Methodological Advances in Regulatory Real World Evidence Generation Systems: Perspectives from Sentinel and DARWIN-EU

Rishi Desai, PhD[1], Robert Ball, MD, MPH[2], ScM, Patrice Verpillat, MD, MPH, PhD[3], Sebastian Schneeweiss, MD, ScD[1], Talita Duarte-Salles, PhD[4], Daniel Prieto-Alhambra, MD, Msc, PhD[5]

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States

[2] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration Silver Spring, United States

[3] European Medicines Agency, Amsterdam, Netherlands

[4] Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

[5] University of Oxford, United Kingdom

Panel introduction

Robert Ball, MD, MPH, ScM
Deputy Director, Office of Surveillance and Epidemiology (OSE), Center for Drug Evaluation and Research (CDER), FDA

Sebastian Schneeweiss
Professor of Medicine
Division of Pharmacoepidemiology and Pharmacoeconomics
Harvard Medical School
Brigham & Women's Hospital

Rishi J Desai, MS, PhD
Associate Professor of Medicine
Division of Pharmacoepidemiology and Pharmacoeconomics
Harvard Medical School
Brigham & Women's Hospital

Patrice Verpillat, MD, MPH, PhD
Head of the Real-World Evidence (RWE) Workstream of the Data Analytics Taskforce at the European Medicines Agency (EMA)

Talita Duarte-Salles, MPH, PhD
Epidemiologist
Senior Epidemiologist, IDIAPJGol and Assistant professor of Medical Informatics, Erasmus MC

Dani Prieto-Alhambra, MD MSc(Oxf) PhD
Section Head - Health Data Sciences at Botnar Research Centre and Professor at University of Oxford and Erasmus MC

Disclaimer

- This work was supported by Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA).

- The views expressed in this presentation represent those of the presenter and do not necessarily represent the official views of the U.S. FDA.

Sentinel

Data Infrastructure Update

(Sebastian Schneeweiss)

Bias as an Obstacle to Causal Inference

1. Confounding
2. Selection bias
3. **Information bias**

The error mechanisms

- **Outcome** misclassification (MC)/ measurement error (ME)
  - Random
  - Differential
  - Dependent
- **Exposure** MC/ ME
  - Random
  - Differential
- **Confounder** MC/ ME
  - Random
  - Differential

Porta M. A Dictionary of Epidemiology 5th ed. 2008

Data Quality Map

Information Bias Mechanisms

Data Curation & Provenance

Measurement

Validation studies

Measurement Characteristics

Quant Bias Analysis

Data Quality Dimensions Relevant for Causal Inference

Data Continuity	Patients receive treatments/assessments by a range of providers during their journey through the healthcare continuum: • More longitudinally complete data throughout the care continuum will reduce surveillance related issues/bias
Data Granularity	Detailed clinical and other information improves the measurement of exposure, confounders, and outcomes: • More granular data are preferred for a broad range of etiologic studies
Data Chronology	The accurate chronology of confounder, exposure and outcome measurement is critical for causal inference: • Unclear chronology can lead to a range of biases, like reverse causation, adjustment for intermediates, immortal time

Schneeweiss, Desai, Ball AJE 2024

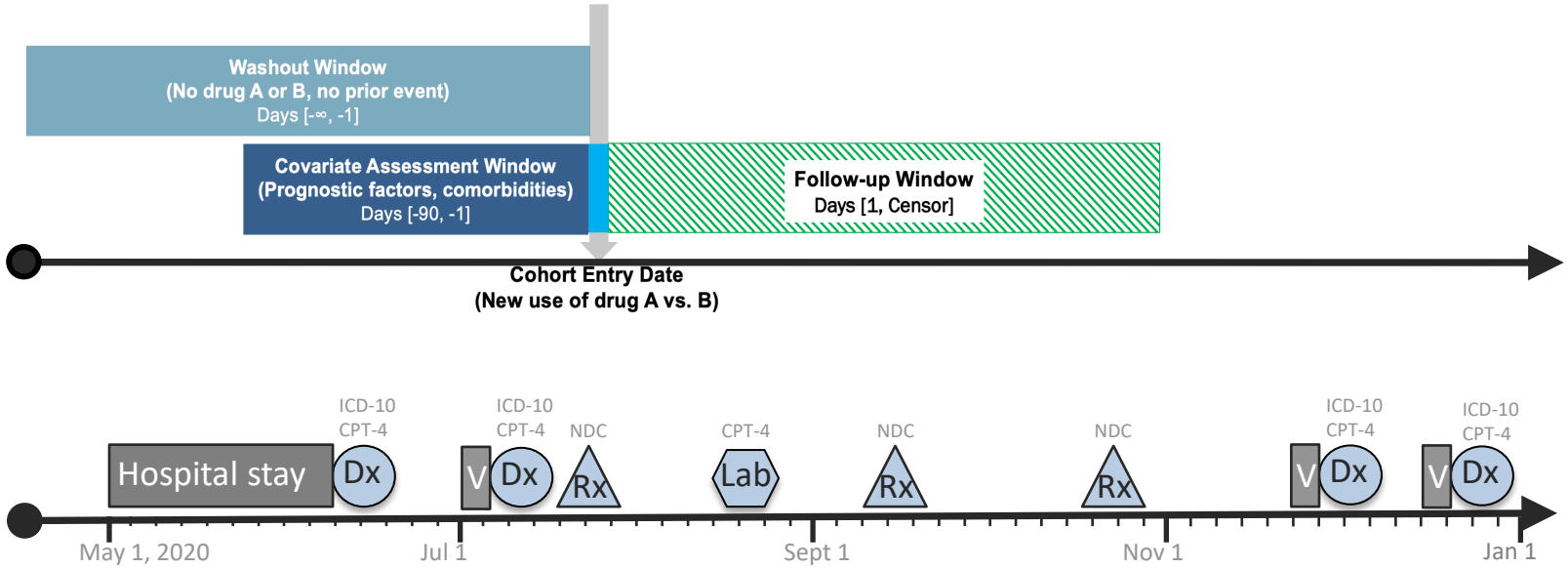
Data continuity

Ⓑ A **causal study design** is implemented in event time anchored at the cohort entry date

Data granularity

Ⓐ **Claims data** provide a longitudinal recording of all encounters with the professional healthcare system

Data chronology



Note: This figure focuses on elements relevant for a discussion of inferential studies embedded in EHR+claims data. It purposefully disregards many informatics aspects that are required but would distract from this discussion.

Data continuity ✓

Data granularity ✓

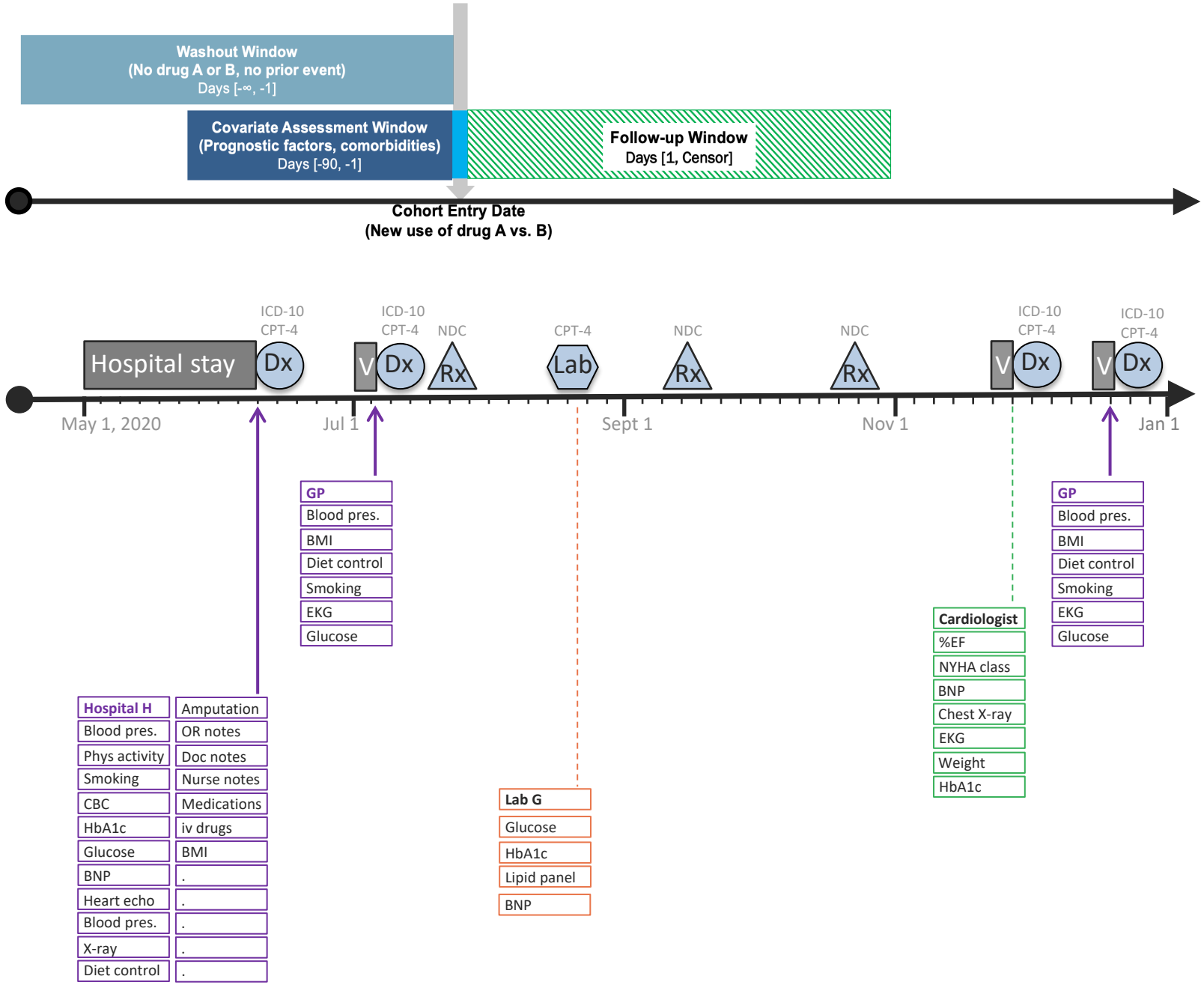
Data chronology ✓

Ⓑ A **causal study design** is implemented in event time anchored at the cohort entry date

Ⓐ **Claims data** provide a longitudinal claims recording of all encounters with the professional healthcare system

Ⓒ **EHR data within system** from Hospital and General Practitioner (GP)

Ⓓ **EHR data from outside** system Laboratory G and Cardiologist remain unobservable to the investigator



Note: This figure focuses on elements relevant for a discussion of inferential studies embedded in EHR+claims data. It purposefully disregards many informatics aspects that are required but would distract from this discussion.

Data continuity ✓

Ⓑ A **causal study design** is implemented in event time anchored at the cohort entry date

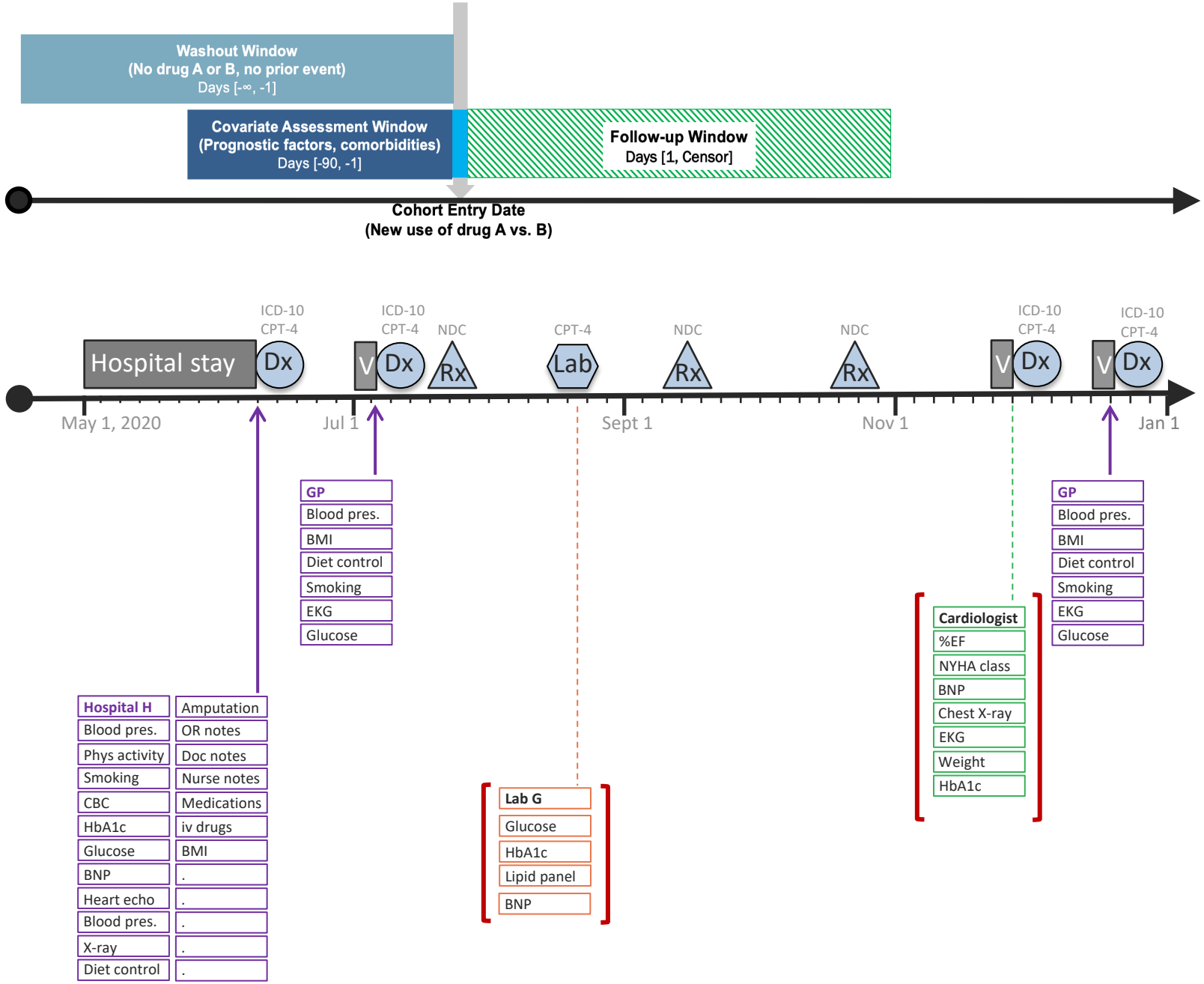
Data granularity ✓

Ⓐ **Claims data** provide a longitudinal claims recording of all encounters with the professional healthcare system

Data chronology ✓

Ⓒ **EHR data within system** from Hospital and General Practitioner (GP)

Ⓓ **EHR data from outside** system Laboratory G and Cardiologist remain unobservable to the investigator



Note: This figure focuses on elements relevant for a discussion of inferential studies embedded in EHR+claims data. It purposefully disregards many informatics aspects that are required but would distract from this discussion.

Real-World Evidence Data Enterprise (RWE DE)- An Overview

Commercial Network (21 million lives)

healthverity

TriNetX

EHR+claims linked lives in SCDM for routine querying

Development Network

Mass General Brigham

Duke Clinical Research Institute

VANDERBILT UNIVERSITY MEDICAL CENTER

KAISER PERMANENTE

Kaiser Permanente Washington Health Research Institute

EHR+claims linked lives in SCDM + investigator access to free-text EHRs through a standardized storage process across sites for methods and tool development

- ✓ Both networks operational
- ✓ Several demonstration projects ongoing

Broadening the Reach of Sentinel Inferential Queries with RWE-DE

A. Primary analysis in the RWE-DE; analysis is not feasible using claims data from the SDD

Commercial Network
(21 million lives)

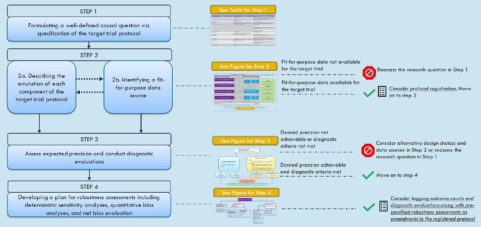
healthverity · TriNetX

EHR+claims linked lives in SCDM for routine querying

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

Central analytic hub

Inferential query execution following PRINCIPLED framework



B. Primary analysis in the SDD; supporting analyses in RWE-DE

Commercial Network
(21 million lives)

healthverity · TriNetX

EHR+claims linked lives in SCDM for routine querying

Development Network
(4.5 million lives)

Mass General Brigham · Duke Clinical Research Institute · Vanderbilt University Medical Center · Kaiser Permanente (Kaiser Permanente Washington Health Research Institute)

EHR+claims linked lives in SCDM + investigator access to free-text EHRs through a standardized storage process across sites for methods and tool development

Use cases where direct access to free-text notes is needed

Use cases relying on structured EHR data

Central analytic hub

Rapid balance evaluation for confounder unmeasured in claims data but available in structured EHRs

Statistical adjustment for unmeasured confounders that are measured for a subset in structured EHR data using calibration approaches

Co-ordinated local analyses

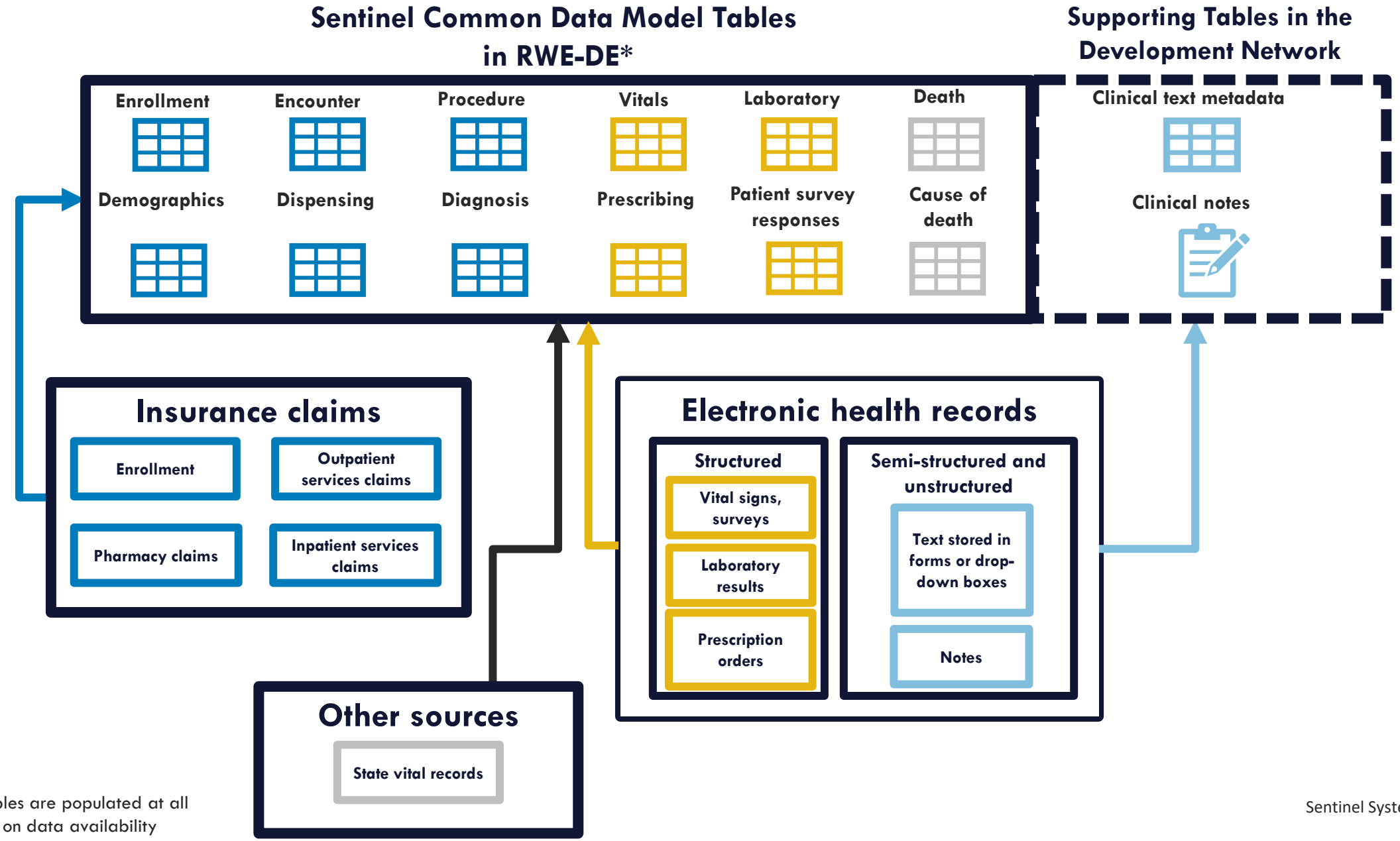
Rapid balance evaluation for confounder unmeasured in claims data but available in structured or unstructured EHRs

Statistical adjustment for unmeasured confounders that are measured for a subset in structured or unstructured EHR data using calibration approaches

Expedited endpoint validation using NLP assistance

Phenotyping algorithm development

Data Sources and Availability in RWE DE



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

Methodological Initiatives in Sentinel

(Rishi Desai)

Causal Inference Requirements

Design Layer

Achieve causal study design

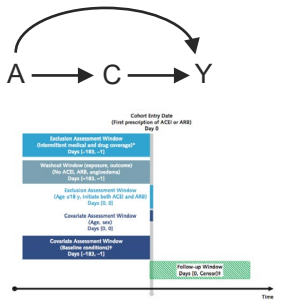
Considering:
- Study question
- Exposure variation
- Measurement quality

DESIGN CHOICE

1) Controlled 2) self-controlled 3) scanning
- Medically-informed target pop^{n}
- Patient-informed outcomes
- Biologically-informed effect window

BIAS REDUCTION

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



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

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

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

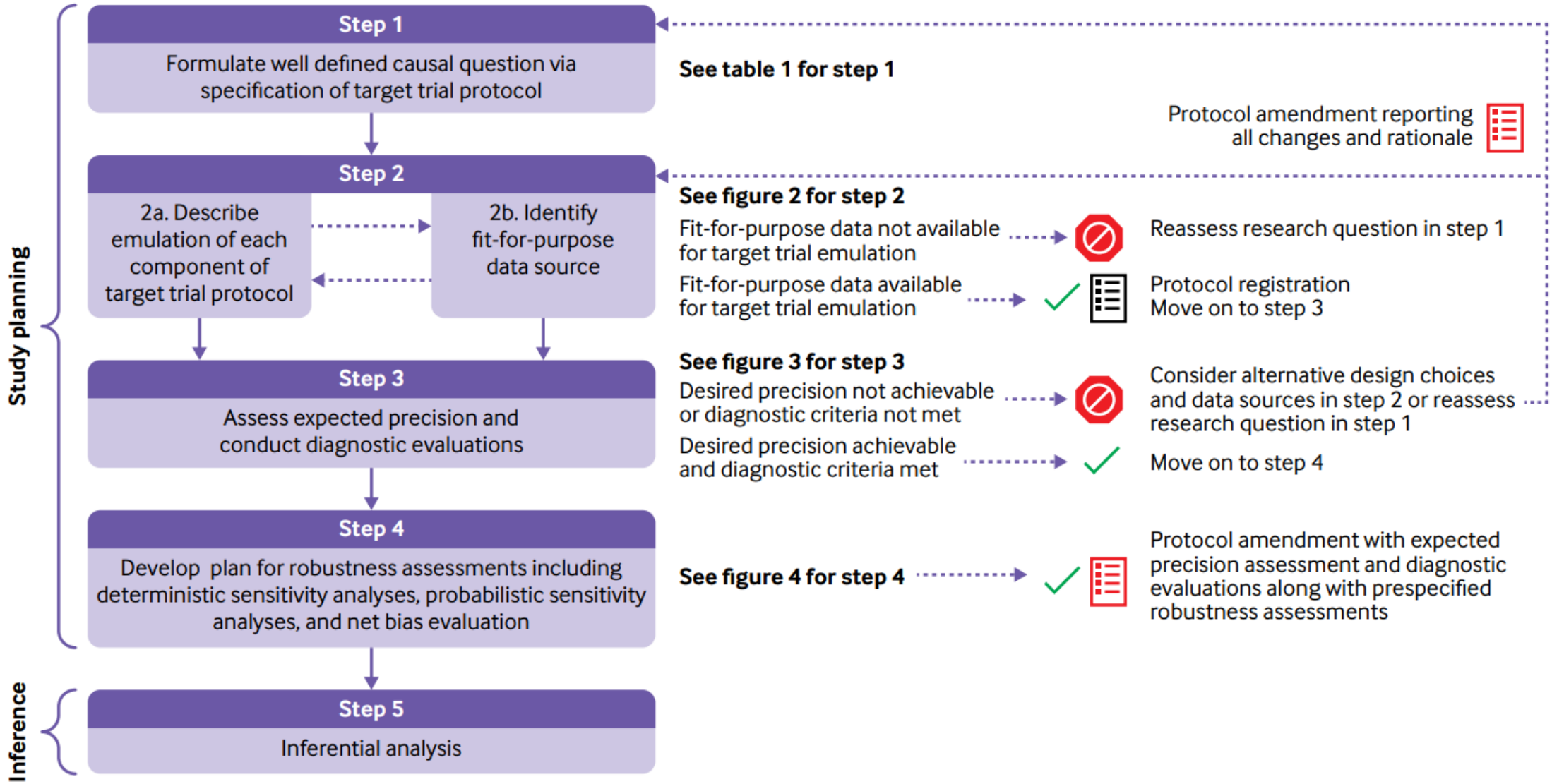


Fig 1 | Overview of the process guide for inferential studies using healthcare data from routine clinical practice

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: Computable phenotyping to identify health conditions of interest incompletely captured with Dx, Px, or Rx codes

Analytics Layer

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Article Contents

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

A general framework for developing computable clinical phenotype algorithms

David S Carrell, PhD, James S Floyd, MD, MS, Susan Gruber, PhD, Brian L Hazlehurst, PhD, Patrick J Heagerty, PhD, Jennifer L Nelson, PhD, Brian D Williamson, PhD, Robert Ball, MD, MPH, ScM Author Notes

Journal of the American Medical Informatics Association, ocae121, https://doi.org/10.1093/jamia/ocae121

Published: 15 May 2024 **Article history ▾**

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)

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).

Initially submitted August 11, 2021; accepted for publication October 11, 2022.

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

Follow this preprint

Scalable Incident Detection via Natural Language Processing and Probabilistic Language Models

Colin G. Walsh, Drew Wilimitis, Qingxia Chen, Aileen Wright, Jhansi Kolli, Katelyn Robinson, Michael A. Ripperger, Kevin B. Johnson, David Carrell, Rishi J. Desai, Andrew Mosholder, Sai Dharmarajan, Sruthi Adimadhyam, Daniel Fabbri, Danijela Stojanovic, Michael E. Matheny, Cosmin A. Bejan

doi: https://doi.org/10.1101/2023.11.30.23299249

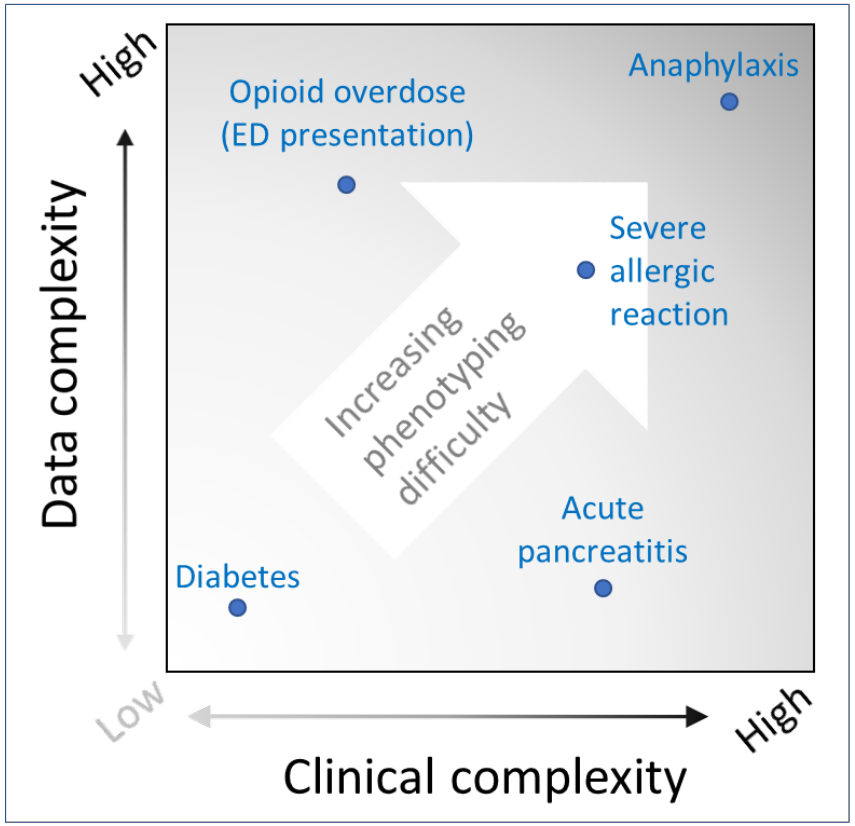
Computable Phenotyping

What do we mean by computable phenotyping?

- An attempt to accurately identify a health condition of interest from healthcare data using combination of various sources of information eg diagnosis codes, procedures, medications, symptoms in physician notes (aka “features”)
- For many conditions, complex algorithms are needed to integrate various sources of information to assign probabilities of having the condition of interest in a patient given her profile
- When these algorithms are created, we typically need to validate our predictions against some “gold-standard” truth to determine the best approach

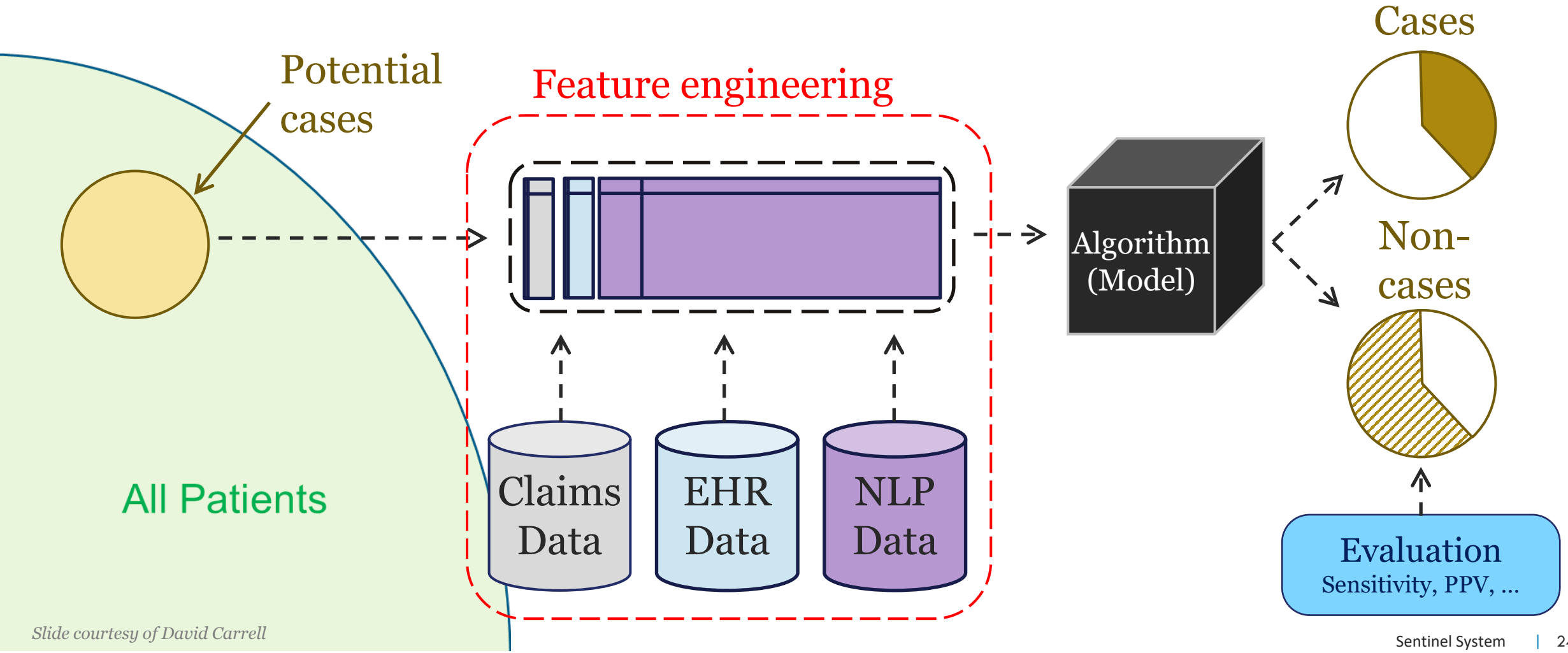
Computable Phenotyping: General Framework

- 5 stages of model development
  - Fitness-for-purpose assessment
  - Creating gold standard data
  - Feature engineering
  - Model development
  - Model Evaluation and reporting
- Avoid unnecessary complexity
- Leverage automation when feasible
- Design for transportability/reusability



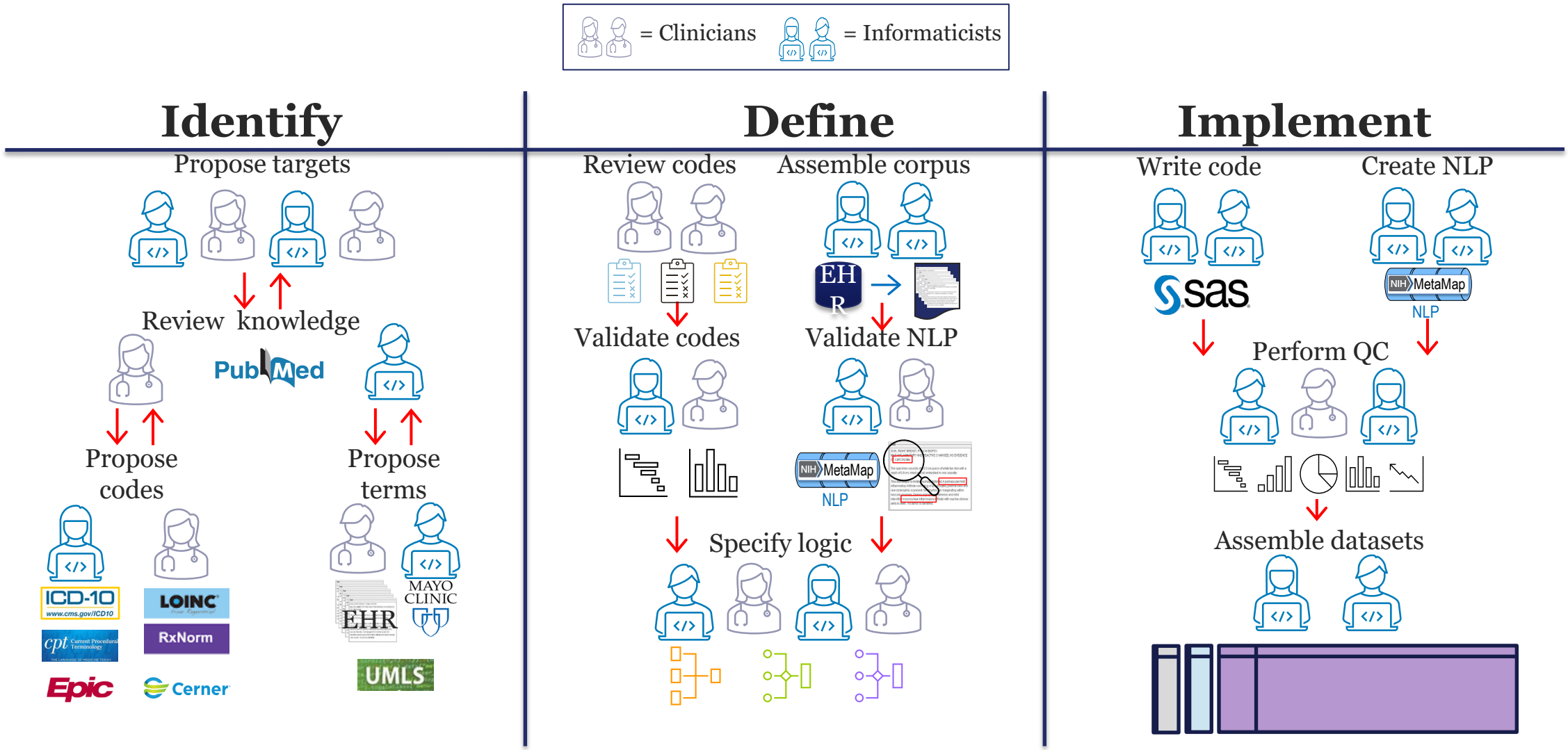
Computable Phenotype: *Development Process*

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



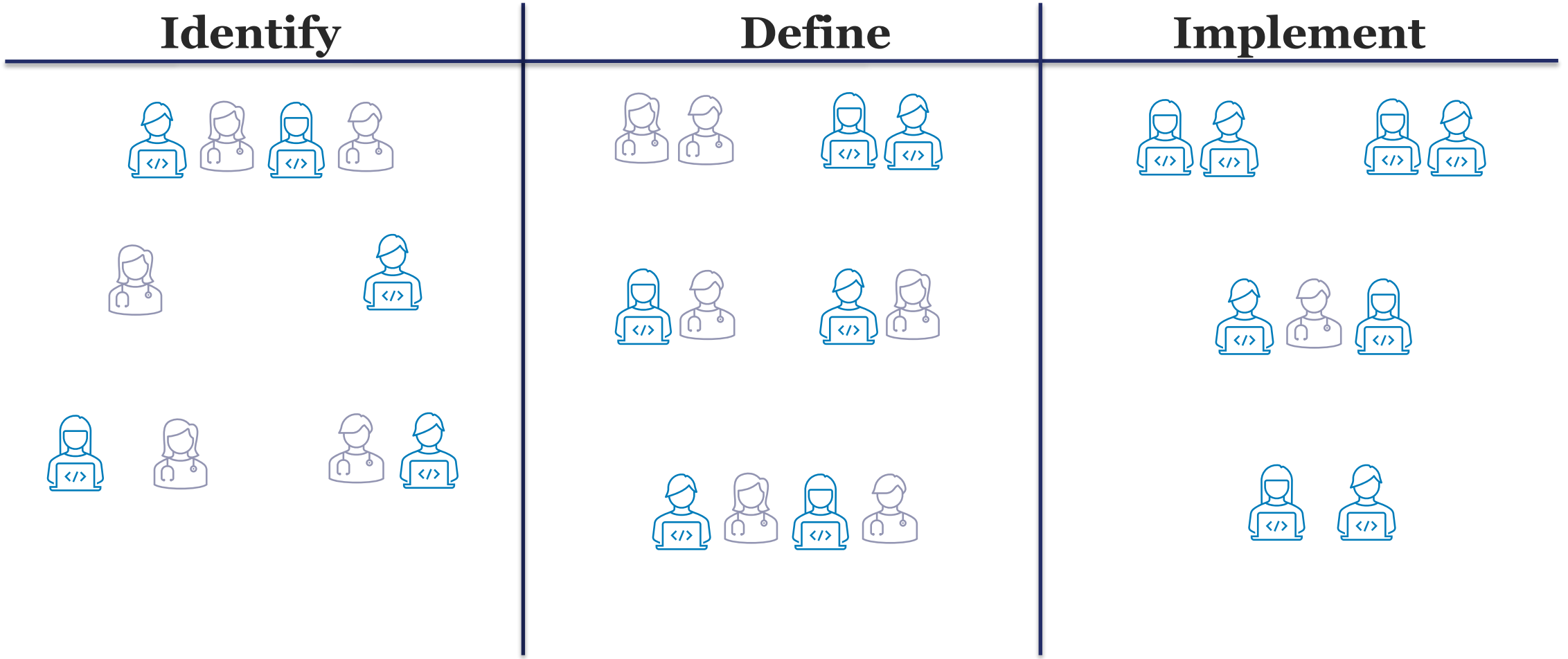
Slide courtesy of David Carrell

Manual Feature Engineering



Slide courtesy of David Carrell

Manual Feature Engineering



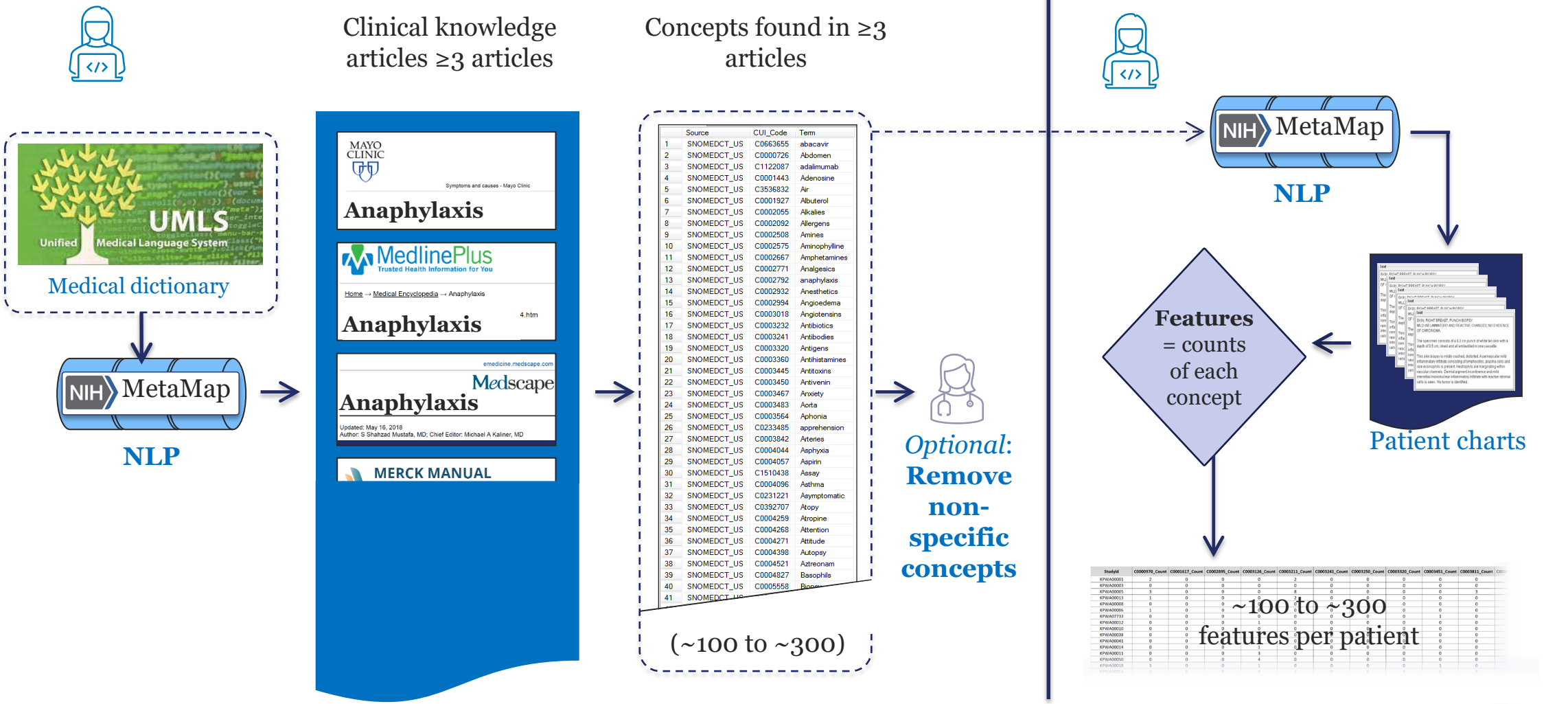
Slide courtesy of David Carrell

Feature Engineering: *Automated*

= Clinicians = Informaticists

Identify & Define*

Implement



* Yu et al. JAMIA 2015

Slide courtesy of David Carrell

Feature Engineering: *Automated*

= Clinicians = Informaticists

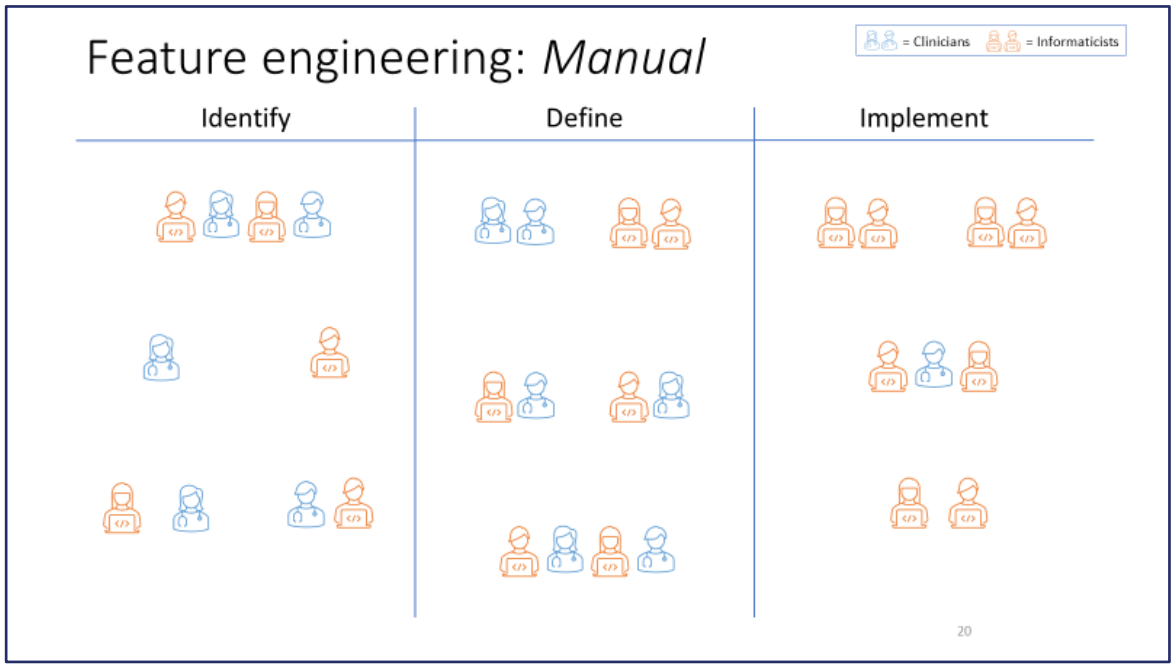
Identify & Define

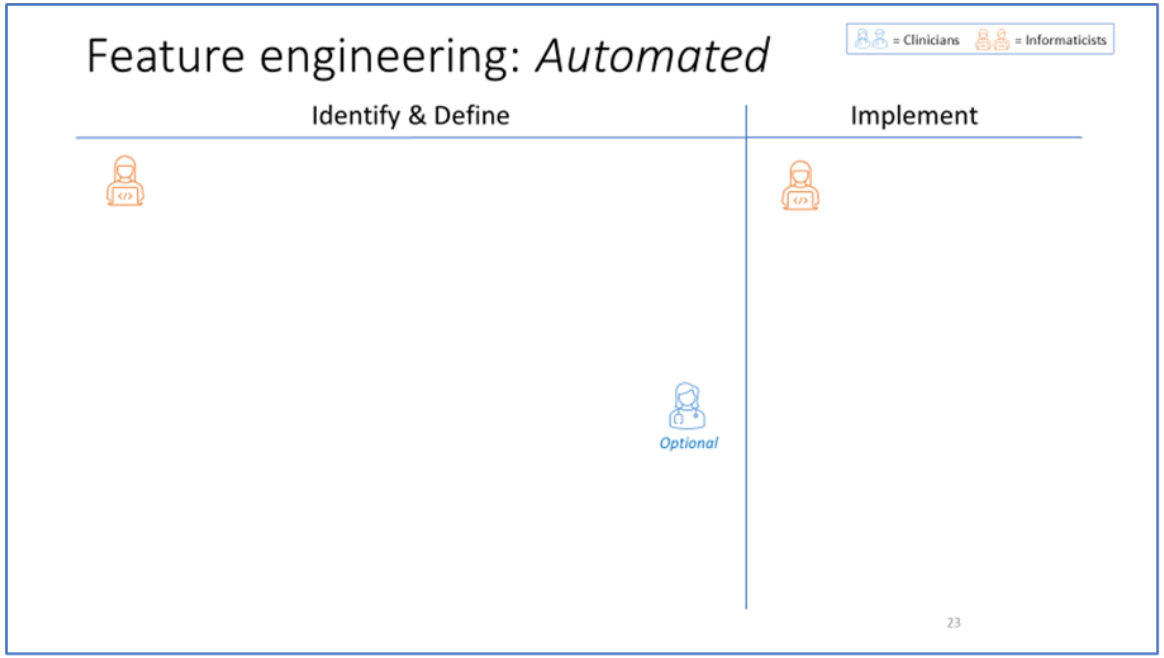
Optional

Implement

Slide courtesy of David Carrell

Feature Engineering: Manual vs. Automated





Automation advantages:

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

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

Slide courtesy of David Carrell

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: 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.

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?	*Summarize and visualize missingness:*	*Identify patterns visually*:*
smdi_check_covar()	*smdi_summarize()*	*gg_miss_upset()*
smdi_na_indicator()	*smdi_vis()*	*md_pattern()*

Inferential Three Group Diagnostics

Group 1 Diagnostics	**Group 2 Diagnostics**	**Group 3 Diagnostics**	**Group 1-3 Diagnostics**
smdi_amsd()	*smdi_rf()*	*smdi_outcome()*	*smdi_diagnose()*
smdi_hotelling()			*smdi_style_gt()*
smdi_little()			

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

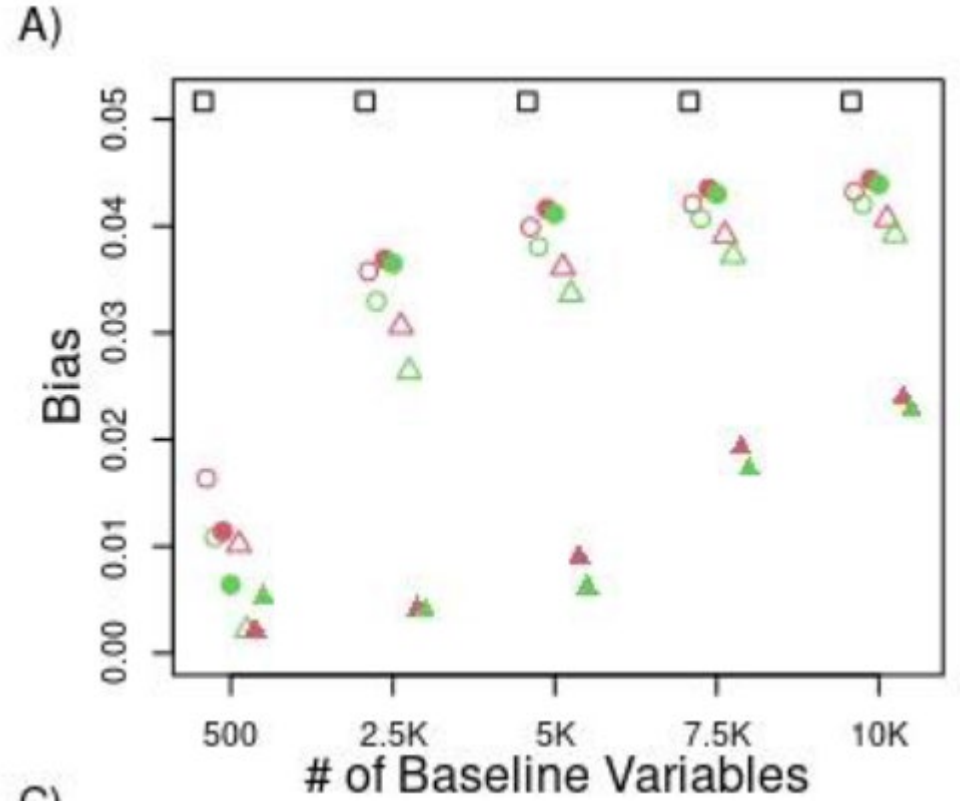
Use of cross-fitting to manage overfitting

- 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



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

- 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

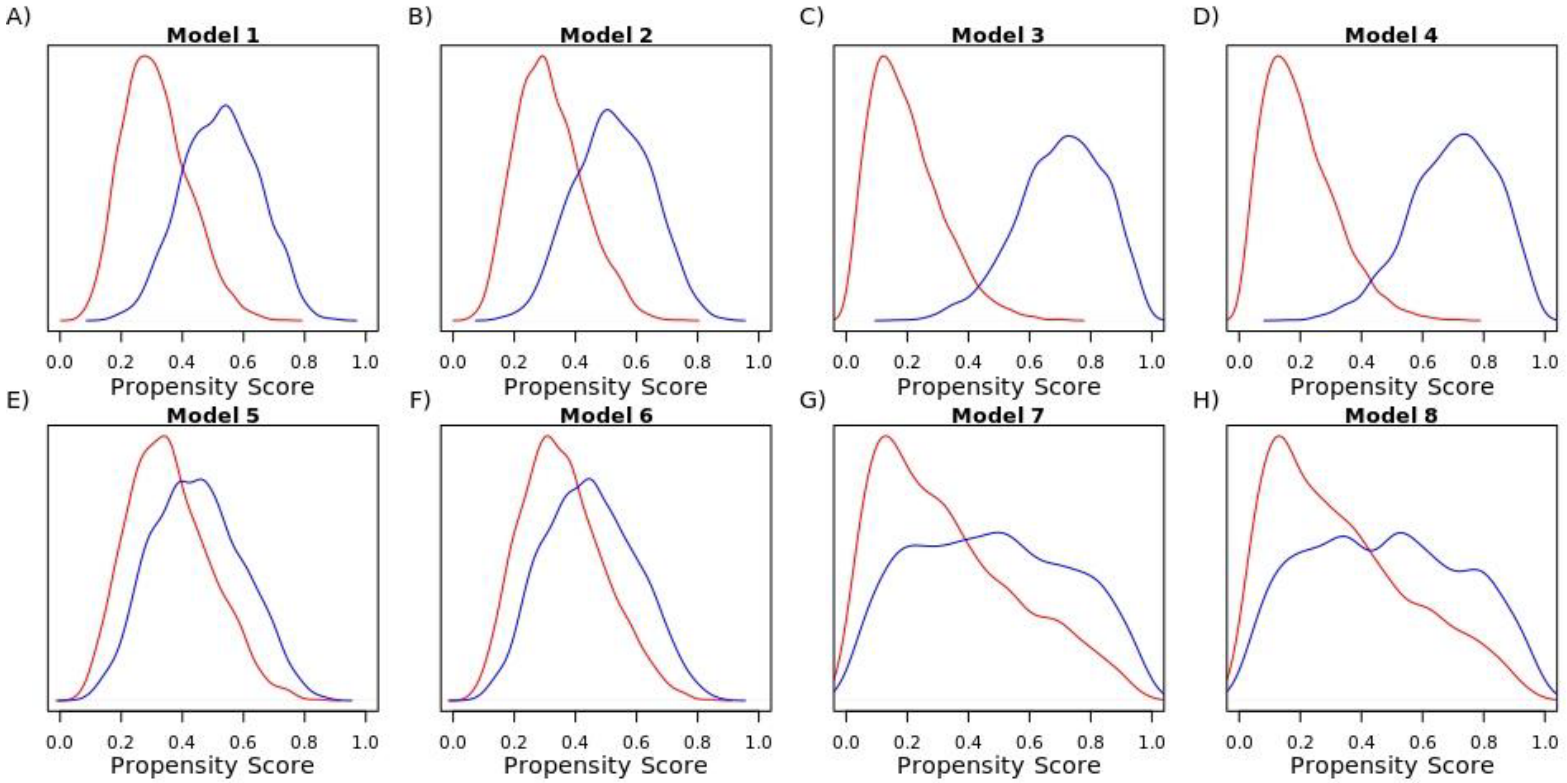
Cross-fit of trt model

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

No cross-fit of trt model

- Model 5: Cross-Fit (CF) Lasso
- Model 6: CF OAL
- Model 7: Collaborative Controlled CF Lasso
- Model 8: Collaborative Controlled CF OAL

Cross-fit of trt model

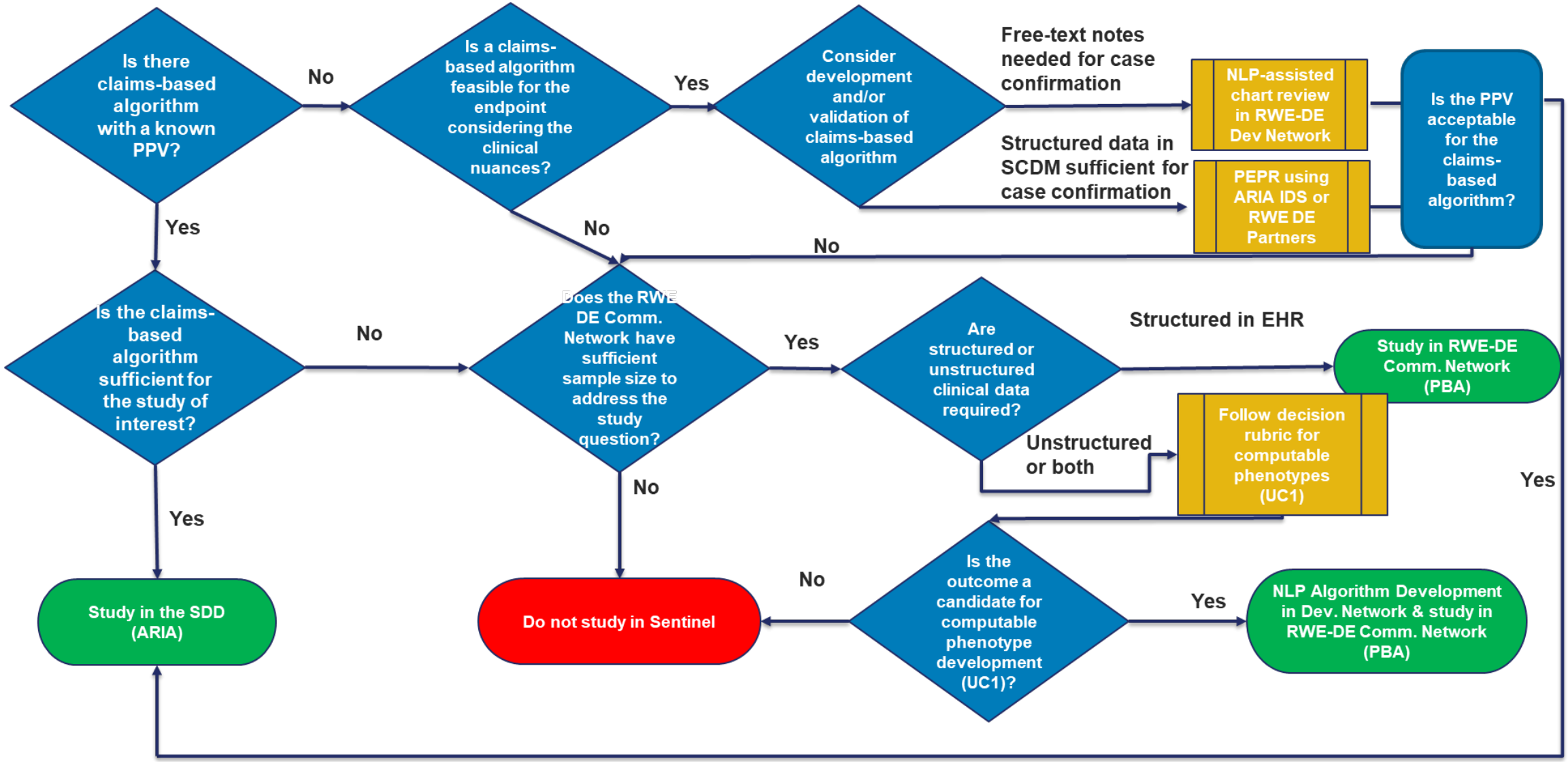
What (likely) explains robust performance:

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

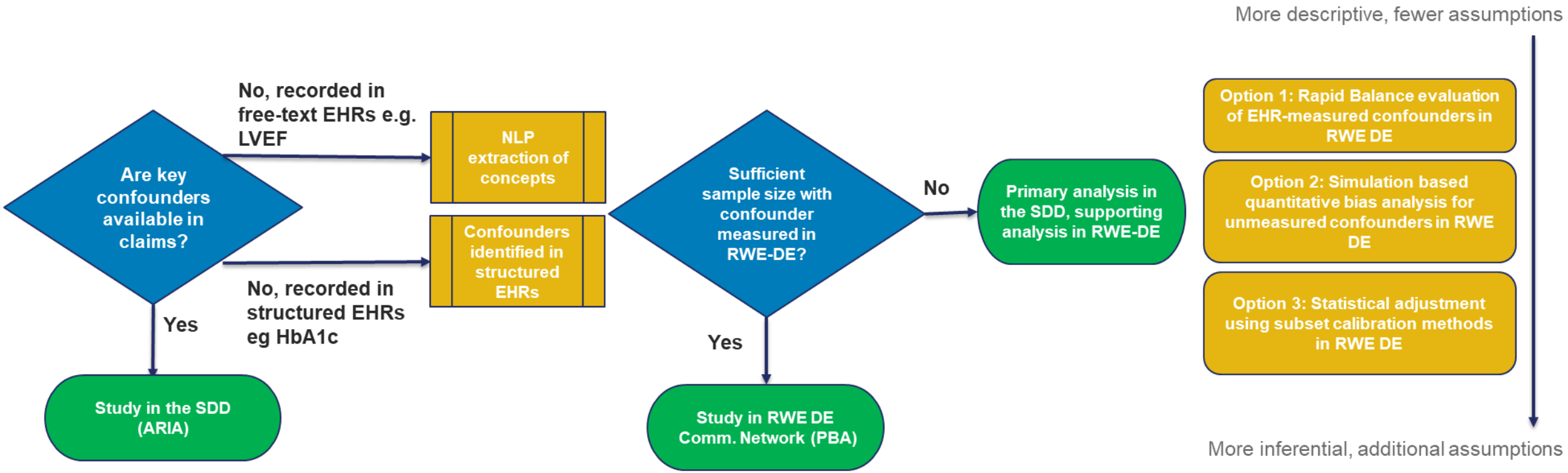
Wyss et al. AJE 2024

Decision Guides to Integrate Methodologic Advances with Practice

Draft Decision Guide for Evaluating Data Fitness for Purpose in Sentinel: Focus on Outcomes*



Draft Decision Guide for Evaluating Data Fitness for Purpose in Sentinel: Focus on Confounders



Summary

- Large scale data infrastructure where EHRs are linked to claims data will offer visibility into additional clinical information that is not available in claims data alone
- Methodological innovations will allow investigators to readily leverage the infrastructure as needed
- All these activities ultimately 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

Thank You