	0	0	0	0	0
KPWA00014	0	0	0	1	0	0	0	0	0	0
KPWA00011	0	0	0	3	0	0	0	0	0	0
KPWA00050	0	0	0	4	0	0	0	0	0	0
KPWA00018	3	0	0	1	0	0	0	0	1	0
KPWA00054	2	0	0	0	1	0	0	0	0	0

Slide courtesy of David Carrell

= Clinicians = Informaticists

Feature Engineering: *Manual*



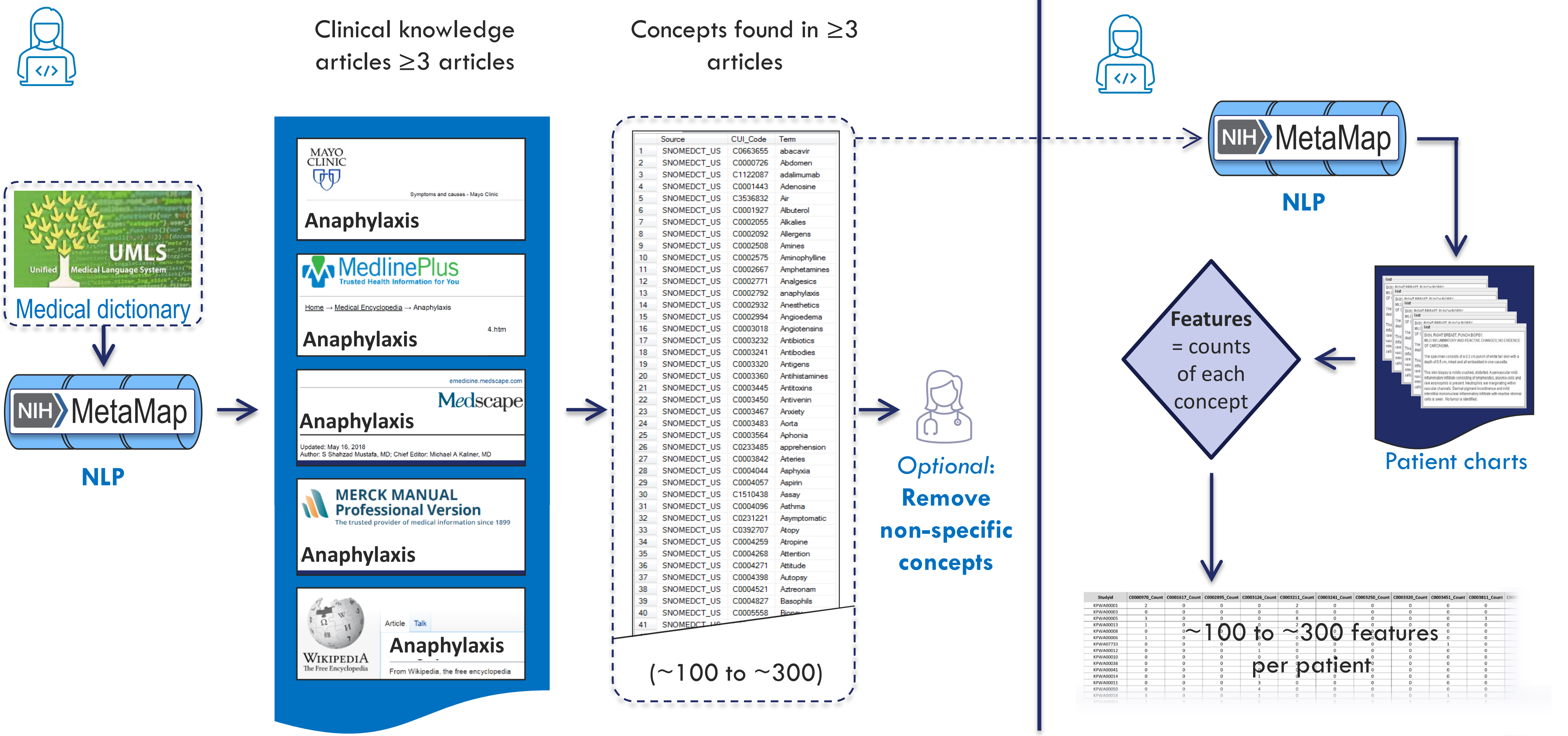
Slide courtesy of David Carrell

Feature Engineering: *Automated*

= Clinicians = Informaticists

Identify & Define*

Implement



Medical dictionary
UMLS
Unified Medical Language System
NIH MetaMap
NLP
Clinical knowledge articles ≥3 articles
MAYO CLINIC
Symptoms and causes - Mayo Clinic
Anaphylaxis
MedlinePlus
Trusted Health Information for You
Home → Medical Encyclopedia → Anaphylaxis
Anaphylaxis
emedicine.medscape.com
Medscape
Anaphylaxis
Updated: May 16, 2018
Author: S Shahzad Mustafa, MD; Chief Editor: Michael A Kaliner, MD
MERCK MANUAL Professional Version
The trusted provider of medical information since 1899
Anaphylaxis
Article Talk
Anaphylaxis
From Wikipedia, the free encyclopedia
Wikipedia The Free Encyclopedia
Concepts found in ≥3 articles

	Source	CUI_Code	Term
1	SNOMEDCT_US	C0663655	abacavir
2	SNOMEDCT_US	C0000726	Abdomen
3	SNOMEDCT_US	C1122087	adalimumab
4	SNOMEDCT_US	C0001443	Adenosine
5	SNOMEDCT_US	C3536832	Air
6	SNOMEDCT_US	C0001927	Albuterol
7	SNOMEDCT_US	C0002055	Alkalies
8	SNOMEDCT_US	C0002092	Allergens
9	SNOMEDCT_US	C0002508	Amines
10	SNOMEDCT_US	C0002575	Aminophylline
11	SNOMEDCT_US	C0002667	Amphetamines
12	SNOMEDCT_US	C0002771	Analgesics
13	SNOMEDCT_US	C0002792	anaphylaxis
14	SNOMEDCT_US	C0002932	Anesthetics
15	SNOMEDCT_US	C0002994	Angioedema
16	SNOMEDCT_US	C0003018	Angiotensins
17	SNOMEDCT_US	C0003232	Antibiotics
18	SNOMEDCT_US	C0003241	Antibodies
19	SNOMEDCT_US	C0003320	Antigens
20	SNOMEDCT_US	C0003360	Antihistamines
21	SNOMEDCT_US	C0003445	Antitoxins
22	SNOMEDCT_US	C0003450	Antivenin
23	SNOMEDCT_US	C0003467	Anxiety
24	SNOMEDCT_US	C0003483	Aorta
25	SNOMEDCT_US	C0003564	Aphonia
26	SNOMEDCT_US	C0233485	apprehension
27	SNOMEDCT_US	C0003842	Arteries
28	SNOMEDCT_US	C0004044	Asphyxia
29	SNOMEDCT_US	C0004057	Aspirin
30	SNOMEDCT_US	C1510438	Assay
31	SNOMEDCT_US	C0004096	Asthma
32	SNOMEDCT_US	C0231221	Asymptomatic
33	SNOMEDCT_US	C0392707	Atopy
34	SNOMEDCT_US	C0004259	Atropine
35	SNOMEDCT_US	C0004268	Attention
36	SNOMEDCT_US	C0004271	Attitude
37	SNOMEDCT_US	C0004398	Autopsy
38	SNOMEDCT_US	C0004521	Aztreonam
39	SNOMEDCT_US	C0004827	Basophils
40	SNOMEDCT_US	C0005558	Biopsy
41	SNOMEDCT_US		

(~100 to ~300)

Optional: **Remove non-specific concepts**

Features = counts of each concept

Studyid	C0000970_Count	C0001617_Count	C0002895_Count	C0003126_Count	C0003211_Count	C0003241_Count	C0003250_Count	C0003320_Count	C0003451_Count	C0003811_Count
KPWA00001	2	0	0	0	2	0	0	0	0	0
KPWA00003	0	0	0	0	0	0	0	0	0	0
KPWA00005	3	0	0	0	8	0	0	0	0	3
KPWA00013	1	0	[illegible]	0	2	[illegible]	0	0	0	0
KPWA00008	0	0	0	[illegible]	[illegible]	0	[illegible]	0	0	0
KPWA00006	1	0	[illegible]	[illegible]	0	[illegible]	[illegible]	0	0	0
KPWA07733	0	0	0	0	0	0	0	0	1	0
KPWA00012	0	0	0	1	0	0	0	0	0	0
KPWA00010	0	0	0	0	0	[illegible]	0	0	0	0
KPWA00038	0	0	0	[illegible]	[illegible]	[illegible]	0	0	0	0
KPWA00041	0	0	0	[illegible]	[illegible]	[illegible]	0	0	0	0
KPWA00014	0	0	0	1	0	0	0	0	0	0
KPWA00011	0	0	0	3	0	0	0	0	0	0
KPWA00050	0	0	0	4	0	0	0	0	0	0
KPWA00018	3	0	0	1	0	0	0	0	1	0
KPWA00054	2	0	0	0	1	0	0	0	0	0

~100 to ~300 features per patient

* Yu et al. JAMIA 2015

Slide courtesy of David Carrell

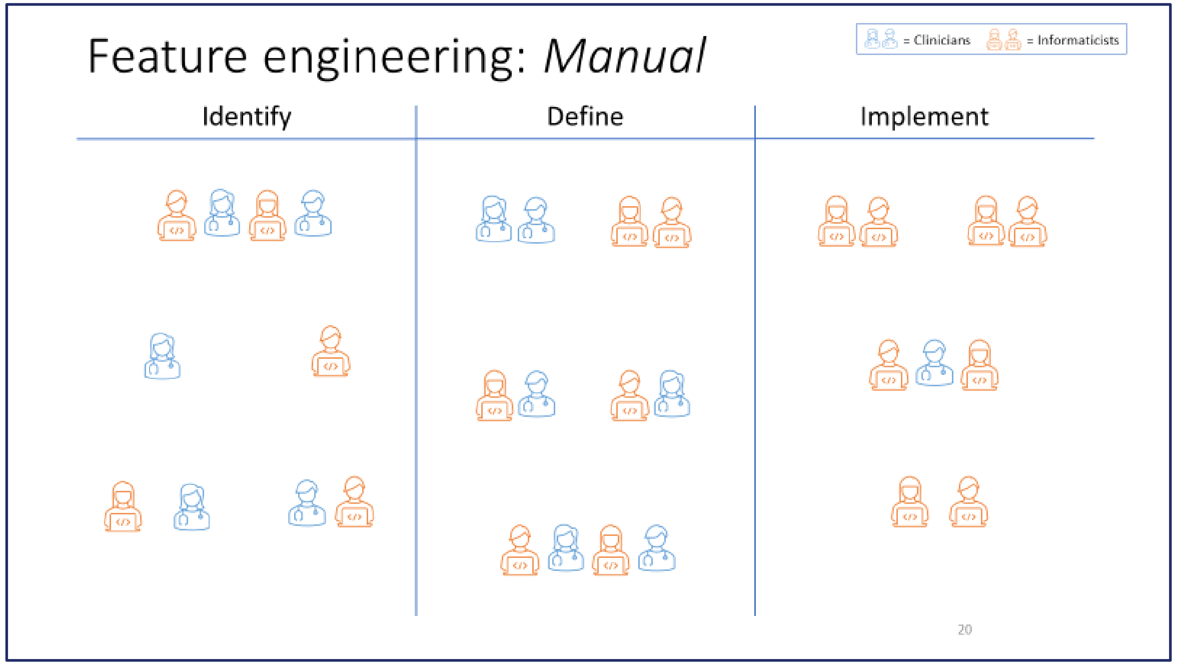
Feature Engineering: *Automated*

= Clinicians = Informaticists

Identify & Define	Implement
Optional	

Slide courtesy of David Carrell

Feature Engineering: Manual vs. Automated

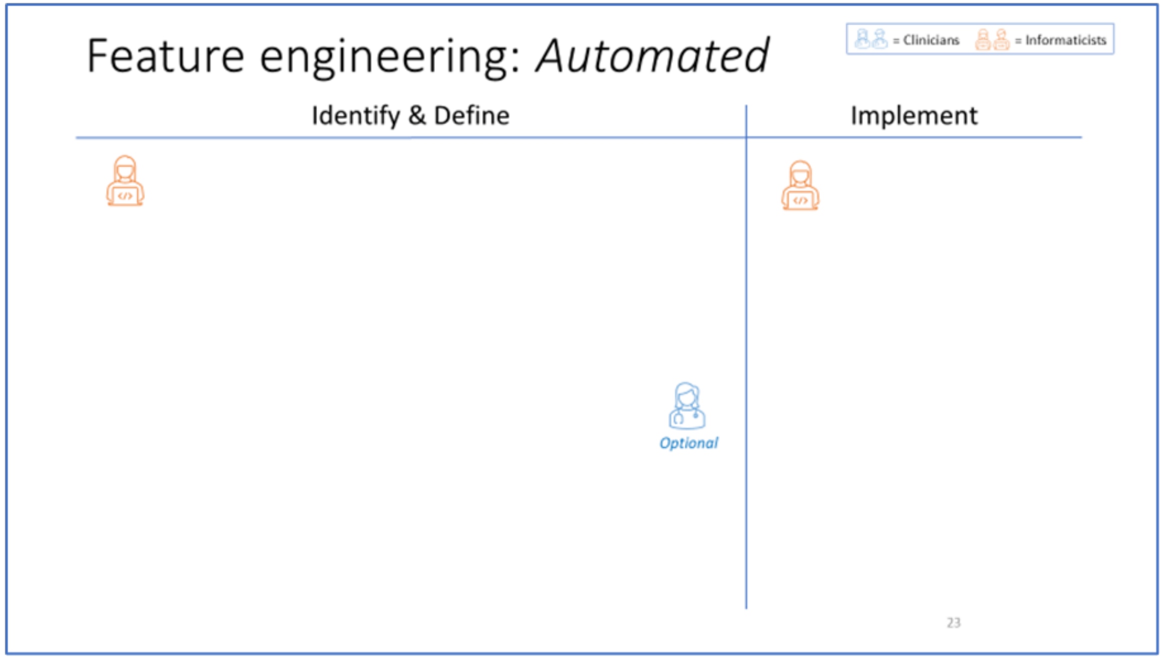


Feature engineering: *Manual*

= Clinicians = Informaticists

Identify | Define | Implement

20



Feature engineering: *Automated*

= Clinicians = Informaticists

Identify & Define | Implement

Optional

23

Slide courtesy of David Carrell

Breakout activity

What are some of the strengths and limitations of the automated approach versus manual approach?

Strengths and limitations

Automation advantages:

- Short development time
- Low/no expenditure for domain expertise
- Reduced operator dependence
- Highly replicable

Automation limitations:

- Unclear if the performance is compromised versus a manual approach

Will it work? As a starting point? As an overall solution?

Feature Engineering Example: Automated (NLP)

High-severity COVID-19 disease (red, N=51)

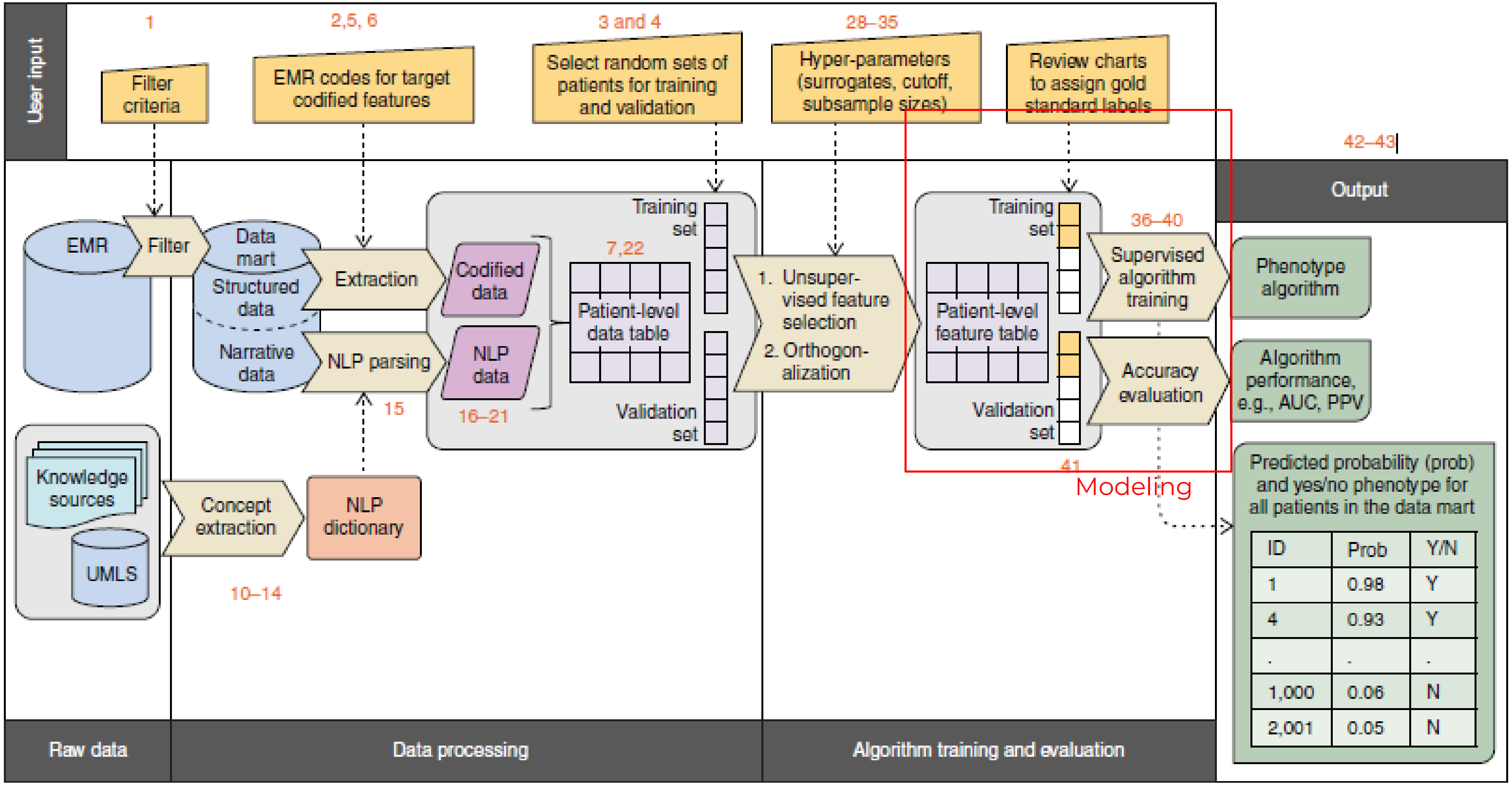
#	CONCEPT	CUI
1	acetaminophen	C0000970
2	Adrenal Cortex Hormones	C0001617
3	air	C3536832
4	Anemia, Sickle Cell	C0002895
5	Angiotensin II receptor antagonist	C0521942
6	animal allergen extracts	C3540698
7	Anosmia	C0003126
8	Antibodies	C0003241
9	Antibodies, Neutralizing	C0475463
10	Antibody studies (procedure)	C0580327
11	Antibody Therapy	C0281176
12	Antigens	C0003320
13	Anti-Inflam. Agents, Non-Steroidal	C0003211
14	Antimicrobial Susceptibility Result	C2827758
15	Antiviral Agents	C0003451
16	Arthralgia	C0003862
17	Asymptomatic (finding)	C0231221
18	At home	C4534363
19	baricitinib	C4044947
20	Blood Clot	C0302148
21	Blood coagulation tests	C0005790
22	Body mass index procedure	C0005893
23	Brain Diseases	C0006111
24	Bronchoalveolar Lavage	C1535502
25	Cardiac Arrhythmia	C0003811
26	Cardiomyopathies	C0878544
27	Cerebrovascular accident	C0038454
28	Chemical Association	C0596306
29	Chest CT	C0202823
30	Chest Pain	C0008031
31	Chills	C0085593
32	chloroquine	C0008269
33	Chronic Kidney Diseases	C1561643
	Chronic Obstructive Airway	

#	CONCEPT	CUI
41	Coronary Arteriosclerosis	C0010054
42	Coughing	C0010200
43	COVID19 (disease)	C5203670
44	COVID-19 drug treatment	C5244048
45	C-reactive protein	C0006560
46	Critical Illness	C0010340
47	Cystic Fibrosis	C0010674
48	Death (finding)	C1306577
49	Death Related to Adverse Event	C1705232
50	Decreased translucency	C0029053
51	Delta-Like Protein 1, human	C3815527
52	Device Alert Level - Serious	C1551395
53	Device Alert Level - Critical	C1551396
54	dexamethasone	C0011777
55	Diabetes Mellitus	C0011849
56	Diabetes Mell., Non-Ins-Depend.	C0011860
57	Diagnostic Imaging	C0011923
58	Diarrhea and vomiting, symptom	C0474496
59	Diffuse Optical Imaging	C3899379
60	Down Syndrome	C0013080
61	Dyspnea	C0013404
62	Emergency Situation	C0013956
63	Environmental air flow	C0042491
64	Extracorp. Membrane Oxygen.	C0015357
65	Fatigue	C0015672
66	Ferritin	C0015879

#	CONCEPT	CUI
81	Hypersensitivity	C0020517
82	Hypertensive disease	C0020538
83	Hypoxemia	C0700292
84	Hypoxia	C0242184
85	Immune System Finding	C1291764
86	Immunocompromised Host	C0085393
87	Immunoglobulins	C0021027
88	Improved - answer to question	C4084203
89	Inflammation	C0021368
90	Interferons	C0021747
91	interleukin-6	C0021760
92	Isolation procedure	C0204727
93	ivermectin	C0022322
94	Lactate Dehydrogenase	C0022917
95	lopinavir / ritonavir	C0939237
96	Loss of taste or smell	C5382033
97	Lung consolidation	C0521530
98	Lung diseases	C0024115
99	Lymphopenia	C0024312
100	M Protein, multiple myeloma	C0700271
101	Malaise	C0231218
102	Mechanical ventilation	C0199470
103	Mechanical Ventilator	C0042497
104	methylprednisolone	C0025815
105	Mild Adverse Event	C1513302
106	Monoclonal Antibodies	C000325

#	CONCEPT	CUI
121	Pharyngitis	C0031350
122	Plain chest X-ray	C0039985
123	Plasma Product	C4521445
124	Pneumonia	C0032285
125	Pneumonia, Viral	C0032310
126	Pressure- physical agent	C0033095
127	Pulmonary (intended site)	C4522268
128	Quarantine	C0034386
129	receptor	C0597357
130	Reduction procedure	C1293152
131	remdesivir	C4726677
132	Respiration Disorders	C0035204
133	Respiratory distress	C0476273
134	Respiratory Distress Synd., Adult	C0035222
135	Respiratory Failure	C1145670
136	Respiratory System Finding	C0425442
137	Rhinorrhea	C1260880
138	RNA, Messenger	C0035696
139	Self-Quarantine	C5392942
140	Septic Shock	C0036983
141	Severe (severity modifier)	C0205082
142	Severe Acute Resp. Syndrome	C1175175
143	Severe disease	C4740692
144	Shock	C0036974
145	Signs and Symptoms, Respiratory	C0037090
146	Sneezing	C0037383
147	Steroids	C0038317
148	Supplemental oxygen	C4534306
149	Symptom mild	C0436343
150	Symptom severe	C0436345
151	Symptomatic Presentation	C5238876
152	Thromboembolism	C0040038

High throughput phenotyping - steps


User input
1
Filter criteria
2,5, 6
EMR codes for target codified features
3 and 4
Select random sets of patients for training and validation
28–35
Hyper-parameters (surrogates, cutoff, subsample sizes)
Review charts to assign gold standard labels
42–43
Output
EMR
Filter
Data mart
Structured data
Narrative data
Extraction
NLP parsing
15
Codified data
NLP data
16–21
7,22
Patient-level data table
Training set
Validation set
1. Unsupervised feature selection
2. Orthogonalization
Patient-level feature table
Training set
Validation set
41
36–40
Supervised algorithm training
Accuracy evaluation
Modeling
Phenotype algorithm
Algorithm performance, e.g., AUC, PPV
Predicted probability (prob) and yes/no phenotype for all patients in the data mart
Knowledge sources
UMLS
Concept extraction
NLP dictionary
10–14
Raw data
Data processing
Algorithm training and evaluation

ID	Prob	Y/N
1	0.98	Y
4	0.93	Y
.	.	.
1,000	0.06	N
2,001	0.05	N

Zhang et al. *Nat protocols*. 2019

Modeling Overview (Illustrative)

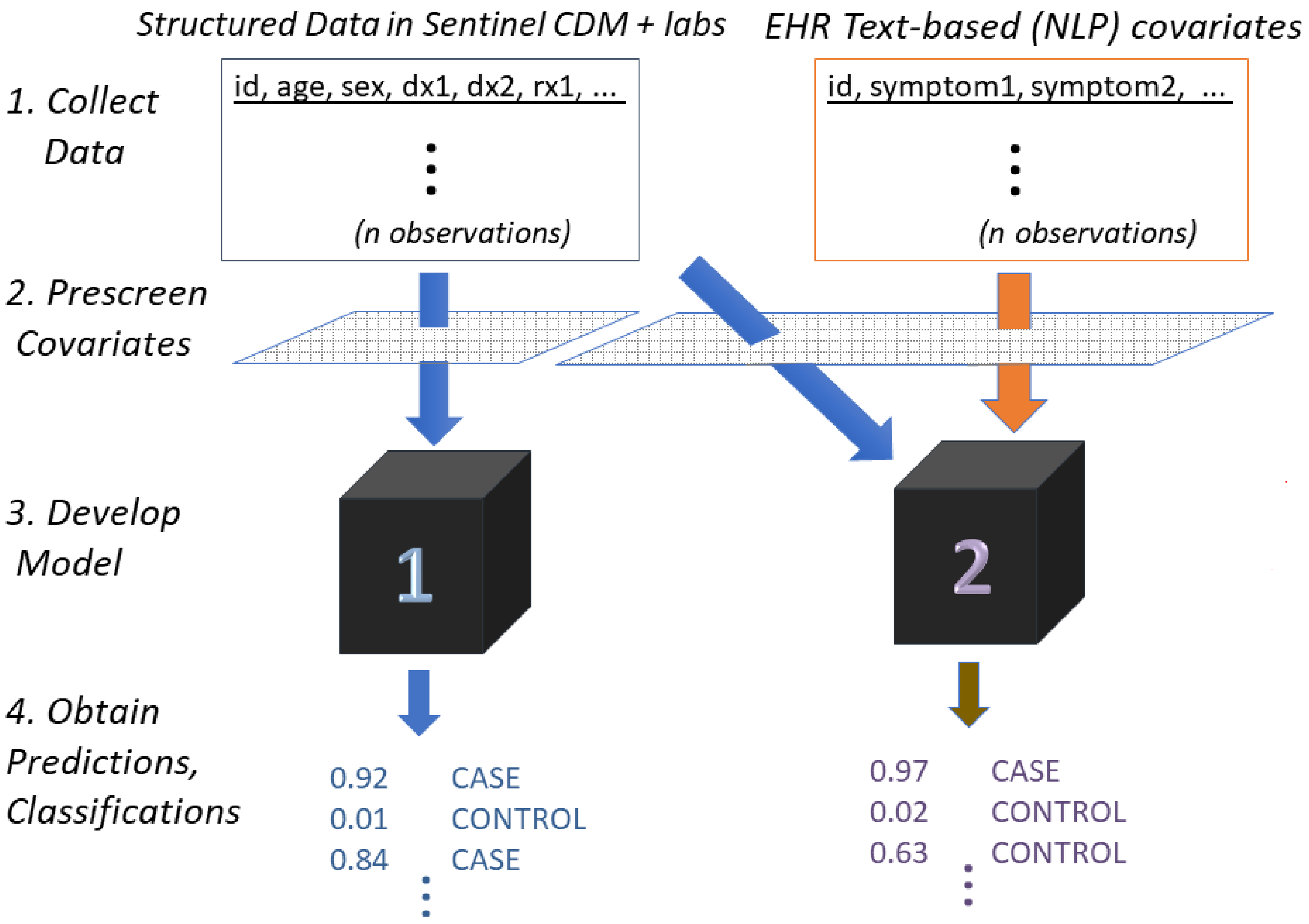

Structured Data in Sentinel CDM + labs
EHR Text-based (NLP) covariates
1. Collect Data
id, age, sex, dx1, dx2, rx1, ...
⋮
(n observations)
id, symptom1, symptom2, ...
⋮
(n observations)
2. Prescreen Covariates
3. Develop Model
1
2
4. Obtain Predictions, Classifications
0.92 CASE
0.01 CONTROL
0.84 CASE
⋮
0.97 CASE
0.02 CONTROL
0.63 CONTROL
⋮

Image courtesy of Susan Gruber

Modeling Overview (Illustrative)

Dimension reduction	
1	Retain All
2	PAM
3	LASSO

×

Algorithms	
1	GLM
2	Elastic net
3	XGBoost v1
4	XGBoost v2
5	BART v1
6	BART v2
7	Neural net v1
8	Neural net v2

=

Combinations	
1	GLM-Retain-All
2	GLM-PAM
3	GLM-LASSO
4	Elastic-net-Retain-All
5	Elastic-net-PAM
6	Elastic-net-LASSO
7	XGBoost-v1-Retain-All
8	XGBoost-v1-PAM
9	XGBoost-v1-LASSO
10	XGBoost-v2-Retain-All
11	XGBoost-v2-PAM
12	XGBoost-v2-LASSO
13	BART-v1-Retain-All
14	BART-v1-PAM
15	BART-v1-LASSO
16	BART-v2-Retain-All
17	BART-v2-PAM
18	BART-v2-LASSO
19	Neural-net-v1-Retain-All
20	Neural-net-v1-PAM
21	Neural-net-v1-LASSO
22	Neural-net-v2-Retain-All
23	Neural-net-v2-PAM
24	Neural-net-v2-LASSO

+

Super Learner	
25	A weighted combination of the other 24 combinations

Slide courtesy of David Carrell

Example Results: Computable Phenotyping for Anaphylaxis

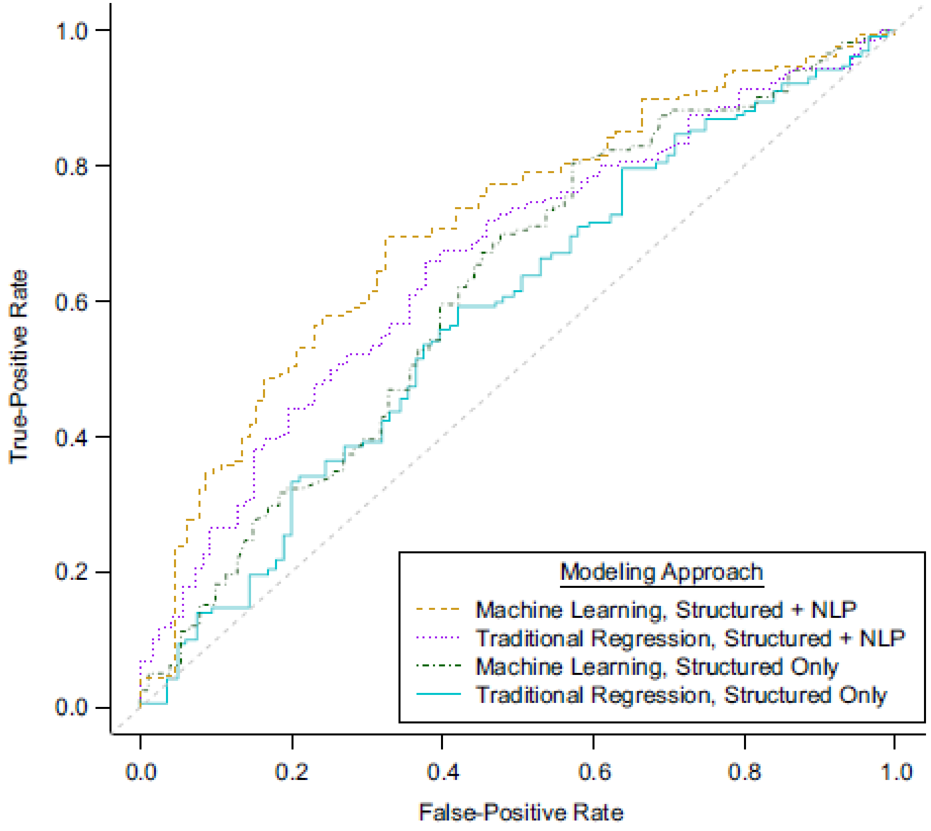
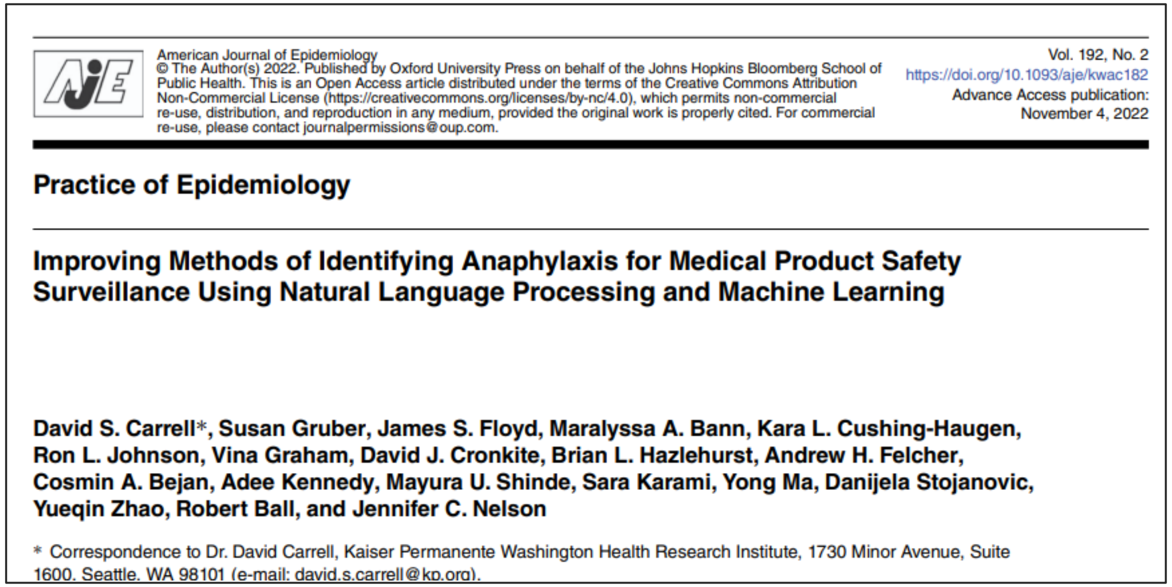


Figure 1. Weighted cross-validated area under the receiver operating characteristic curve for Kaiser Permanente Washington algorithms identifying actual anaphylaxis events in Kaiser Permanente Washington data (2015–2019) using the best machine-learning approach applied to structured and all natural language processing (NLP) data, traditional logistic regression approach applied to structured and all NLP data, machine-learning approach applied to structured data only, and traditional logistic regression approach applied to structured data only.

Carrell et al. AJE 2023

Computable Phenotyping & NLP Activities in Sentinel



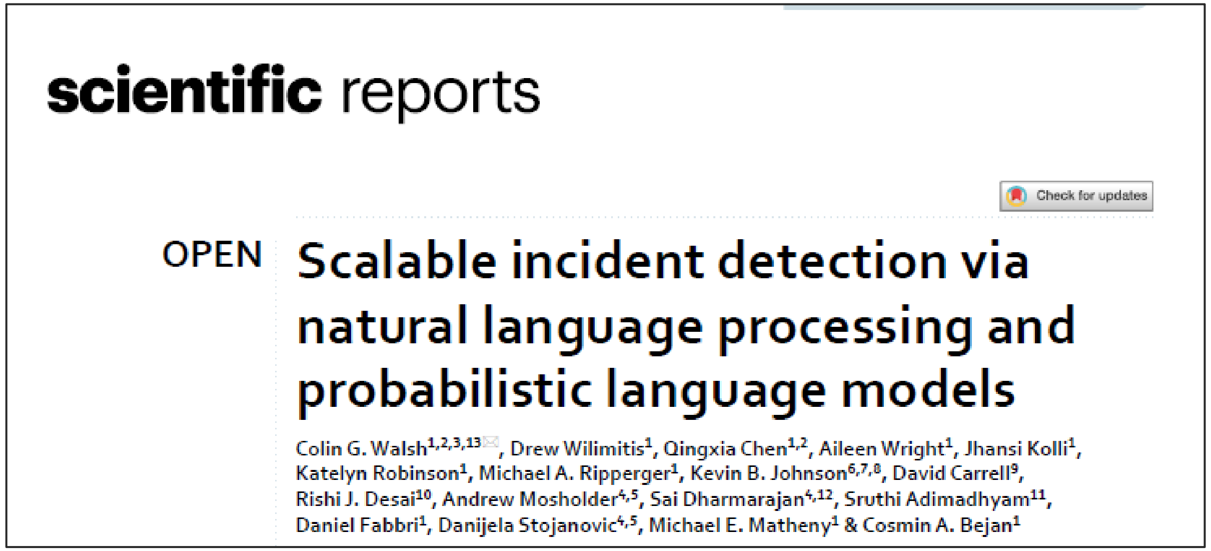
American Journal of Epidemiology
Vol. 192, No. 2
https://doi.org/10.1093/aje/kwac182
Advance Access publication:
November 4, 2022

Practice of Epidemiology

Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning

David S. Carrell*, Susan Gruber, James S. Floyd, Maralyssa A. Bann, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Brian L. Hazlehurst, Andrew H. Felcher, Cosmin A. Bejan, Adee Kennedy, Mayura U. Shinde, Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball, and Jennifer C. Nelson

* Correspondence to Dr. David Carrell, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: david.s.carrell@kp.org).



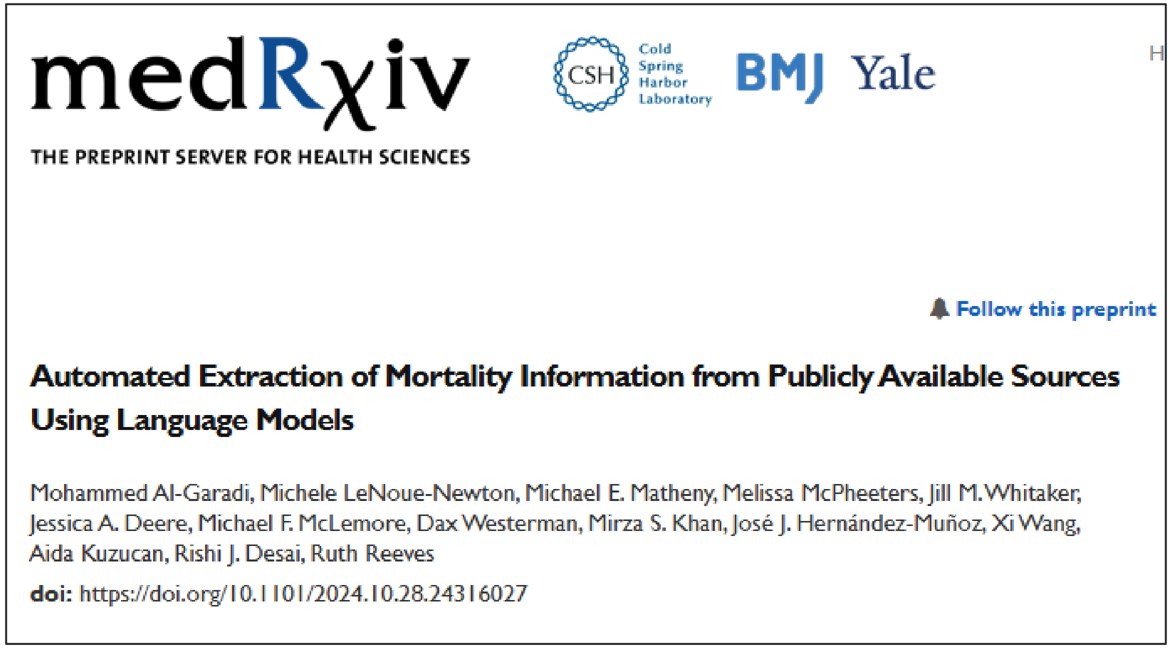
scientific reports

Check for updates

OPEN

Scalable incident detection via natural language processing and probabilistic language models

Colin G. Walsh[1,2,3,13], Drew Wilimitis[1], Qingxia Chen[1,2], Aileen Wright[1], Jhansi Kolli[1], Katelyn Robinson[1], Michael A. Ripperger[1], Kevin B. Johnson[6,7,8], David Carrell[9], Rishi J. Desai[10], Andrew Mosholder[4,5], Sai Dharmarajan[4,12], Sruthi Adimadhyam[11], Daniel Fabbri[1], Danijela Stojanovic[4,5], Michael E. Matheny[1] & Cosmin A. Bejan[1]



medRχiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

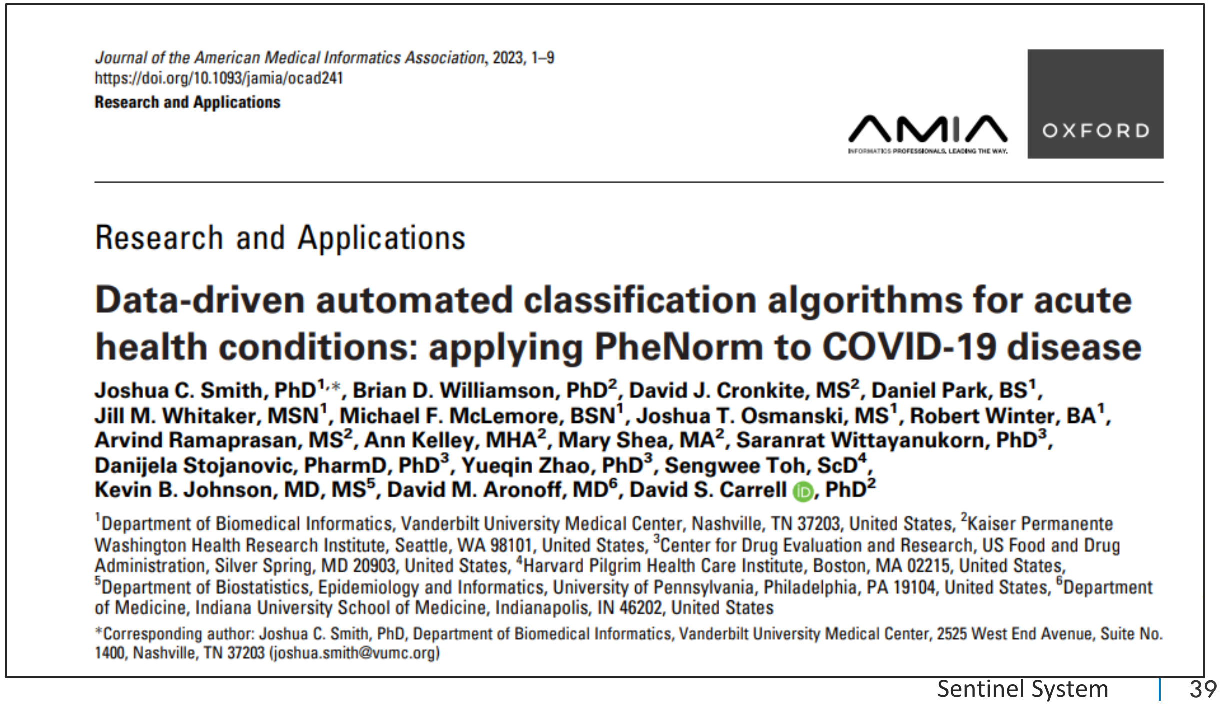
H

Follow this preprint

Automated Extraction of Mortality Information from Publicly Available Sources Using Language Models

Mohammed Al-Garadi, Michele LeNoue-Newton, Michael E. Matheny, Melissa McPheeters, Jill M. Whitaker, Jessica A. Deere, Michael F. McLemore, Dax Westerman, Mirza S. Khan, José J. Hernández-Muñoz, Xi Wang, Aida Kuzucan, Rishi J. Desai, Ruth Reeves

doi: https://doi.org/10.1101/2024.10.28.24316027



Journal of the American Medical Informatics Association, 2023, 1–9
https://doi.org/10.1093/jamia/ocad241
Research and Applications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

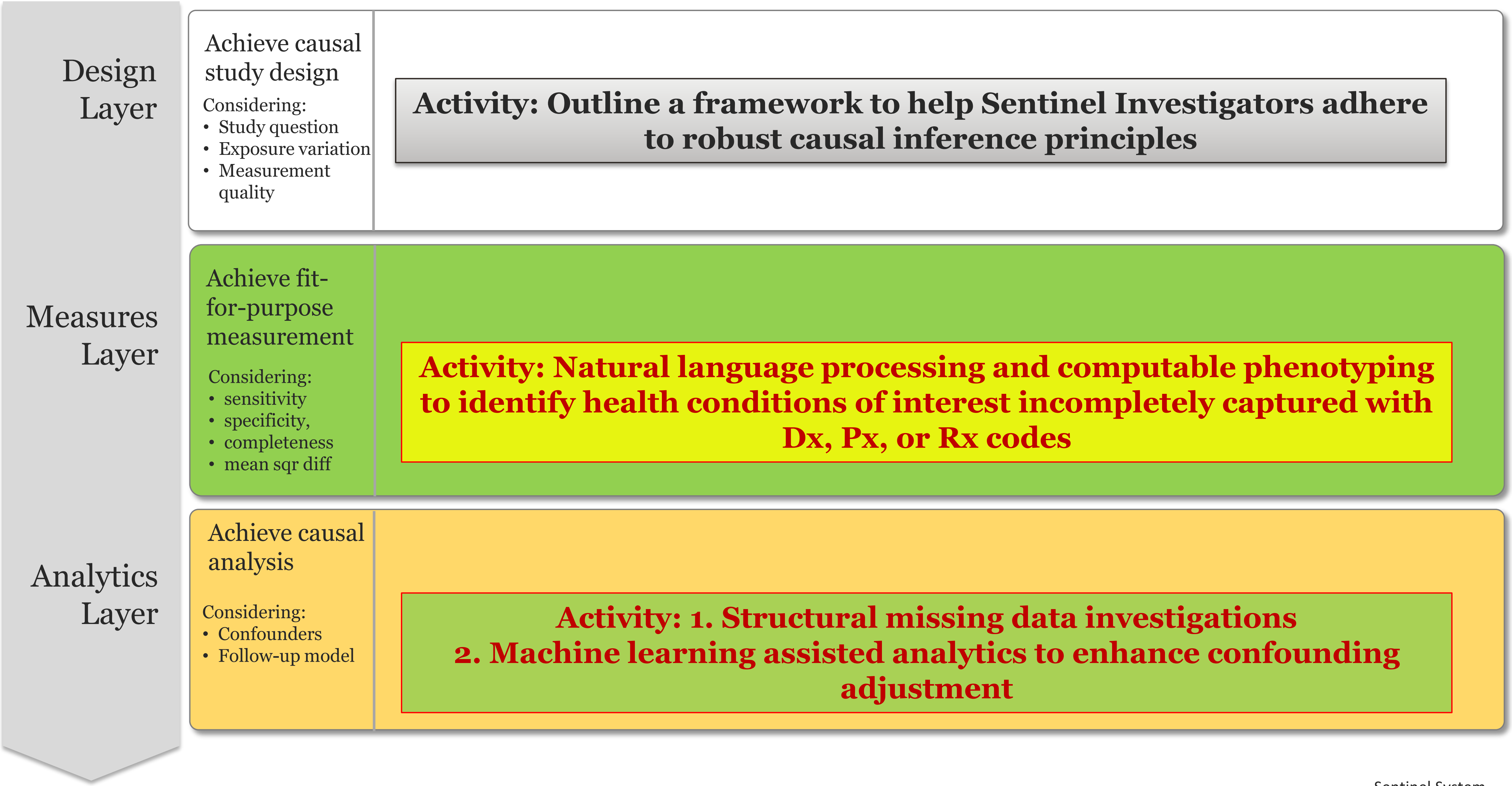
Data-driven automated classification algorithms for acute health conditions: applying PheNorm to COVID-19 disease

Joshua C. Smith, PhD[1,*], Brian D. Williamson, PhD[2], David J. Cronkite, MS[2], Daniel Park, BS[1], Jill M. Whitaker, MSN[1], Michael F. McLemore, BSN[1], Joshua T. Osmanski, MS[1], Robert Winter, BA[1], Arvind Ramaprasan, MS[2], Ann Kelley, MHA[2], Mary Shea, MA[2], Saranrat Wittayanukorn, PhD[3], Danijela Stojanovic, PharmD, PhD[3], Yueqin Zhao, PhD[3], Sengwee Toh, ScD[4], Kevin B. Johnson, MD, MS[5], David M. Aronoff, MD[6], David S. Carrell, PhD[2]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, [2]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States, [3]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States, [4]Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States, [5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, United States, [6]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, United States

*Corresponding author: Joshua C. Smith, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite No. 1400, Nashville, TN 37203 (joshua.smith@vumc.org)

Causal Inference Requirements: Role of Advanced Methods



Design Layer

Achieve causal study design

Considering:
- Study question
- Exposure variation
- Measurement quality

Activity: Outline a framework to help Sentinel Investigators adhere to robust causal inference principles

Measures Layer

Achieve fit-for-purpose measurement

Considering:
- sensitivity
- specificity,
- completeness
- mean sqr diff

Activity: Natural language processing and computable phenotyping to identify health conditions of interest incompletely captured with Dx, Px, or Rx codes

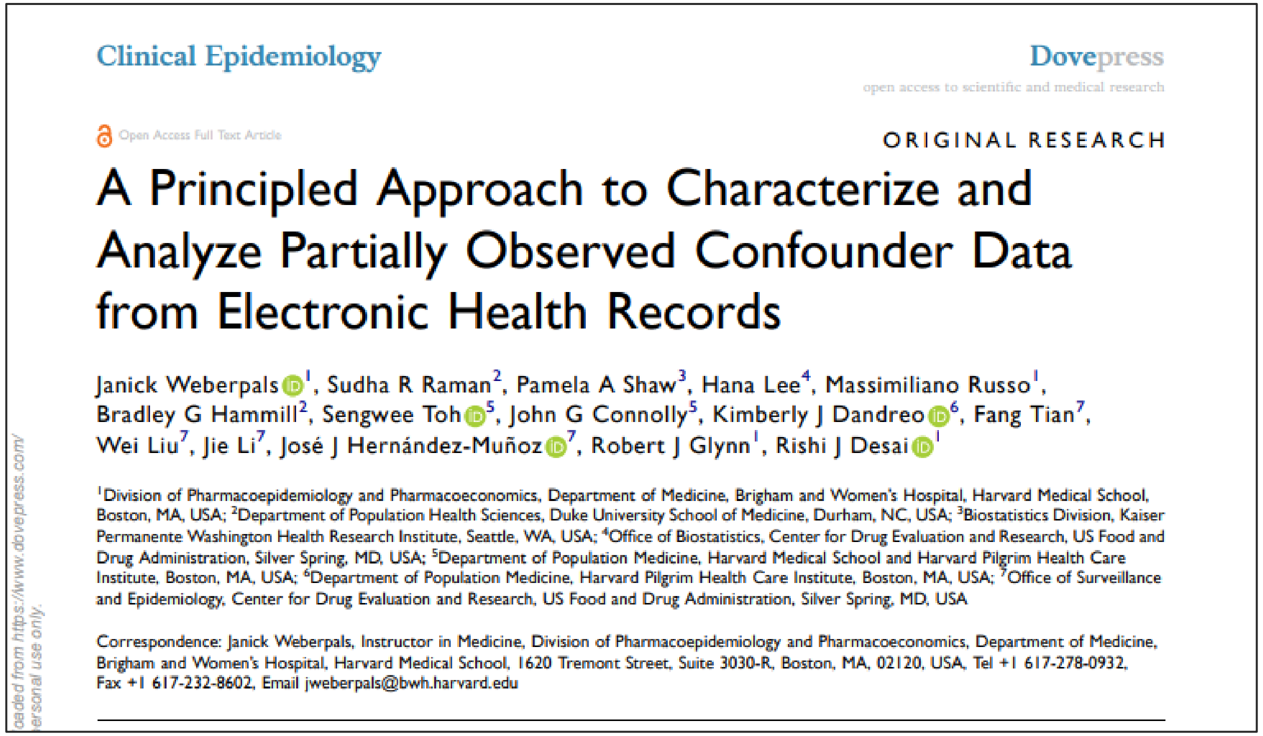
Analytics Layer

Achieve causal analysis

Considering:
- Confounders
- Follow-up model

Activity: 1. Structural missing data investigations
2. Machine learning assisted analytics to enhance confounding adjustment

Activity: 1. Structural Missing Data Investigations



Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

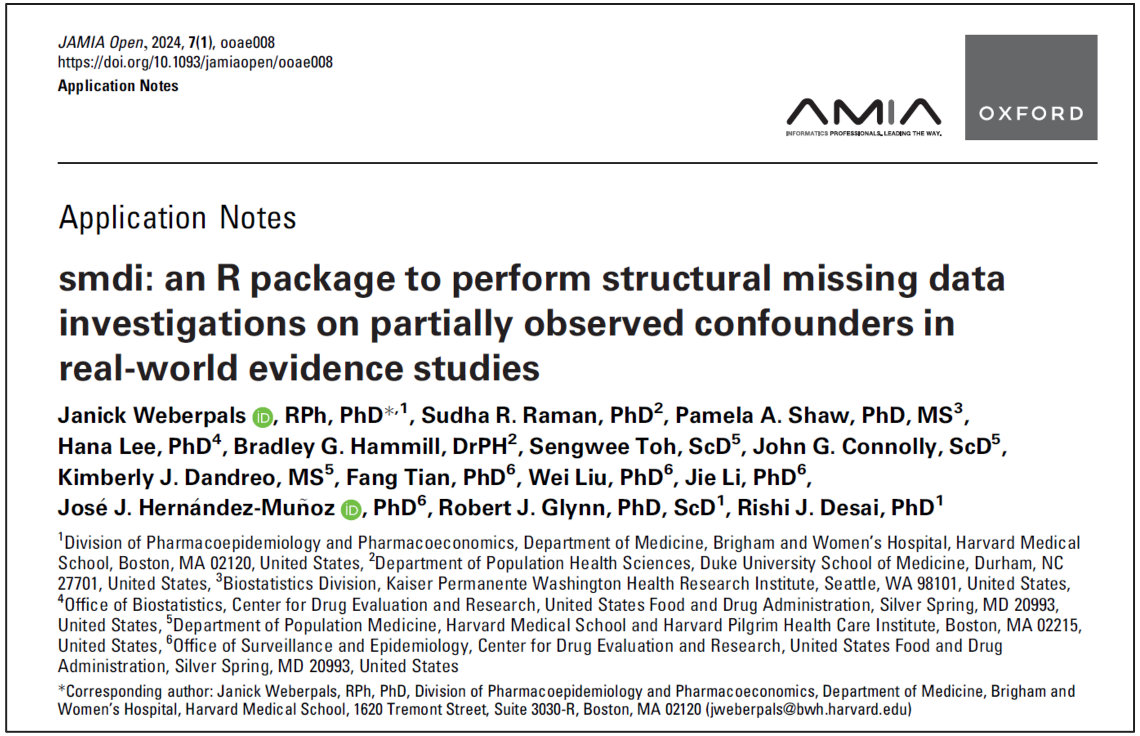
ORIGINAL RESEARCH

A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records

Janick Weberpals[1], Sudha R Raman[2], Pamela A Shaw[3], Hana Lee[4], Massimiliano Russo[1], Bradley G Hammill[2], Sengwee Toh[5], John G Connolly[5], Kimberly J Dandreo[6], Fang Tian[7], Wei Liu[7], Jie Li[7], José J Hernández-Muñoz[7], Robert J Glynn[1], Rishi J Desai[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [2]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA; [3]Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA; [4]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA; [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA; [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA; [7]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

Correspondence: Janick Weberpals, Instructor in Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030-R, Boston, MA, 02120, USA, Tel +1 617-278-0932, Fax +1 617-232-8602, Email jweberpals@bwh.harvard.edu



JAMIA Open, 2024, **7(1)**, ooae008
https://doi.org/10.1093/jamiaopen/ooae008
Application Notes

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Application Notes

smdi: an R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies

Janick Weberpals, RPh, PhD*,[1], Sudha R. Raman, PhD[2], Pamela A. Shaw, PhD, MS[3], Hana Lee, PhD[4], Bradley G. Hammill, DrPH[2], Sengwee Toh, ScD[5], John G. Connolly, ScD[5], Kimberly J. Dandreo, MS[5], Fang Tian, PhD[6], Wei Liu, PhD[6], Jie Li, PhD[6], José J. Hernández-Muñoz, PhD[6], Robert J. Glynn, PhD, ScD[1], Rishi J. Desai, PhD[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, United States, [2]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC 27701, United States, [3]Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States, [4]Office of Biostatistics, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD 20993, United States, [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States, [6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD 20993, United States

*Corresponding author: Janick Weberpals, RPh, PhD, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030-R, Boston, MA 02120 (jweberpals@bwh.harvard.edu)

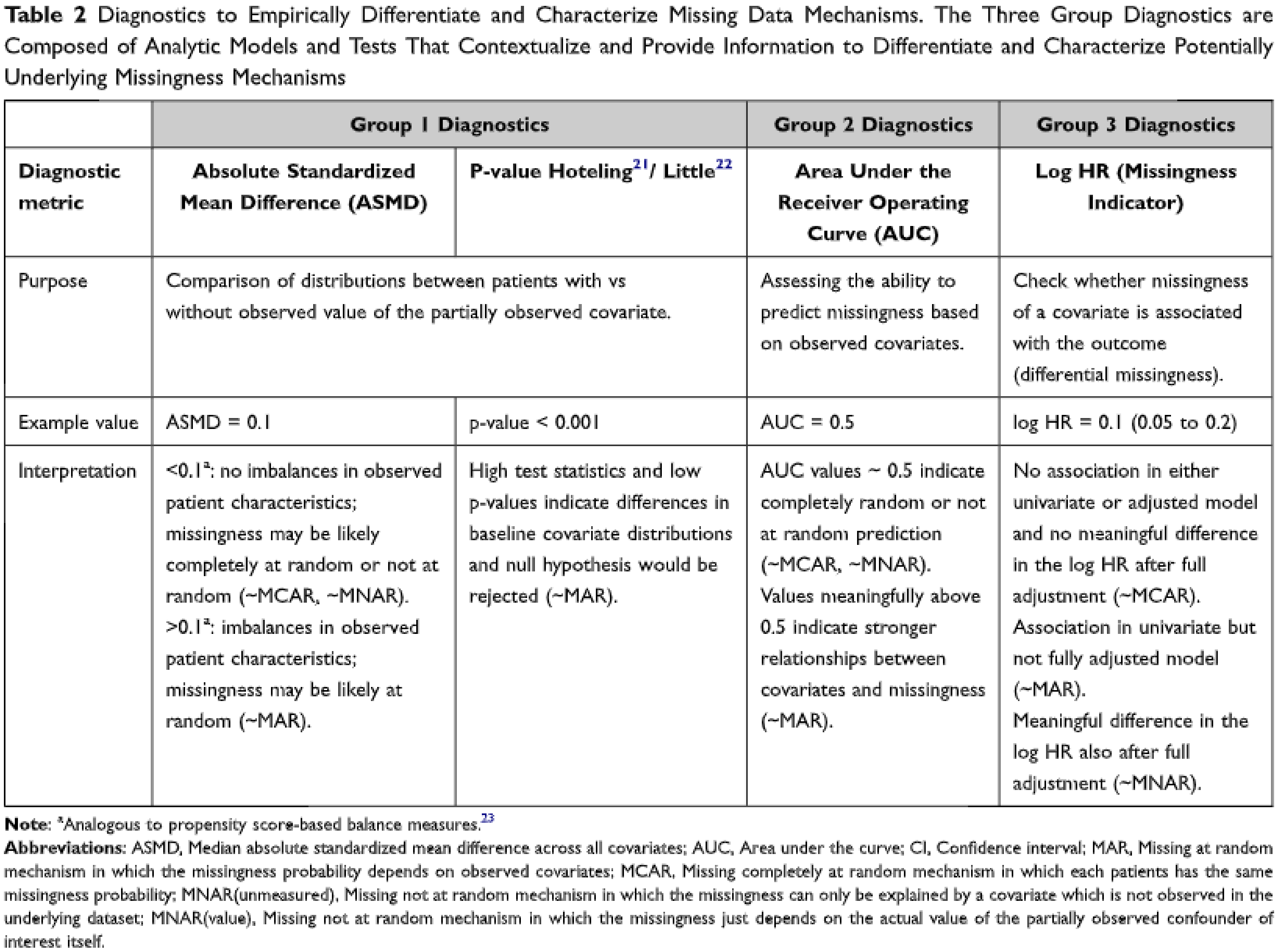


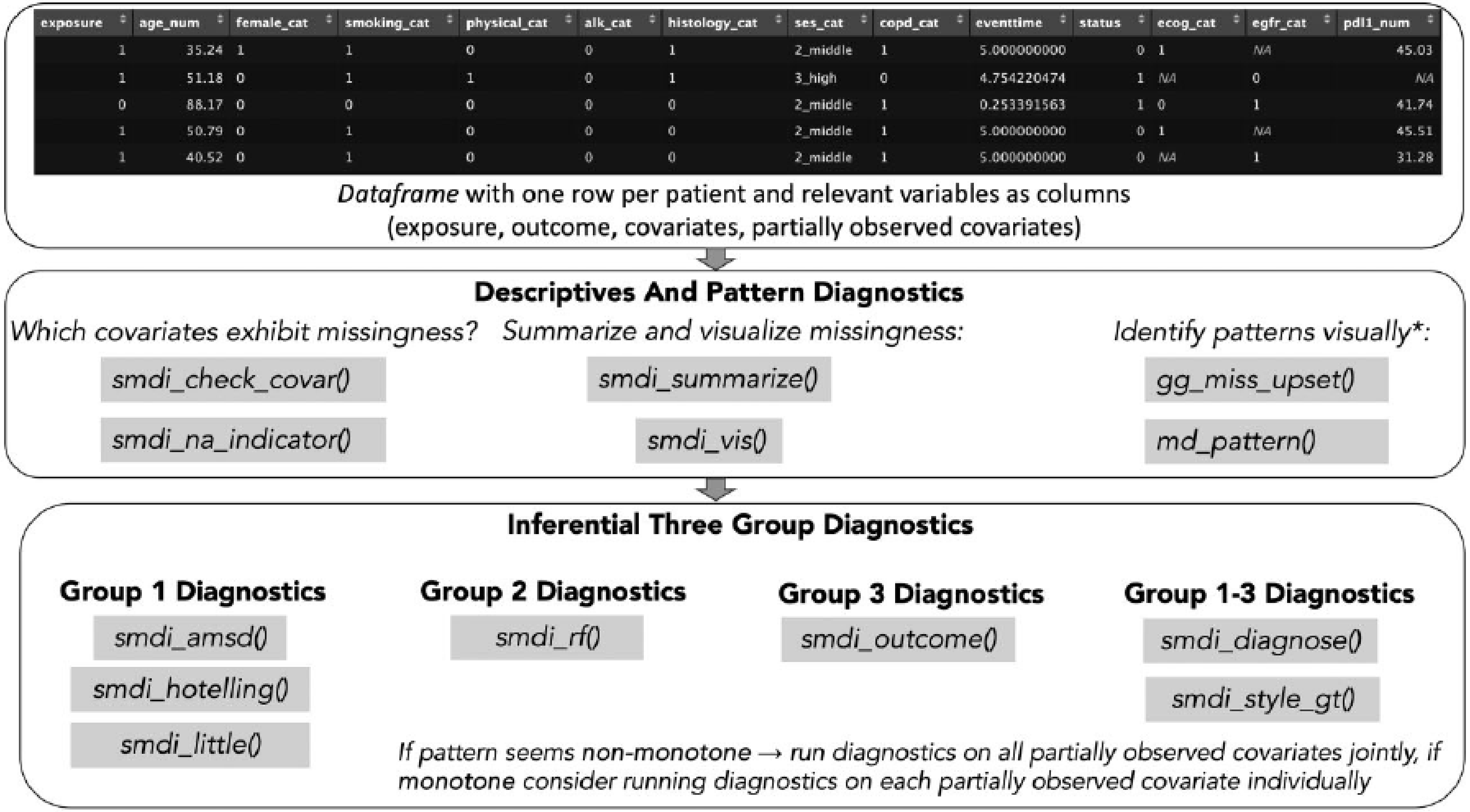
Table 2 Diagnostics to Empirically Differentiate and Characterize Missing Data Mechanisms. The Three Group Diagnostics are Composed of Analytic Models and Tests That Contextualize and Provide Information to Differentiate and Characterize Potentially Underlying Missingness Mechanisms

	Group 1 Diagnostics		**Group 2 Diagnostics**	**Group 3 Diagnostics**
Diagnostic metric	**Absolute Standardized Mean Difference (ASMD)**	**P-value Hoteling[21]/ Little[22]**	**Area Under the Receiver Operating Curve (AUC)**	**Log HR (Missingness Indicator)**
Purpose	Comparison of distributions between patients with vs without observed value of the partially observed covariate.		Assessing the ability to predict missingness based on observed covariates.	Check whether missingness of a covariate is associated with the outcome (differential missingness).
Example value	ASMD = 0.1	p-value < 0.001	AUC = 0.5	log HR = 0.1 (0.05 to 0.2)
Interpretation	<0.1[a]: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR). >0.1[a]: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).	High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR).	AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR). Values meaningfully above 0.5 indicate stronger relationships between covariates and missingness (~MAR).	No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR). Association in univariate but not fully adjusted model (~MAR). Meaningful difference in the log HR also after full adjustment (~MNAR).

Note: [a] Analogous to propensity score-based balance measures.[23]

Abbreviations: ASMD, Median absolute standardized mean difference across all covariates; AUC, Area under the curve; CI, Confidence interval; MAR, Missing at random mechanism in which the missingness probability depends on observed covariates; MCAR, Missing completely at random mechanism in which each patients has the same missingness probability; MNAR(unmeasured), Missing not at random mechanism in which the missingness can only be explained by a covariate which is not observed in the underlying dataset; MNAR(value), Missing not at random mechanism in which the missingness just depends on the actual value of the partially observed confounder of interest itself.

Weberpals J, Raman SR, Shaw PA, et al. A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records. Clin Epidemiol. 2024;16:329-343. Published 2024 May 21. doi:10.2147/CLEP.S436131



exposure	age_num	female_cat	smoking_cat	physical_cat	alk_cat	histology_cat	ses_cat	copd_cat	eventtime	status	ecog_cat	egfr_cat	pdl1_num
1	35.24	1	1	0	0	1	2_middle	1	5.000000000	0	1	*NA*	45.03
1	51.18	0	1	1	0	1	3_high	0	4.754220474	1	*NA*	0	*NA*
0	88.17	0	0	0	0	0	2_middle	1	0.253391563	1	0	1	41.74
1	50.79	0	1	0	0	0	2_middle	1	5.000000000	0	1	*NA*	45.51
1	40.52	0	1	0	0	0	2_middle	1	5.000000000	0	*NA*	1	31.28

Dataframe with one row per patient and relevant variables as columns
(exposure, outcome, covariates, partially observed covariates)

Descriptives And Pattern Diagnostics

Which covariates exhibit missingness?

smdi_check_covar()

smdi_na_indicator()

Summarize and visualize missingness:

smdi_summarize()

smdi_vis()

Identify patterns visually:*

gg_miss_upset()

md_pattern()

Inferential Three Group Diagnostics

Group 1 Diagnostics

smdi_amsd()

smdi_hotelling()

smdi_little()

Group 2 Diagnostics

smdi_rf()

Group 3 Diagnostics

smdi_outcome()

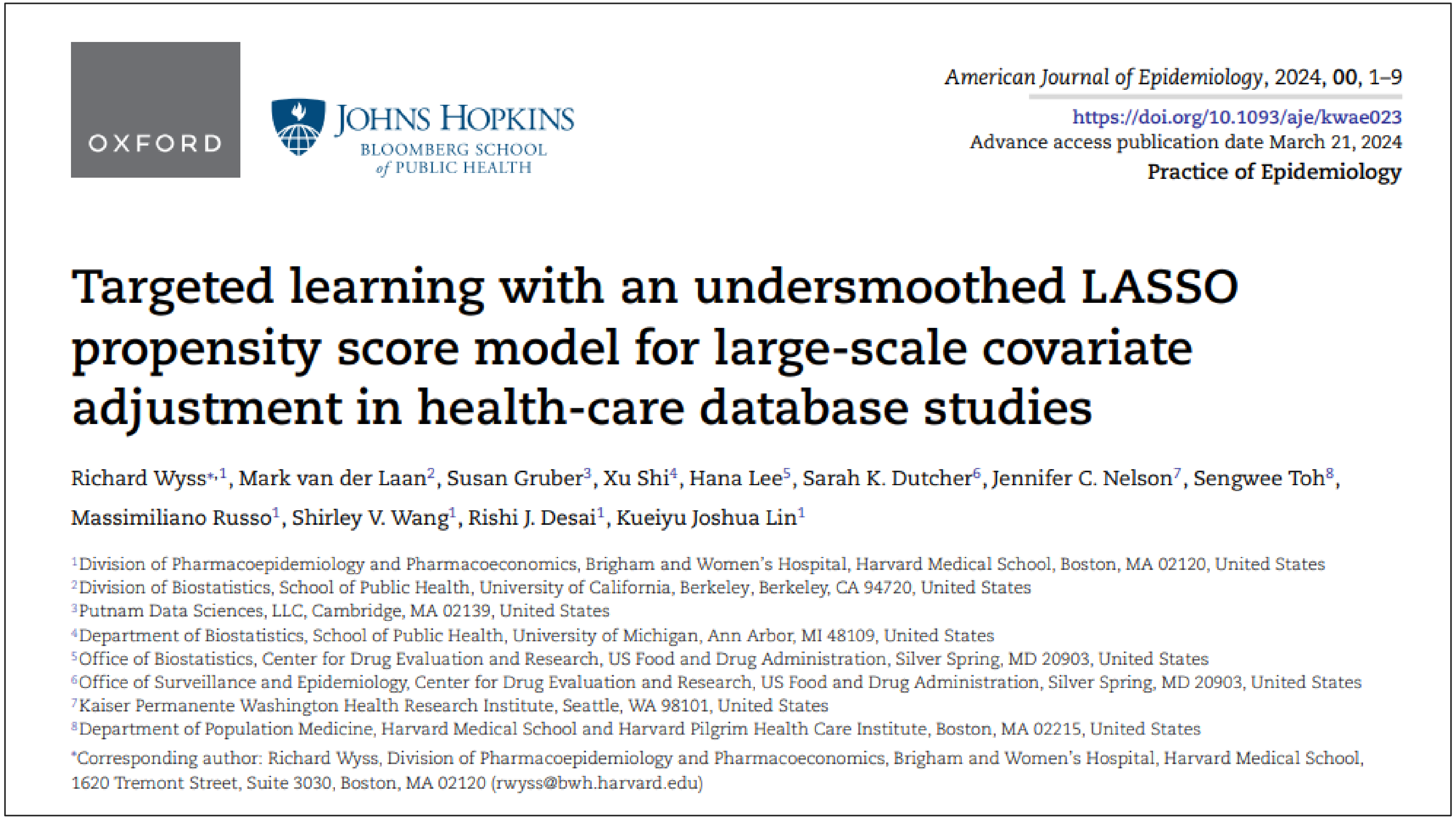
Group 1-3 Diagnostics

smdi_diagnose()

smdi_style_gt()

If pattern seems non-monotone → run diagnostics on all partially observed covariates jointly, if monotone consider running diagnostics on each partially observed covariate individually

Activity 2. Machine Learning Assisted Analytics to Enhance Confounding Adjustment



American Journal of Epidemiology, 2024, **00**, 1–9
https://doi.org/10.1093/aje/kwae023
Advance access publication date March 21, 2024
Practice of Epidemiology

Targeted learning with an undersmoothed LASSO propensity score model for large-scale covariate adjustment in health-care database studies

Richard Wyss*,[1], Mark van der Laan[2], Susan Gruber[3], Xu Shi[4], Hana Lee[5], Sarah K. Dutcher[6], Jennifer C. Nelson[7], Sengwee Toh[8], Massimiliano Russo[1], Shirley V. Wang[1], Rishi J. Desai[1], Kueiyu Joshua Lin[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, United States
[2]Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, United States
[3]Putnam Data Sciences, LLC, Cambridge, MA 02139, United States
[4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, United States
[5]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States
[7]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States
[8]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States
*Corresponding author: Richard Wyss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (rwyss@bwh.harvard.edu)

Leveraging Unstructured EHRs for Large-Scale Proxy Adjustment

(ultra-high dimensional data)

NLP tools turn free-text notes from EHR data into structured features that can serve as proxy confounding adjustment

Table. Example data structure for 2 cohort studies that include linked claims with NLP generated EHR features

	Sample Size			Outcome	Baseline Covariates		
Cohort	N_{Total}	$N_{Treated}$	$N_{Comparator}$	N_{Total}	N_{Total}	$N_{Predefined}$	$N^{**}_{Proxies}$
Study 1:[A]	21,343	13,576	7,767	899 (4.2%)	14,937	91	**14,846**
Study 2:[B]	35,031	12,872	22,159	251 (0.7%)	12,464	91	**12,373**

[A] Study 1: Effect of NSAIDs versus opioids on acute kidney injury

[B] Study 2: Effect of high vs low-dose proton pump inhibitors (PPIs) on gastrointestinal bleeding

** Number of claims and EHR features after screening those with prevalence <0.001

Propensity Score (PS) Models with Ultra-High Dimensional Data

Overfit PS models that include too many variables could lead to reduced covariate overlap, positivity violations

Some degree of dimension reduction is necessary– BUT ideally, without compromising bias reducing properties

Various approaches for fitting PS models available for this purpose

1. Traditional LASSO (L1 regularization with loss function based on minimizing prediction error of treatment)
2. Outcome adaptive LASSO (forces all variables that predict the outcome in the LASSO PS model)
3. Collaborative controlled LASSO (variable selection based on minimizing empirical loss of the estimate for the target causal parameter i.e treatment effect)
4. Collaborative controlled, outcome adaptive LASSO (combination of 2 & 3)

Wyss et al. AJE (In Press)

Propensity Score Models with Ultra-High Dimensional Data

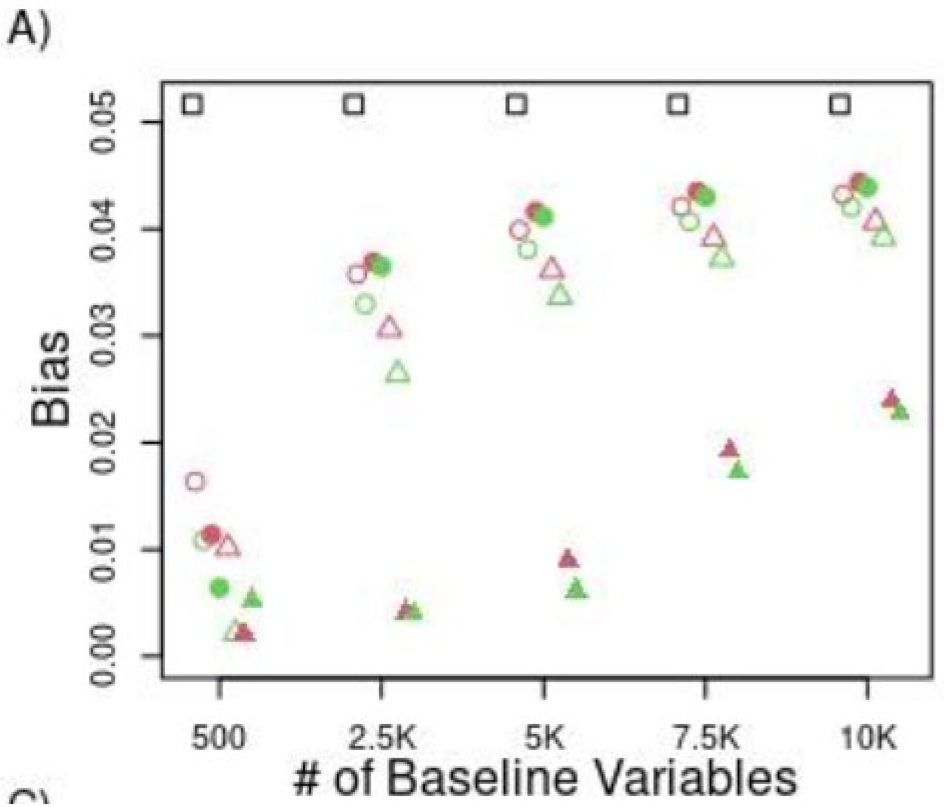
Use of cross-fitting to manage overfitting

- Randomly split the data into 10 equally sized non-overlapping groups. The given Lasso model trained in 9 of the groups. The trained model was then applied to the held-out group to assign PS.
- Same models described on the previous slides with cross-fitting

5. Traditional LASSO (L1 regularization with loss function based on minimizing prediction error of treatment)
6. Outcome adaptive LASSO (forces all variables that predict the outcome in the LASSO PS model)
7. Collaborative controlled LASSO (variable selection based on minimizing empirical loss of the estimate for the target causal parameter i.e treatment effect)
8. Collaborative controlled, outcome adaptive LASSO (combination of 2 & 3)

Wyss et al. AJE (In Press)

Propensity Score Models with Ultra-High Dimensional Data: Simulation Results



No cross-fit of trt model

Cross-fit of trt model

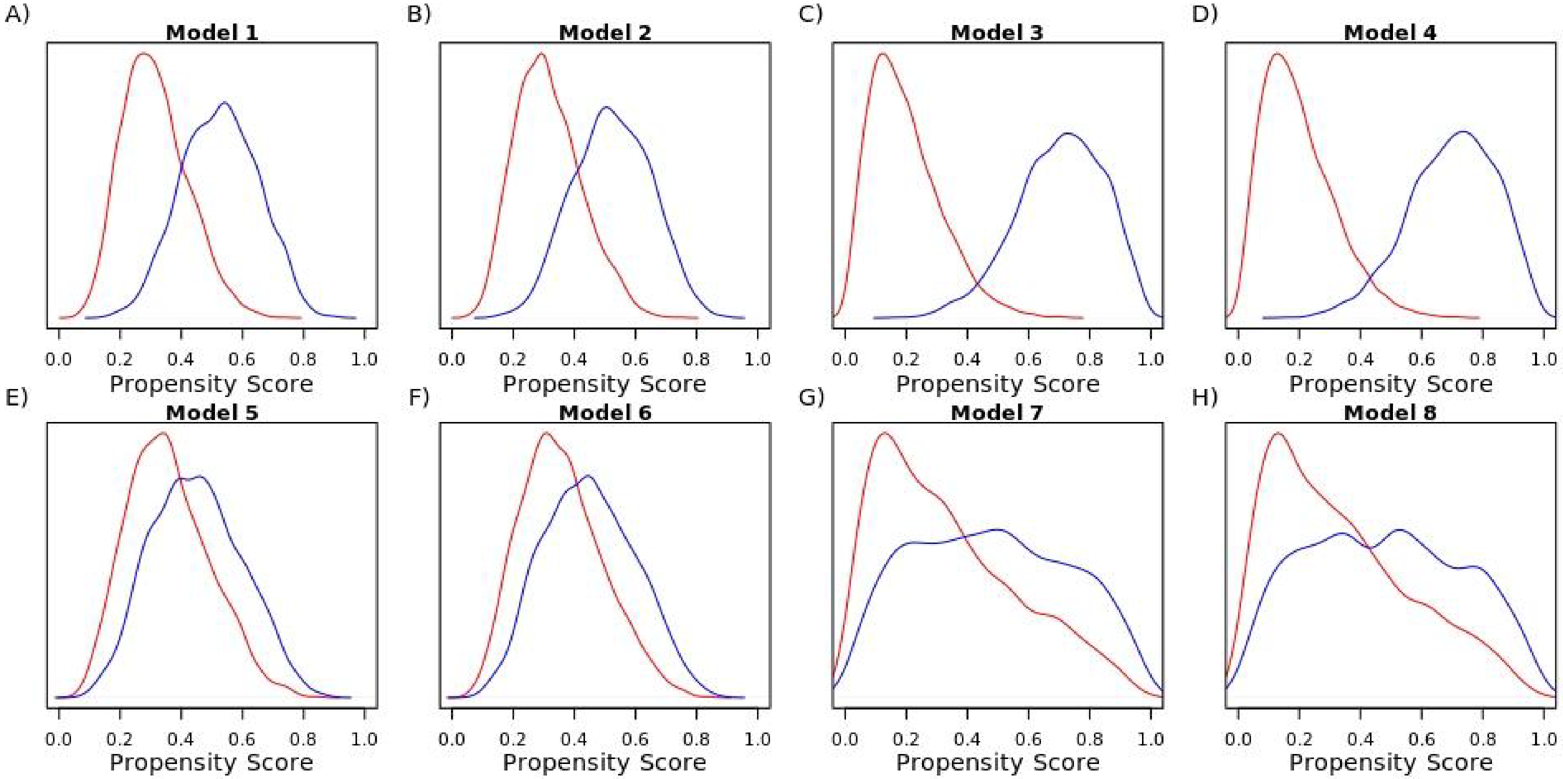
As overfitting increases, models with cross-fitting, especially 7 & 8, tend to outperform other models

Take home point:

Advanced analytical approaches can allow for enhanced confounding adjustment using granular data from EHRs

Wyss et al. AJE 2024

Propensity Score Models with Ultra-High Dimensional Data: Simulation Results



Propensity score distributions for treated (blue) and comparator (red) groups for one simulated dataset consisting of 9,500 spurious variables and 500 baseline confounders that ranged in the strength of covariate effects on treatment and outcome (Scenario 5 consisting of 10,000 total baseline variables)

- □ Crude (Unadjusted)
- ○ Model 1: Lasso
- ○ Model 2: Outcome Adaptive Lasso (OAL)
- ● Model 3: Collaborative Controlled Lasso
- ● Model 4: Collaborative Conrolled OAL
- △ Model 5: Cross-Fit (CF) Lasso
- △ Model 6: CF OAL
- ▲ Model 7: Collaborative Controlled CF Lasso
- ▲ Model 8: Collaborative Controlled CF OAL

No cross-fit of trt model

Cross-fit of trt model

What (likely) explains robust performance:

Cross fitting allows for reducing non-overlap for the overfit collaborative-controlled models

Wyss et al. AJE 2024

Software and other materials available for use

1. Analytical and data processing software

Goal	Tool	References
Descriptive evaluation and diagnostics for missingness in EHR-based confounding variables	SMDI (IC-developed R package)	Weberpals J, Raman SR, Shaw PA, et al. smdi: An R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. *JAMIA Open*. 2024;7(1):ooae008. doi:10.1093/jamiaopen/ooae008.
Simulation-based descriptive analysis for an unmeasured confounding to assess its impact on study results	Sim.BA (IC-developed R package)	Desai RJ, Bradley MC, Lee H et al. A simulation-based bias analysis to assess the impact of unmeasured confounding when designing nonrandomized database studies. *Am J Epidemiol*. 2024 Nov 4;193(11):1600-1608. doi: 10.1093/aje/kwae102. PMID: 38825336.
Statistical adjustment for a partially measured confounding variable with multiple imputations	MICE, MatchThem (Existing R packages used by prior Sentinel investigations)	Pishgar F, Greifer N, Leyrat C, Stuart E. MatchThem:: Matching and weighting after multiple imputation. Published online September 24, 2020. doi:10.48550/arXiv.2009.11772.
Statistical adjustment for a partially measured confounding variable with two-stage approaches (TMLE/Raking weights)	MarginalEffects (IC-developed reusable R codes)	Williamson BD, Krakauer C, Johnson E, et al. Assessing treatment effects in observational data with missing confounders: A comparative study of practical doubly-robust and traditional missing data methods. *arXiv*.2024/12/19;doi:10.48550/arXiv.2412.15012
Large-scale propensity scores with undersmoothing for high-dimensional confounding adjustment	CI5 (IC-developed reusable R codes)	Wyss et al. Targeted learning with an undersmoothed lasso propensity score model for large-scale covariate adjustment in healthcare database studies. *Am J Epidemiol*. 2024 doi:10.1093/aje/kwae023.
NLP assisted chart review tool	CORA (Clinical Optimized Record Annotation)	Wang et al. (In Review)

2. Phenotype library and other models for off-the-shelf use

Phenotype	Description	References
COVID19	Algorithm using elements from structured and unstructured EHRs (Phenorm approach)	Smith JC, Williamson BD, Cronkite DJ, Park D, Whitaker JM, McLemore MF, Osmanski JT, Winter R, Ramaprasan A, Kelley A, Shea M. Data-driven automated classification algorithms for acute health conditions: applying PheNorm to COVID-19 disease. Journal of the American Medical Informatics Association. 2024 Mar 1;31(3):574-82.
Suicidal attempt Sleep related behaviors	NLP score-based approach, requires free-text notes	Walsh CG, Wilimitis D, Chen Q, Wright A, Kolli J, Robinson K, Ripperger MA, Johnson KB, Carrell D, Desai RJ, Mosholder A, Dharmarajan S, Adimadhyam S, Fabbri D, Stojanovic D, Matheny ME, Bejan CA. Scalable incident detection via natural language processing and probabilistic language models. Sci Rep. 2024 Oct 8;14(1):23429. doi: 10.1038/s41598-024-72756-7. PMID: 39379449; PMCID: PMC11461638.
Acute pancreatitis	Algorithm using structured dx, labs, and free-text; a version without free-text features is also validated, with has similar PPV	Bann et al. (in review)
Acute kidney injury	Algorithm using structured features from claims data only (PhenoSCALE approach)	Pradhan et al. (in review)
Anaphylaxis	Algorithm using elements from structured and unstructured EHRs (Phenorm approach)	Smith et al. (in review)
Cause of death	Model using structured and free-text EHR data to probabilistically assign cause of death	Al-Garadi et al. (in review)

Summary

Summary

- Large scale data infrastructure of the RWE-DE where EHRs are linked to claims data from 6 diverse data sources covering 25.5 million lives is available for use in Sentinel

- RWE-DE will offer opportunities to improve the validity of studies of medical products in clinical practice and to expand the range of questions that can be answered through Sentinel