



Innovation Day

April 12, 2023

Sentinel Innovation Center

Sentinel Innovation Center



Brigham and Women's Hospital
Founding Member, Mass General Brigham



Duke University
School of Medicine

VANDERBILT  UNIVERSITY
MEDICAL CENTER



KAISER PERMANENTE®

Panelists



Session 1

A General Framework for Developing Computable Phenotyping Algorithms from Electronic Health Records

David S. Carrell, PhD

Associate Investigator,
Kaiser Permanente Washington Health
Research Institute (KPWHRI)



Session 2

Harmonizing Electronic Health Record and Claims Data Across FDA Sentinel Initiative Data Partners: Case Study and Lessons Learned

Xu Shi, PhD

Assistant Professor,
Department of Biostatistics at the University
of Michigan



Session 3

A PProcess guide for INferential studies using healthcare data from routine ClinIcal Practice to EvaLuate causal Effects of Drugs (PRINCIPLED)

Rishi J Desai, MS, PhD

Assistant Professor
Division of Pharmacoepidemiology &
Pharmacoeconomics, Brigham and Women's Hospital,
Harvard Medical School



Session 4

Approaches to Handling Partially Observed Confounder Data from Electronic Health Records (EHR) in Non-randomized Studies of Medication Outcomes

Janick Weberpals, RPh, PhD

Instructor
Division of Pharmacoepidemiology &
Pharmacoeconomics, Brigham and Women's Hospital,
Harvard Medical School

Session Logistics



Questions



Presentation availability



A General Framework for Developing Computable Phenotyping Algorithms from Electronic Health Records

David S. Carrell, PhD

Kaiser Permanente Washington Health Research Institute

on behalf of the

Sentinel Advanced Phenotyping Framework Team

and the

Scalable Natural Language Processing (NLP) Team

Outline

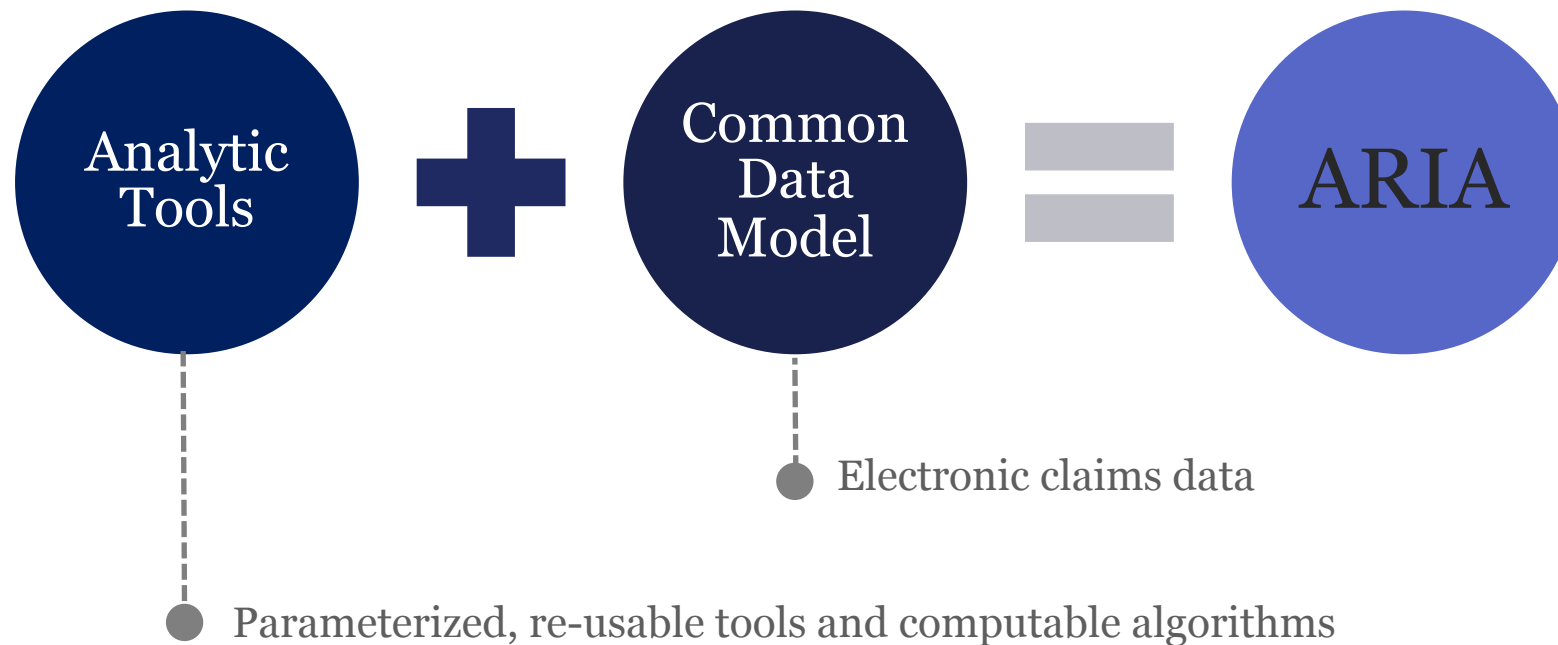
- Motivation
- A general framework for scalable development
 - Assessing fitness for purpose
 - Creating gold standard data
 - Feature engineering
 - Model development
 - Model evaluation and reporting



Motivation

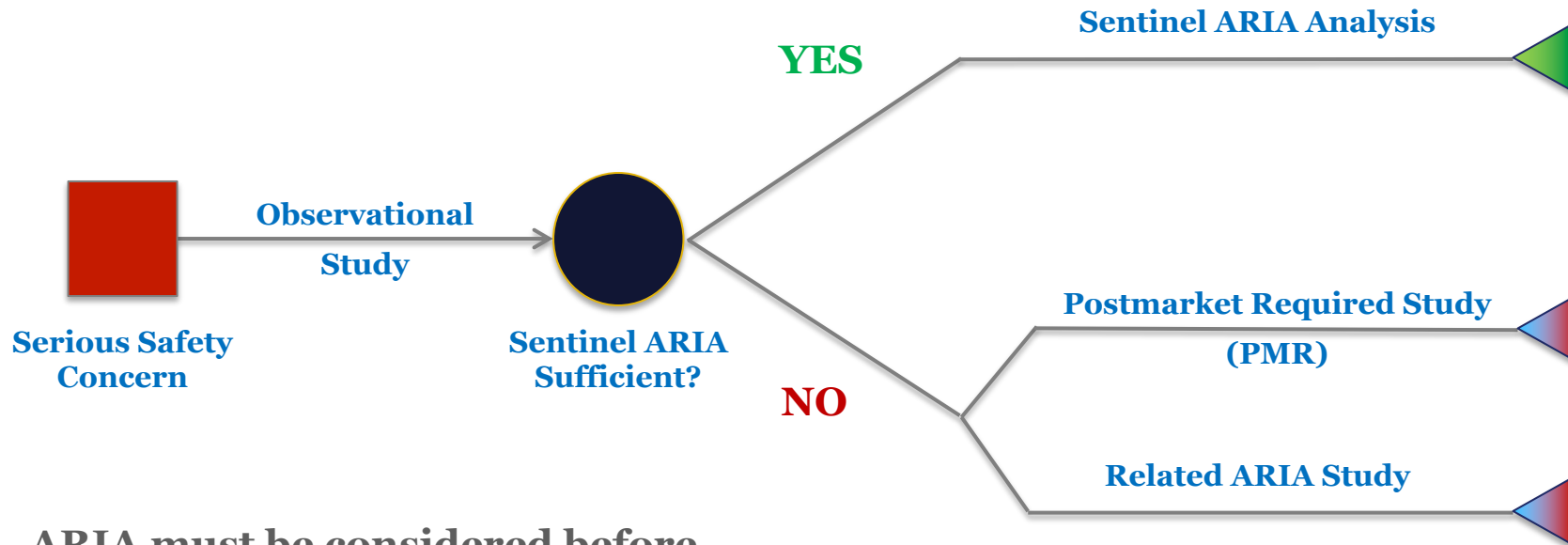
Motivation

- Goal: improve safety surveillance using observational data
- Active Risk Identification and Analysis (ARIA) system:



Motivation

When is the ARIA Process Needed?



ARIA must be considered before a sponsor PMR can be issued

Motivation

- ARIA sufficiency is achieved when:
 - Study population is available in the data
 - Outcome & exposure of interest, covariates can be identified from data
 - Methods/analytic tools can assess exposure-related risk *with satisfactory precision*
- 2016—2021: ARIA **insufficient** for **60% of safety concerns**
 - Availability of outcome data is a primary reason for insufficiency
 - 67% of all ARIA insufficiencies were insufficient (in part*) due to lack of outcome data

Example ARIA **sufficient**** outcomes:

- GI bleeding
- Heart failure
- Lymphoma
- Major adverse cardiac events (MACE)
- Myocardial infarction
- Multiple sclerosis relapse
- Non-melanoma skin cancer
- Seizure
- Stroke

Example ARIA **insufficient*** outcomes:

- Acute pancreatitis
- Anaphylaxis
- Drug-induced liver injury
- Fatal MACE
- Malignancies (several)
- Nerve injury
- Suicide or suicidal ideation

*Reasons for insufficiency are not mutually exclusive

**Sufficiency is highly dependent on the scientific question and regulatory context

Motivation

- Our focus: Improving ARIA sufficiency by improving methods of *outcome identification* (phenotyping)
- Key considerations:
 - Assessing “fitness for purpose” of a phenotyping effort
 - Gold-standard data creation
 - Feature engineering
 - Model development
 - Model evaluation and reporting
- Challenge: Traditional approaches to phenotyping are expensive and time-consuming
- Approach: A general framework is needed to guide *scalable development* of phenotype algorithms
- Case studies: Anaphylaxis, acute pancreatitis, COVID-19 disease



Assessing Fitness for Purpose

Assessing Fitness for Purpose: Key Points

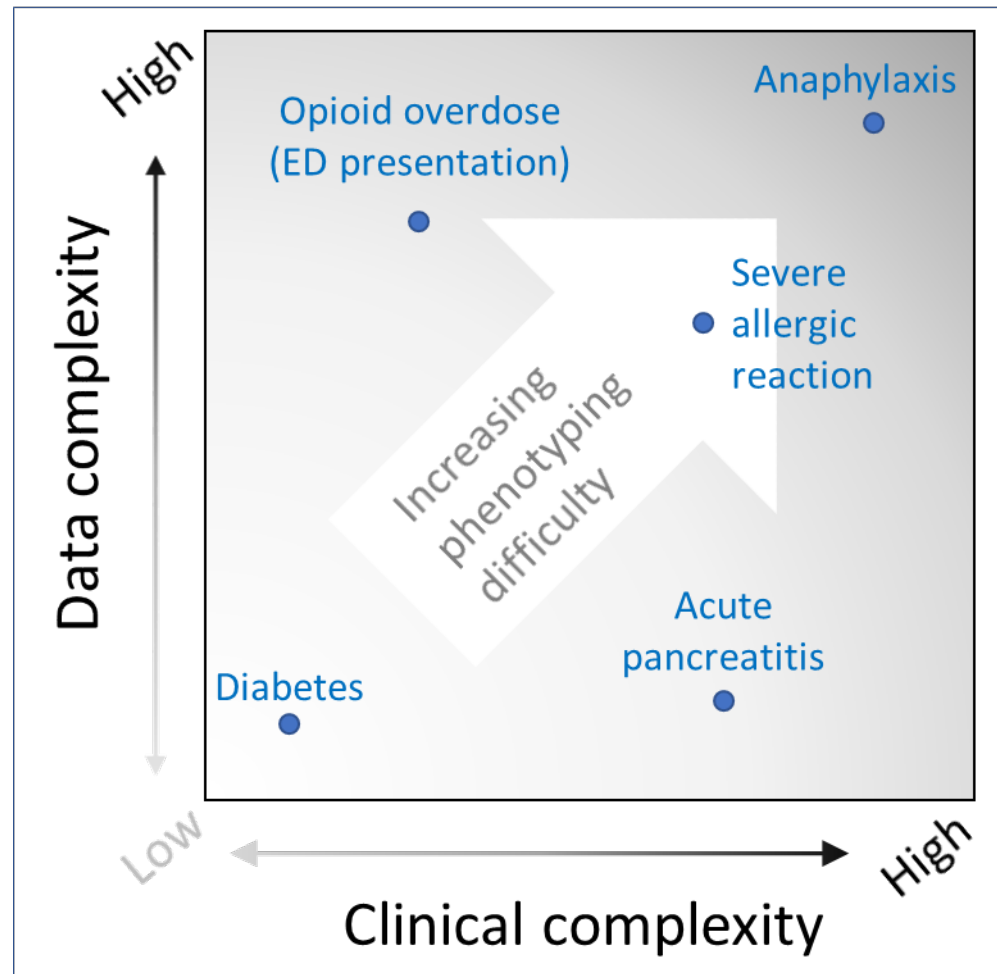
- Assessing fitness for purpose” in the process of determining whether a particular phenotyping effort has a reasonably likely chance of success before the phenotype development work begins.
 - A critical first step intended to identify—and avoid expending scarce resources on—phenotyping efforts highly likely to be unsuccessful (i.e., developing a phenotype model whose performance is insufficient for ARIA purposes)
 - A type of feasibility assessment (or “premortem”) with specific guidance as to how it should be done
- Based on the combined expert judgement of members of the *development team* (*clinicians, statisticians, informaticists, chart reviewers, EHR data experts*)
- Considers each stage of development
 - Creating gold standard data
 - Feature engineering
 - Model development
 - Model evaluation & reporting

} Impact of *clinical complexity*?
Impact of *data complexity*?

Assessing Fitness for Purpose: Complexities

Data complexity sources

- Data heterogeneity
- Data obscurity
- Data imprecision
- Data irregularity
- Data instability
- High dimensionality
- Lack of structure



Clinical complexity sources

- Competing diagnoses
- Lack of definitive diagnostic tests
- Lack of consensus about diagnostic criteria
- Limited knowledge, time, or technology

Figure 1. Relationship between clinical complexity, data complexity, and increasing phenotyping difficulty with illustrative phenotypes.

Assessing Fitness for Purpose: Key Points

- “Assessing fitness for purpose” in the process of determining whether a particular phenotyping effort has a reasonably likely chance of success before the phenotype development work begins.
 - A critical first step intended to identify—and avoid expending scarce resources on—phenotyping efforts highly likely to be unsuccessful (i.e., developing a phenotype model whose performance is insufficient for ARIA purposes)
 - A type of feasibility assessment (or “premortem”) with specific guidance as to how it should be done
- Based on the combined expert judgement of members of the *development team* (*clinicians, statisticians, informaticists, chart reviewers, EHR data experts*)
- Considers each stage of development
 - Creating gold standard data
 - Feature engineering
 - Model development
 - Model evaluation & reporting

} Impact of *clinical complexity*?
} Impact of *data complexity*?
- ➔ • Yields a “Go/No-go” decision
 - “No-go” → Efficiency by avoiding wasted effort
 - “Go” → Higher likelihood of success; insights into challenges, opportunities
- Future work
 - Methods for estimating **amount of training data needed** for model development



Creating Gold Standard Data

Creating Gold Standard Data

- Goal: identify true cases and controls
 - Gold standard data is *always* for *evaluation*
 - ... and often also needed for model *training*
- Challenge: Creating gold standard data is very expensive, and this limits the quantity available for any given study
 - *Note:* Unlike gold standard data, silver standard data is inexpensive and abundant because it does not require manual chart review; instead, it is created by specifying simple rules to create measures that are believed to be highly correlated with actual (“gold”) outcomes
- Best practices:
 - Chart abstraction guidelines should reflect established clinical diagnostic criteria
 - Clinician oversight of (non-clinician) chart abstractors can enhance efficiency
 - Dual independent review of a representative sample of charts is important for assessing replicability when some chart are reviewed by only one person
 - Efficiency is also served by reusing existing abstraction tools (e.g., REDCap forms)
- Future work:
 - Efficiencies of NLP-assisted methods?
 - Incorporating silver standard surrogate outcomes into model training?
 - Can sampling strategies reduce -the quantity of gold standard data needed for evaluation by focusing reviews on those events that best reflect a model’s performance?
 - *Example:* Can sampling be guided by predicted probabilities of models trained on silver labels?

Creating Gold Standard Data

Model training
minority class
issues

Phenotype	Setting	Data type	Cases	Non-cases
Anaphylaxis	Kaiser Washington	Gold	154	85
Anaphylaxis	Kaiser NW	Gold	180	97
Acute pancreatitis	Kaiser NW	Gold	182	118
COVID-19	Vanderbilt	Silver	24,355	
		Gold	266	153
COVID-19	Kaiser Washington	Silver	8,329	
		Gold	269	168



Feature Engineering

Feature Engineering

- Goal: Measure things that help distinguish true cases from non-cases
- Challenge: Manual approaches are **time/expert-intensive, operator-dependent,**
 - **Wasted effort** if based on idiosyncratic local data, don't improve model performance

Feature Engineering: *Manual*

 = Clinicians  = Informaticists

Identify

Propose targets



Review knowledge



Propose codes

Propose terms



Define

Review code lists



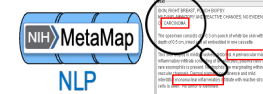
Validate code usage



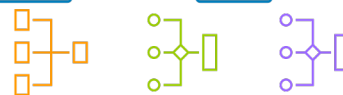
Assemble corpus



Validate NLP



Specify logic



Implement

Write code



Create NLP



Perform QC



Assemble datasets



StudyId	CO000975_Count	CO001817_Count	CO002895_Count	CO001126_Count	CO001211_Count	CO001241_Count	CO001210_Count	CO001812_Count	CO001811_Count	CO001811_Count
KPFA000001	2	0	0	0	0	2	0	0	0	0
KPFA000002	0	0	0	0	0	0	0	0	0	0
KPFA000003	3	3	0	0	0	8	8	0	0	0
KPFA000004	1	0	0	0	2	0	0	0	0	0
KPFA000005	0	0	0	0	0	0	0	0	0	0
KPFA000006	1	0	0	0	0	0	0	0	0	0
KPFA000007	0	0	0	0	0	0	0	0	0	0
KPFA000008	0	0	0	1	0	0	0	0	0	0
KPFA000009	0	0	0	0	0	0	0	0	0	0
KPFA000010	0	0	0	0	0	0	0	0	0	0
KPFA000011	0	0	0	0	0	0	0	0	0	0
KPFA000012	0	0	0	0	0	0	0	0	0	0
KPFA000013	0	0	0	0	0	0	0	0	0	0
KPFA000014	0	0	0	0	0	0	0	0	0	0
KPFA000015	0	0	0	1	0	0	0	0	0	0
KPFA000016	0	0	0	3	0	0	0	0	0	0
KPFA000017	0	0	0	4	0	0	0	0	0	0
KPFA000018	3	0	0	1	0	0	0	0	1	0
KPFA000019	1	0	0	0	0	0	0	0	0	0

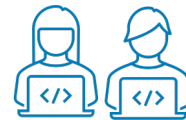
Feature Engineering: *Manual*

  = Clinicians   = Informaticists

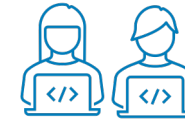
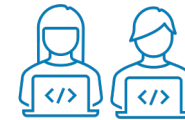
Identify



Define



Implement



Feature Engineering: Automated

AFEP

Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources

Sheng Yu^{1,2,3,*}, Katherine P Liao^{2,3}, Stanley Y Shaw⁴, Vivian S Gainer⁵, Susanne E Churchill⁵, Peter Szolovits⁶, Shawn N Murphy^{4,5}, Isaac S Kohane^{3,7}, Tianxi Cai⁸

ABSTRACT

Objective Analysis of narrative (text) data from electronic health records (EHRs) can improve population-scale phenotypic research. Currently, selection of text features for phenotyping algorithms is slow and laborious, requiring extensive manual expert review. This paper introduces a method to develop phenotyping algorithms in an unbiased manner by automatically selecting informative features, which can be comparable to expert-curated ones in classification accuracy.

Materials and methods Comprehensive medical concepts were collected from publicly available knowledge sources. Natural language processing (NLP) revealed the occurrence patterns of these concepts in EHR narrative notes. Informative features for phenotype classification. When combined with additional codified features, a penalized logit model trained to classify the target phenotype.

Results The authors applied our method to develop algorithms to identify patients with rheumatoid arthritis and CAD among those with rheumatoid arthritis from a large multi-institutional EHR. The area under the receiver operating characteristic curve for classifying RA and CAD using models trained with automated features were 0.951 and 0.929, respectively, compared to 0.929 by models trained with expert-curated features.

Discussion Models trained with NLP text features selected through an unbiased, automated procedure achieved comparable accuracy to those trained with expert-curated features. The majority of the selected model features were interpretable.

Conclusion The proposed automated feature extraction method, generating highly accurate phenotyping algorithms is a significant step toward high-throughput phenotyping.

INTRODUCTION

Electronic health record (EHR) adoption has increased dramatically in recent years. By 2013, 59% of private acute care hospitals in the United States had adopted an EHR system, up from 9% in 2008.¹ Secondary use of EHR data has emerged as a powerful approach for a variety of biomedical research, including comparative effectiveness and stratifying patients for risk of comorbidities or adverse outcomes.^{2–10} More recently, the linking of genotype and biomarker data to EHR data has facilitated translational studies, such as genetic association studies.^{11–17} Compared to conventionally assembled epidemiologic and genomic cohorts that require individual patient recruitment, EHR-based studies can provide large sample sizes at a lower cost and shorter time frames. Furthermore, results from EHR-based genetic as-

narrative notes such as physician notes, or pathologic studies, or hospital discharge summaries provide a rich source of complementary information. Natural language processing (NLP) can efficiently extract occurrences of terms of clinical concepts and also used as features for algorithmic phenotyping algorithms that use both codified and accuracy relative to algorithms using codified features (e.g., ICD-9 billing codes).^{19–22}

Today, algorithms that identify a disease or condition are often constructed in two rather different ways. The first is to rely on human expertise to suggest a logic or model that identifies features that must be present

RECEIVED 24 October 2014
REVISED 25 February 2015
ACCEPTED 24 March 2015
PUBLISHED ONLINE FIRST 30 April 2015



SAFE

Journal of the American Medical Informatics Association, 24(e1), 2017, e143–e149

doi: 10.1093/jamia/ocw135

Advance Access Publication Date: 15 September 2016

Research and Applications



Research and Applications

Surrogate-assisted feature extraction for high-throughput phenotyping

Sheng Yu,^{1,2} Abhishek Chakraborty,³ Katherine P Liao,⁴ Tianrun Cai,⁵ Ashwin N Ananthakrishnan,⁶ Vivian S Gainer,⁷ Susanne E Churchill,⁸ Peter Szolovits,⁹ Shawn N Murphy,^{7,10} Isaac S Kohane,⁸ and Tianxi Cai⁵

¹Center for Statistical Science, Tsinghua University, Beijing, China, ²Department of Industrial Engineering, Tsinghua University, Beijing, China, ³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA, ⁴Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, USA, ⁵Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, USA, ⁶Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, USA, ⁷Research IS and Computing, Partners HealthCare, Charlestown, Massachusetts, USA, ⁸Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA, ⁹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, and ¹⁰Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts, USA

Corresponding Author: Sheng Yu, Center for Statistical Science, Tsinghua University, Beijing, China. Email: shengyu@sem.tsinghua.edu.cn

- Principles:
1. **Clinical text** is the primary data source
 2. **Published knowledge** provides expertise
 3. **Data-driven engineering** methods

PheNorm

Journal of the American Medical Informatics Association, 25(1), 2018, 54–60

doi: 10.1093/jamia/ocx111

Advance Access Publication Date: 3 November 2017

Research and Applications



Research and Applications

Enabling phenotypic big data with PheNorm

Sheng Yu,^{1,2} Yumeng Ma,³ Jessica Gronsbell,⁴ Tianrun Cai,⁵ Ashwin N Ananthakrishnan,⁶ Vivian S Gainer,⁷ Susanne E Churchill,⁸ Peter Szolovits,⁹ Shawn N Murphy,^{7,10} Isaac S Kohane,⁸ Katherine P Liao,¹¹ and Tianxi Cai⁴

¹Center for Statistical Science, Tsinghua University, Beijing, China, ²Department of Industrial Engineering, Tsinghua University, Beijing, China, ³Department of Mathematical Sciences, Tsinghua University, Beijing, China, ⁴Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, ⁵Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA, ⁶Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA, ⁷Research Information Science and Computing, Partners HealthCare, Charlestown, MA, USA, ⁸Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, ⁹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, ¹⁰Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA, and ¹¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Feature Engineering: Automated

 = Clinicians  = Informaticists

Identify & Define*

Implement



Clinical knowledge articles ≥ 3 articles

Concepts found in ≥ 3 articles



NLP

Stack of clinical knowledge articles for "Anaphylaxis":

- MAYO CLINIC: Symptoms and causes - Mayo Clinic
- MedlinePlus: Trusted Health Information for You
- emedicine.medscape.com: Medscape
- MERCK MANUAL Professional Version: The trusted provider of medical information since 1899
- WIKIPEDIA: The Free Encyclopedia

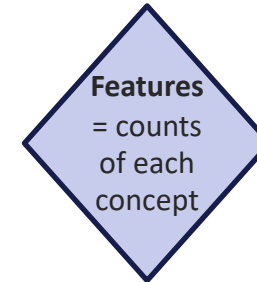
Source	CUI_Code	Term	
1	SNOMEDCT_US	C0663655	abacavir
2	SNOMEDCT_US	C0000726	Abdomen
3	SNOMEDCT_US	C1122087	adalimumab
4	SNOMEDCT_US	C0001443	Adenosine
5	SNOMEDCT_US	C3536832	Air
6	SNOMEDCT_US	C0001927	Albuterol
7	SNOMEDCT_US	C0002055	Alkalies
8	SNOMEDCT_US	C0002092	Allergens
9	SNOMEDCT_US	C0002508	Amines
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	Artifistamines
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	Asayy
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	Basophilia
40	SNOMEDCT_US	C0005558	Binge
41	SNOMEDCT_US		

(~ 100 to ~300)

Optional:
Remove
non-specific
concepts



NLP



Patient charts

StudyId	C000070_Count	C0001617_Count	C0002895_Count	C0003126_Count	C0003211_Count	C0003241_Count	C0003250_Count	C0003320_Count	C0003451_Count	C0003811_Count	C0003812_Count
KPFA00001	2	0	0	0	2	0	0	0	0	0	0
KPFA00003	0	0	0	0	0	0	0	0	0	0	0
KPFA00005	3	0	0	0	0	0	0	0	0	3	0
KPFA00013	1	0	0	0	0	0	0	0	0	0	0
KPFA00008	0	0	0	0	0	0	0	0	0	0	0
KPFA00006	0	0	0	0	0	0	0	0	0	0	0
KPFA00773	0	0	0	0	0	0	0	1	0	0	0
KPFA00012	0	0	0	1	0	0	0	0	0	0	0
KPFA00010	0	0	0	0	0	0	0	0	0	0	0
KPFA00018	0	0	0	0	0	0	0	0	0	0	0
KPFA00041	0	0	0	1	0	0	0	0	0	0	0
KPFA00014	0	0	0	0	3	0	0	0	0	0	0
KPFA00011	0	0	0	0	0	0	0	0	0	0	0
KPFA00050	0	0	0	4	0	0	0	0	0	0	0
KPFA00010	1	0	0	1	0	0	0	0	0	1	0

~ 100 to ~300 features per patient

* Yu et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. JAMIA 2015

Feature Engineering: *Automated*

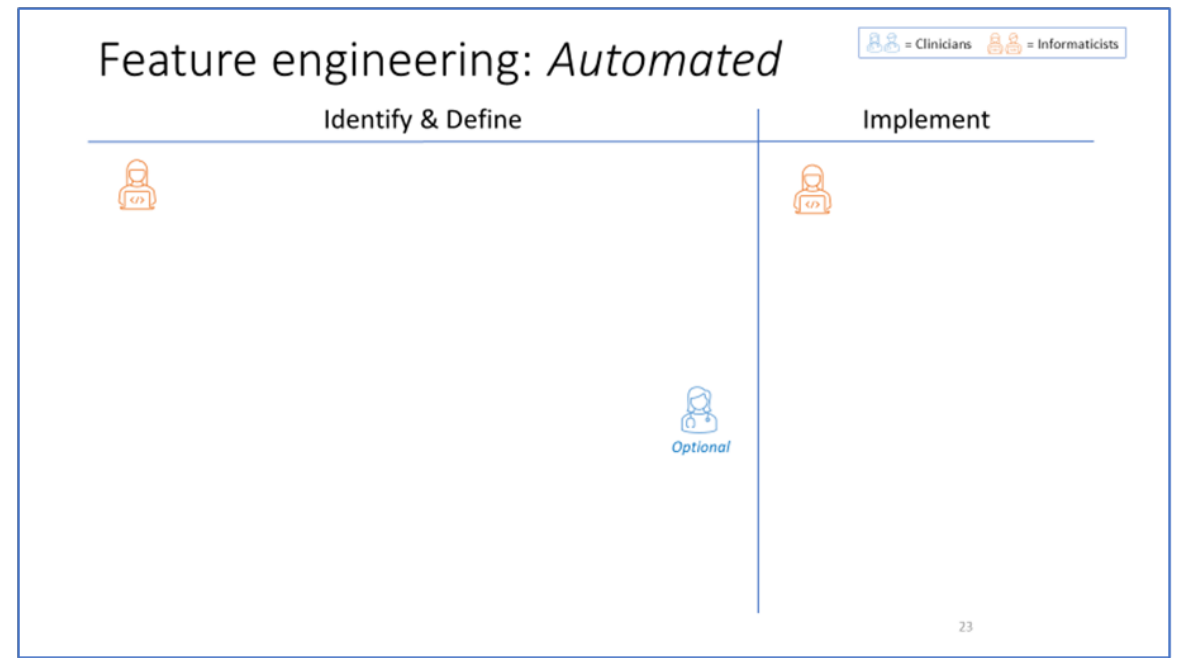
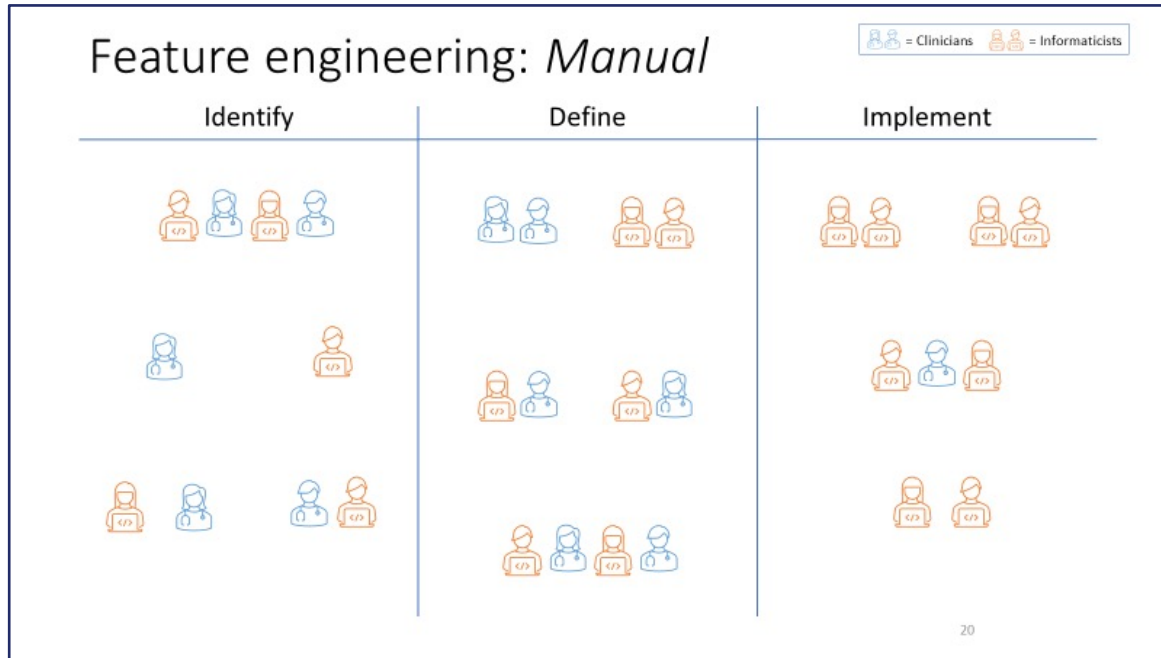
  = Clinicians   = Informaticists

Identify & Define

Implement



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?

Feature Engineering Example: Manual Structured

92 structured features manually engineered for acute pancreatitis

GROUP1	ABDIMG_CT_14	INTEST_OBS_SD	HEPATITIS_CS	MYOCARD_ISCH_PY
GROUP2	ABDIMG_MR_14	INTEST_OBS_CS	HEPATITIS_PY	GALL_BIL_SD
GROUP3	APLAB_MAX_ULN_14BEF_14AFT	INTEST_OBS_PY	INFLUENZA_SD	GALL_BIL_CS
UPCLASS_EDIP	APLAB_MAX_GT3_14BEF_14AFT	ILEUS_SD	INFLUENZA_CS	GALL_BIL_PY
ENRLYEARS	CHR_PANCR_SD	ILEUS_CS	INFLUENZA_PY	GB_CANCER_SD
AGE_AT_EVENT_IN_YRS	CHR_PANCR_CS	ILEUS_PY	FOOD_POIS_SD	GB_CANCER_CS
SEXF	CHR_PANCR_PY	CONSTIPATION_SD	FOOD_POIS_CS	GB_CANCER_PY
RACE_WHITE	PANC_CNCR_SD	CONSTIPATION_CS	FOOD_POIS_PY	IBD_SD
HISPANIC	PANC_CNCR_CS	CONSTIPATION_PY	ASCITES_SD	IBD_CS
SMOKE_CURR_SELF_365	PANC_CNCR_PY	MESENT_ISCH_SD	ASCITES_CS	IBD_PY
SMOKE_FORMER_SELF_365	PEPTIC_ULCER_SD	MESENT_ISCH_CS	ASCITES_PY	GASTRO_SD
SMOKE_DX_PX_365	PEPTIC_ULCER_CS	MESENT_ISCH_PY	NEPHROLITH_SD	GASTRO_CS
ALC_SD	PEPTIC_ULCER_PY	DIVERTICUL_SD	NEPHROLITH_CS	GASTRO_PY
ALC_CS	GASTRITIS_SD	DIVERTICUL_CS	NEPHROLITH_PY	ESOPHAGITIS_SD
ALC_PY	GASTRITIS_CS	DIVERTICUL_PY	DKA_SD	ESOPHAGITIS_CS
HYPERTRIG_SD	GASTRITIS_PY	APPENDIC_SD	DKA_CS	ESOPHAGITIS_PY
HYPERTRIG_CS	GERD_SD	APPENDIC_CS	DKA_PY	
HYPERTRIG_PY	GERD_CS	APPENDIC_PY	MYOCARD_ISCH_SD	
ERCP	GERD_PY	HEPATITIS_SD	MYOCARD_ISCH_CS	

Maximum lipase lab (normalized)
+/-14 days from diagnosis date

Feature Engineering Example: Manual NLP*

Anaphylaxis NLP dictionary for 71 concepts (843 terms)			
<ul style="list-style-type: none"> • BRADYCARDIA (13) • CARDIACARRHYTH (8) • CARDIOCOLLAPSE (2) • COLLAPSE (2) • END ORGAN (2) • HYPOTENSION (77) • PALPITATIONS (3) • SHOCK (3) • SYNCOPE (30) • TACHYCARDIA (9) • ABDOPAIN (3) • VOMIT (1) • AIRWAY (4) • AIRWAY CONstriction (4) • ALTERED MENTATION (1) • APHONIA (3) • BREATH (6) • BRONCHOSPASM (1) • CHEST DISCOMFORT (2) • CHEST TIGHTNESS (9) 	<ul style="list-style-type: none"> • COARSE BREATH SOUND (4) • DYSPHONIA (1) • DYSPNEA (55) • HOARSENESS (7) • HYPOXEMIA (6) • HYPOXIA (3) • IMPENDING DOOM (2) • INTUBATION (6) • LARYNGEAL OEDEMA (1) • RESP COMPROMISE (3) • RESP DISTRESS (2) • RESPFail (1) • RONCHI (2) • STRIDOR (3) • TACHYPNEA (5) • THROAT CLOSURE (14) • THROAT TIGHTNESS (34) • TIGHTNESS BREATHING (1) • VOICE QUALITY (1) • WHEEZE (8) 	<ul style="list-style-type: none"> • ANGIOEDEMA (102) • DIFFICULTY SWALLOWING (14) • DYSPHAGIA (1) • EDEMA (4) • ERYTHEMA (42) • EYE SWELLING (33) • FACIAL SWELLING (20) • FLUSH (38) • HIVES (68) • ITCHING (14) • ITCHY SOFT TISSUE (15) • METALLIC TASTE (1) • MOUTH (1) • MOUTHSWELL (4) • ORALSWELL (4) • PRURITUS (15) • RASH (7) • REACTION (1) • SOFT TISSUE SWELLING (4) • SWELLING (31) 	<ul style="list-style-type: none"> • THROAT (4) • TINGLING (1) • TINGLY SOFT TISSUE (14) • URTICARIA (24) • ALLERGReact (5) • ANAPH (5) • COMPLAINT (12) • DIAGNOSIS (8) • DIFFERENTIAL (1) • HYPO (6) • IMPRESSION (1)
<p>Group: • REDUCED BLOOD PRESSURE • GASTROINTESTINAL • RESPIRATORY COMPROMISE • SKIN/MUCOSAL • OTHER</p>			

* Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning, American Journal of Epidemiology, Volume 192, Issue 2, February 2023, Pages 283–295, <https://doi.org/10.1093/aje/kwac182>

Feature Engineering Example: Automated (NLP)

Symptomatic COVID-19 disease (N=158)

#	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
34	Chronic Obstructive Airway Disease	C0024117
35	Chronic obstructive pulm. disease	C3714496
36	combination - answer to question	C3811910
37	Common Cold	C0009443
38	Communicable Diseases	C0009450
39	Community Transmission	C5392207
40	Complication	C0009566

#	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
67	Fever	C0015967
68	Fever symptoms (finding)	C0424755
69	Fibrin fragment D	C0060323
70	Functional disorder	C0277785
71	Gastrointestinal System Finding	C1333803
72	Glucocorticoids	C0017710
73	Has difficulty doing (qualifier)	C1299586
74	Headache	C0018681
75	Heart Diseases	C0018799
76	Heart failure	C0018801
77	High risk of	C0332167
78	Human Immunodef. Vir. Meas.	C5202935
79	hydrocortisone	C0020268
80	hydroxychloroquine	C0020336

#	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	C0003250
107	Mucocutan. Lymph Node Synd.	C0026691
108	Multiple Organ Failure	C0026766
109	Muscle Fatigue	C0242979
110	Muscle strain	C0080194
111	Myalgia	C0231528
112	Myocarditis	C0027059
113	Nausea or vomiting	C3843946
114	Noninvasive Ventilation	C1997883
115	Nucleic Acid Amplification Tests	C0200932
116	Obesity	C0028754
117	Organ Transplantation	C0029216
118	oxygen	C0030054
119	Oxygen Therapy Care	C0184633
120	Patient in hospital	C0701159

#	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
153	Thrombus	C0087086
154	Tissue damage	C0010957
155	tocilizumab	C1609165
156	Viral Load result	C0376705
157	Virus Diseases	C0042769
158	Worse	C1457868

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
34	Chronic Obstructive Airway Disease	C0024117
35	Chronic obstructive pulm. disease	C3714496
36	combination - answer to question	C3811910
37	Common Cold	C0009443
38	Communicable Diseases	C0009450
39	Community Transmission	C5392207
40	Complication	C0009566

#	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
67	Fever	C0015967
68	Fever symptoms (finding)	C0424755
69	Fibrin fragment D	C0060323
70	Functional disorder	C0277785
71	Gastrointestinal System Finding	C1333803
72	Glucocorticoids	C0017710
73	Has difficulty doing (qualifier)	C1299586
74	Headache	C0018681
75	Heart Diseases	C0018799
76	Heart failure	C0018801
77	High risk of	C0332167
78	Human Immunodef. Vir. Meas.	C5202935
79	hydrocortisone	C0020268
80	hydroxychloroquine	C0020336

#	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	C0003250
107	Mucocutan. Lymph Node Synd.	C0026691
108	Multiple Organ Failure	C0026766
109	Muscle Fatigue	C0242979
110	Muscle strain	C0080194
111	Myalgia	C0231528
112	Myocarditis	C0027059
113	Nausea or vomiting	C3843946
114	Noninvasive Ventilation	C1997883
115	Nucleic Acid Amplification Tests	C0200932
116	Obesity	C0028754
117	Organ Transplantation	C0029216
118	oxygen	C0030054
119	Oxygen Therapy Care	C0184633
120	Patient in hospital	C0701159

#	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
153	Thrombus	C0087086
154	Tissue damage	C0010957
155	tocilizumab	C1609165
156	Viral Load result	C0376705
157	Virus Diseases	C0042769
158	Worse	C1457868

Feature Engineering: Best Practices, Future Work

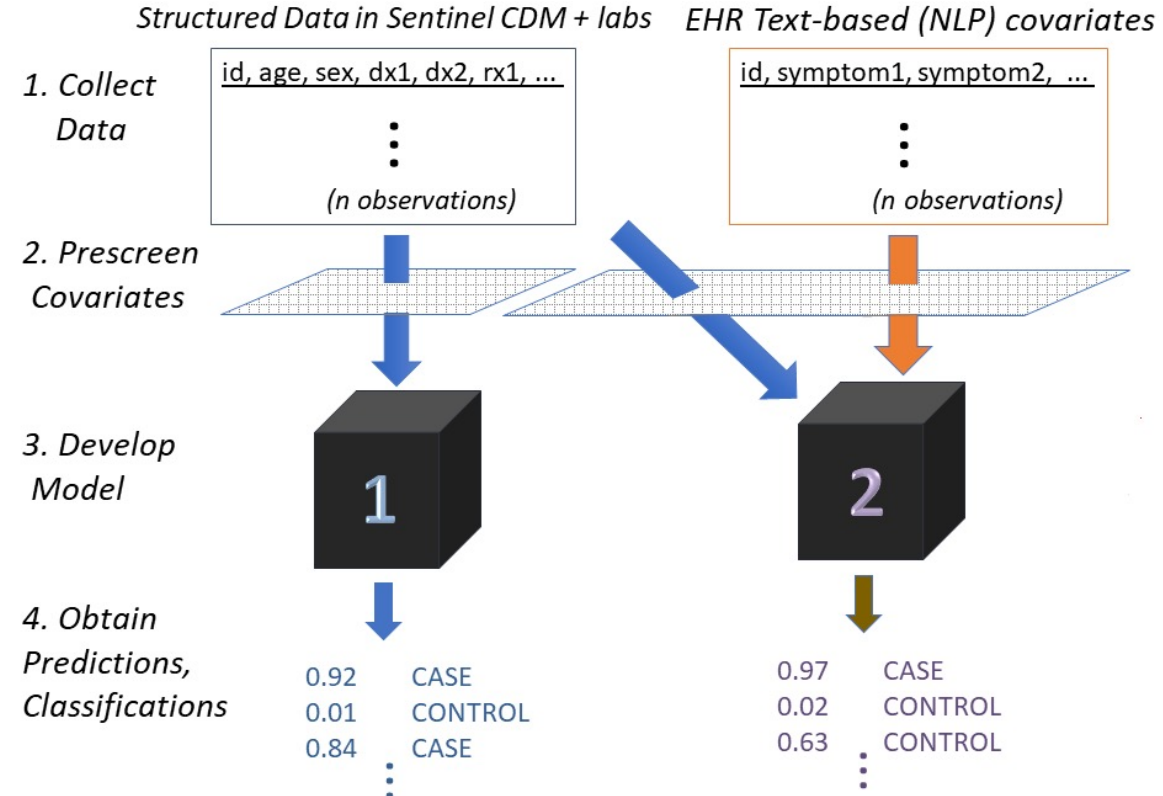
- Goal: Measure things that help distinguish true cases from non-cases
- Challenge: Manual approaches are time/expert-intensive, operator-dependent,
 - Wasted effort if based on idiosyncratic local data, don't improve model performance
- ➔ • Best practices:
 - Feature engineering is enhanced by *domain knowledge*
 - Engineer *many features* to capture information that may help distinguish cases from non-cases
 - Use *manual curation sparingly* (for known, high-value features)
 - Engineer for generalizability across settings
 - If tailoring is needed, design for easy tailoring
- Future work:
 - **Automated** engineering approaches (at least as a **starting point?**)



Model Development

Model Development

- Goal: construct a **useful** prediction model
- Challenges:
 - Clinical complexity and data complexity of many phenotypes
 - Model training requires (expensive) gold-standard data
- Best practices:
 - Incorporate domain knowledge
 - Apply outcome blind dimension reduction without sacrificing predictive power
 - Consider diverse combinations of dimension reduction strategies & algorithms



Model Development

- Consider diverse combinations of dimension reduction strategies & algorithms

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

=

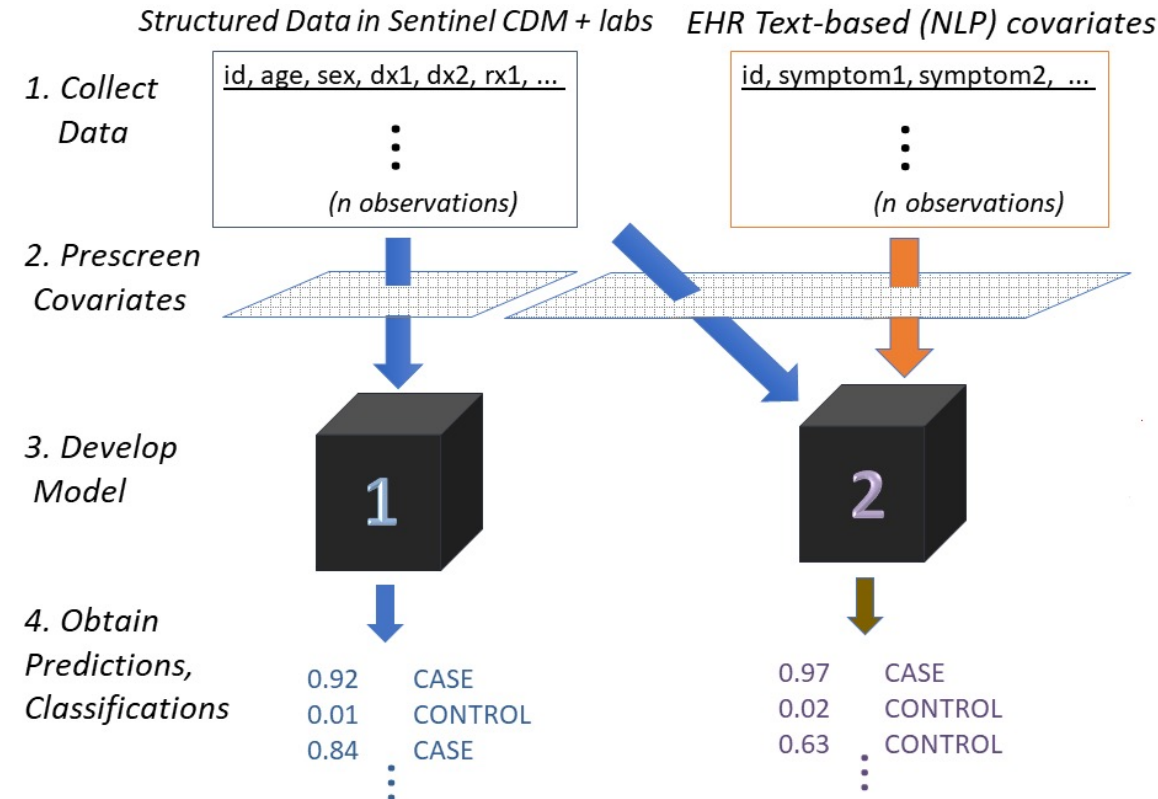
Combinations	
1	GLM-Retain-All
2	GLM-PAM
3	GLM-LASSO
4	Elastic-net-Retain-All
5	Elastic-net-PAM
6	Elastic-net-LASSO
7	XGBoost-v1-Retain-All
8	XGBoost-v1-PAM
9	XGBoost-v1-LASSO
10	XGBoost-v2-Retain-All
11	XGBoost-v2-PAM
12	XGBoost-v2-LASSO
13	BART-v1-Retain-All
14	BART-v1-PAM
15	BART-v1-LASSO
16	BART-v2-Retain-All
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

Model Development

- Goal: construct a **useful** prediction model
- Challenges:
 - Clinical complexity and data complexity of many phenotypes
 - Model training requires (expensive) gold-standard data
- Best practices:
 - Incorporate domain knowledge
 - Apply outcome blind dimension reduction without sacrificing predictive power
 - Consider diverse combinations of dimension reduction strategies & algorithms
- ➔ • Use V-fold cross-validation to make use of all the data
- Future work:
 - How to incorporate **silver standard surrogate outcome labels** in model training?

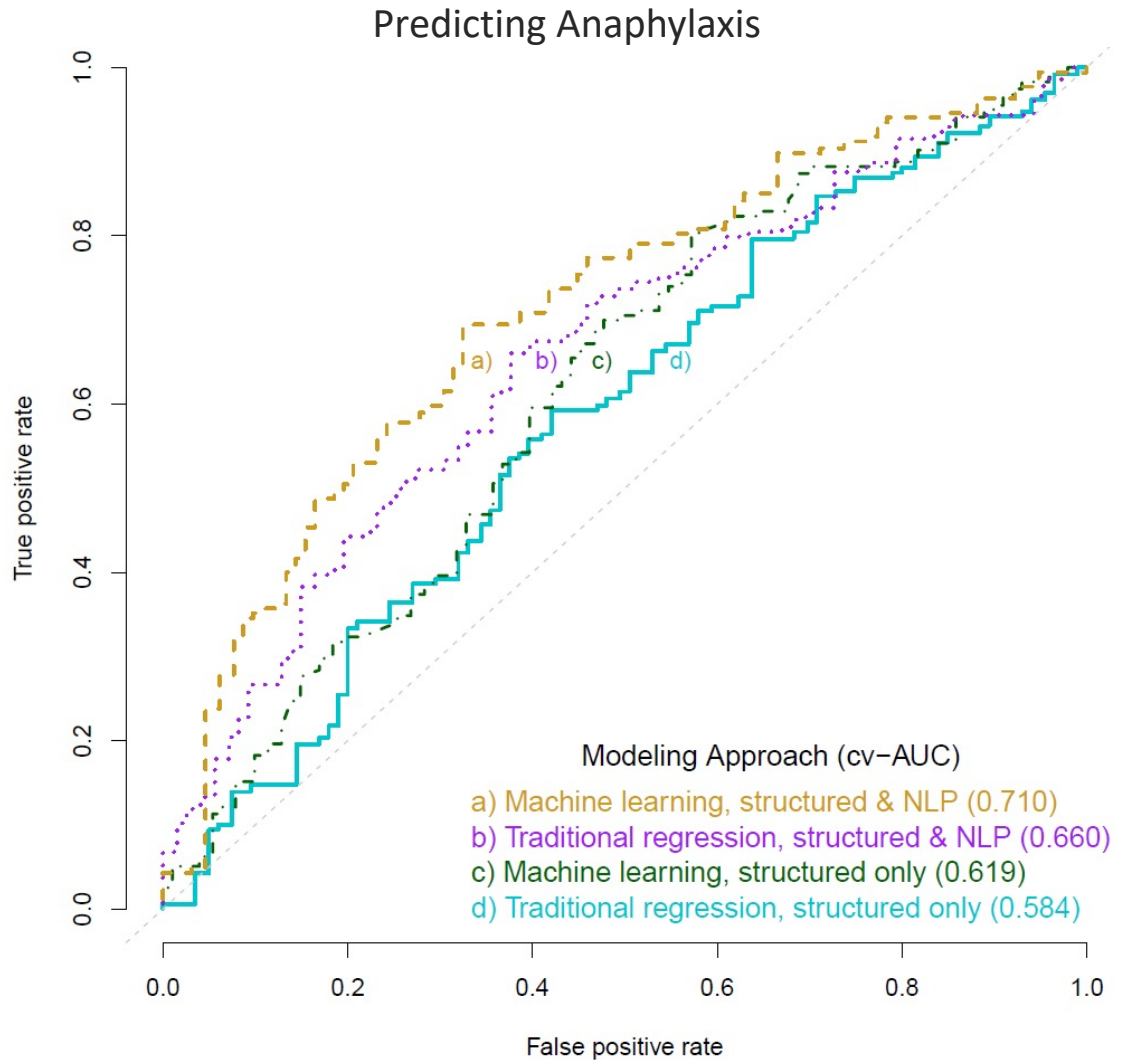




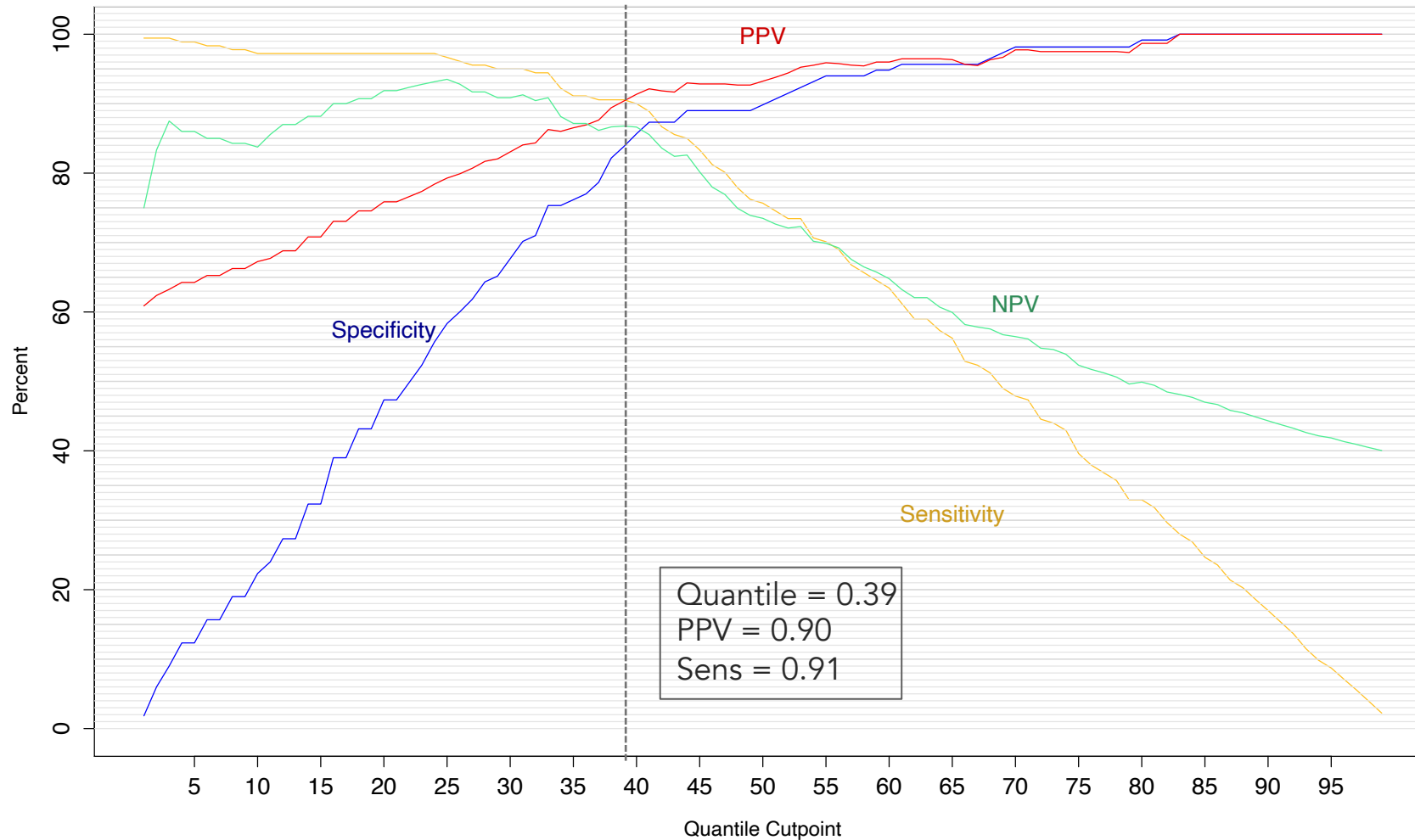
Model Evaluation and Reporting

Model Evaluation and Reporting

- Goal: Understand model performance in *unseen data*
- Challenges:
 - Evaluation requires (expensive) gold-standard outcome data
 - Bias and variance of causal effect estimates depend on model sensitivity and PPV
- Best practices:
 - Consider many performance metrics
 - Use cross-validated performance metrics relevant to use case
 - For FDA safety study outcomes, choosing a cut point of predicted probability to define case status should be informed by sensitivity and PPV at alternative cut points



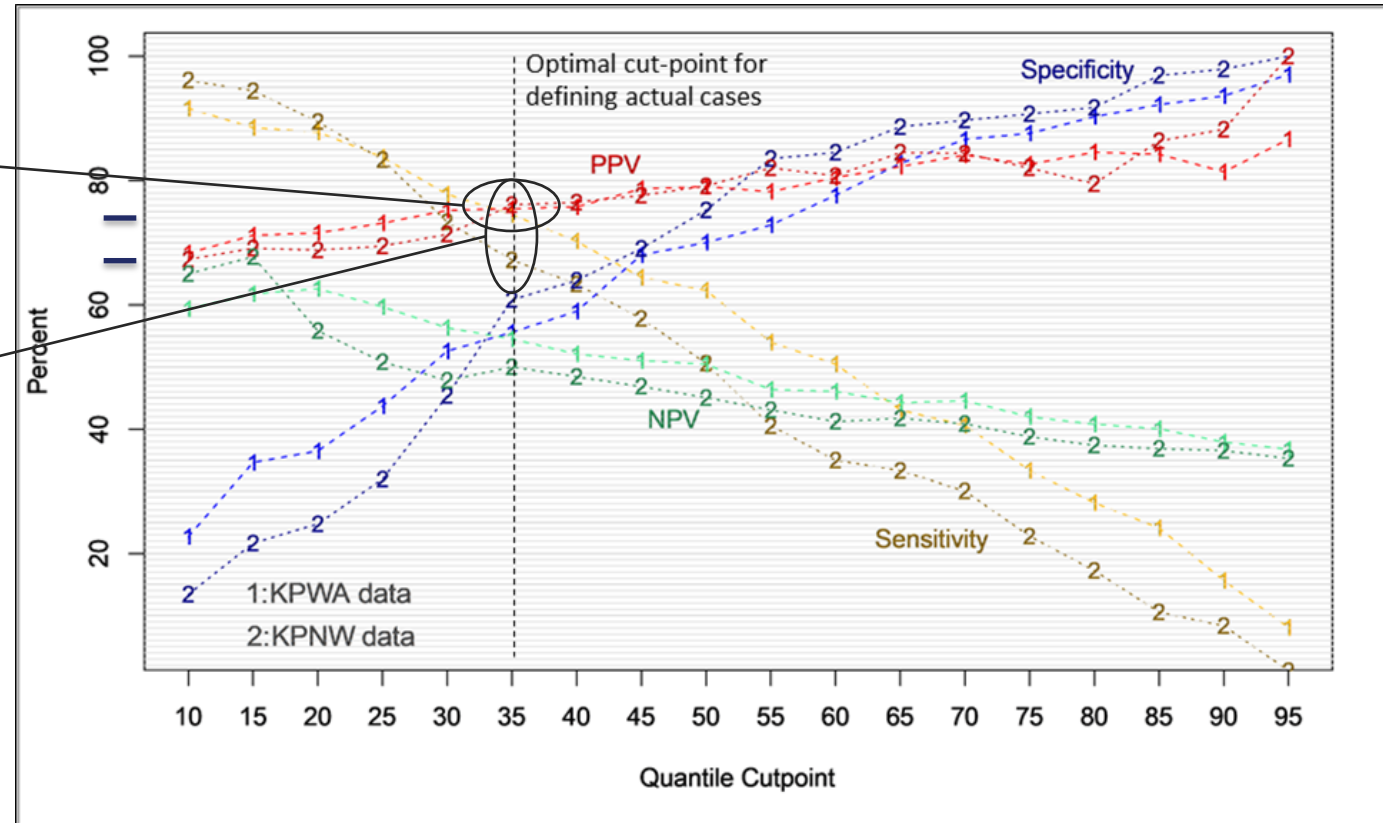
Model Evaluation: Acute Pancreatitis



Cross-validated performance metrics for a best-fitting model using structured data and NLP-derived data, *KPNW*, BART2 with LASSO dimension reduction.

Model Evaluation: Anaphylaxis, External Site

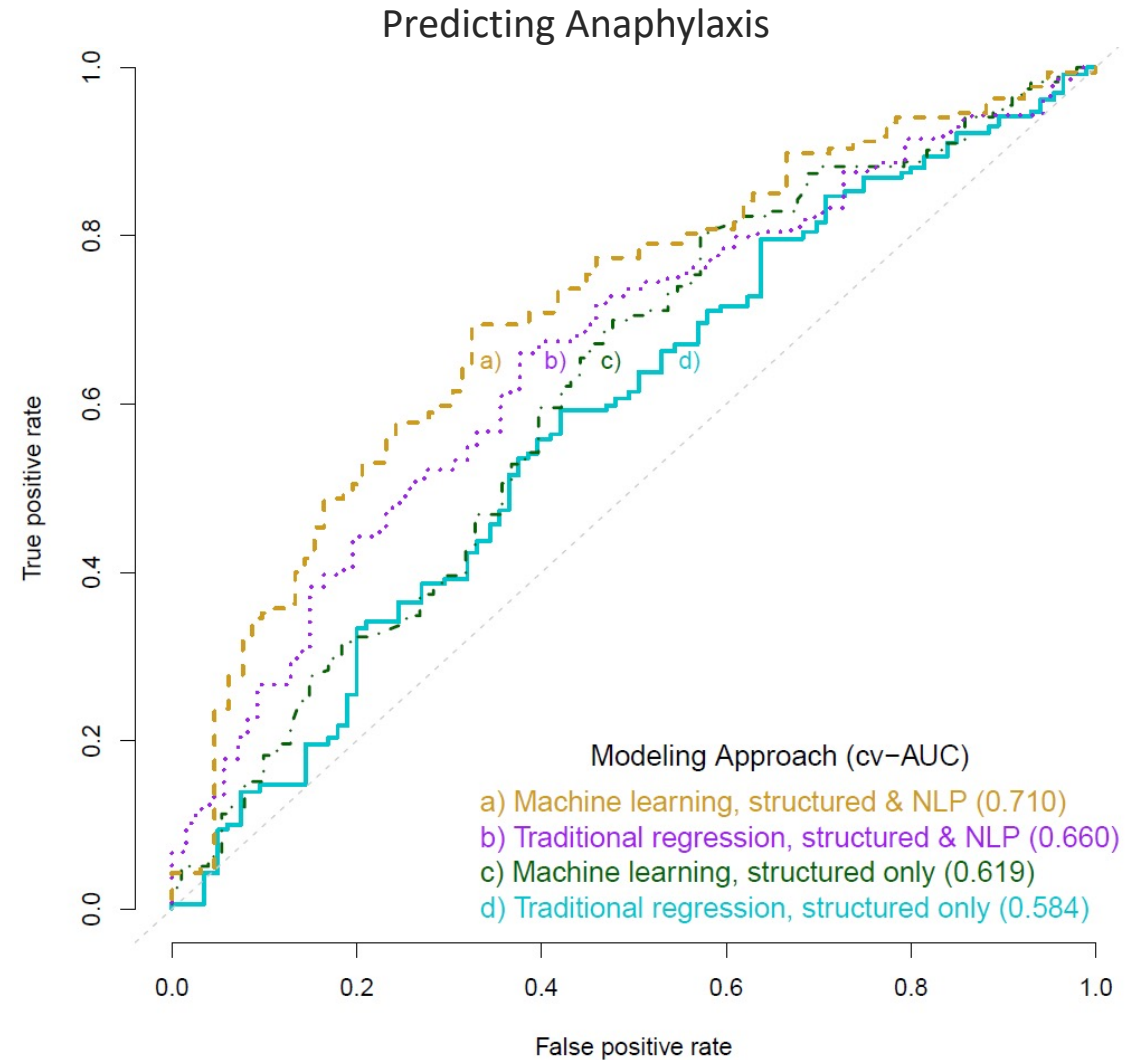
- Comparable PPV
- 7% drop in sensitivity at external site



Anaphylaxis model performance in 1) internal KPWA data and 2) external KPNW data at quantiles of predicted risk (BART2, retain-all)

Model Evaluation and Reporting

- Goal: Understand model performance in *unseen data*
- Challenges:
 - Evaluation requires (expensive) gold-standard outcome data
 - Bias and variance of causal effect estimates depend on model sensitivity and PPV
- Best practices:
 - Consider many performance metrics
 - Use cross-validated performance metrics relevant to use case
 - For FDA safety study outcomes, choosing a cut point of predicted probability to define case status should be informed by sensitivity and PPV at alternative cut points
- ➔ • **Caution:** Narrowly focusing on high PPV may undermine power to detect non-null associations
- Final algorithm choice guided by downstream performance, **transportability**, and **generalizability**



Model Evaluation: Selecting a Final Model

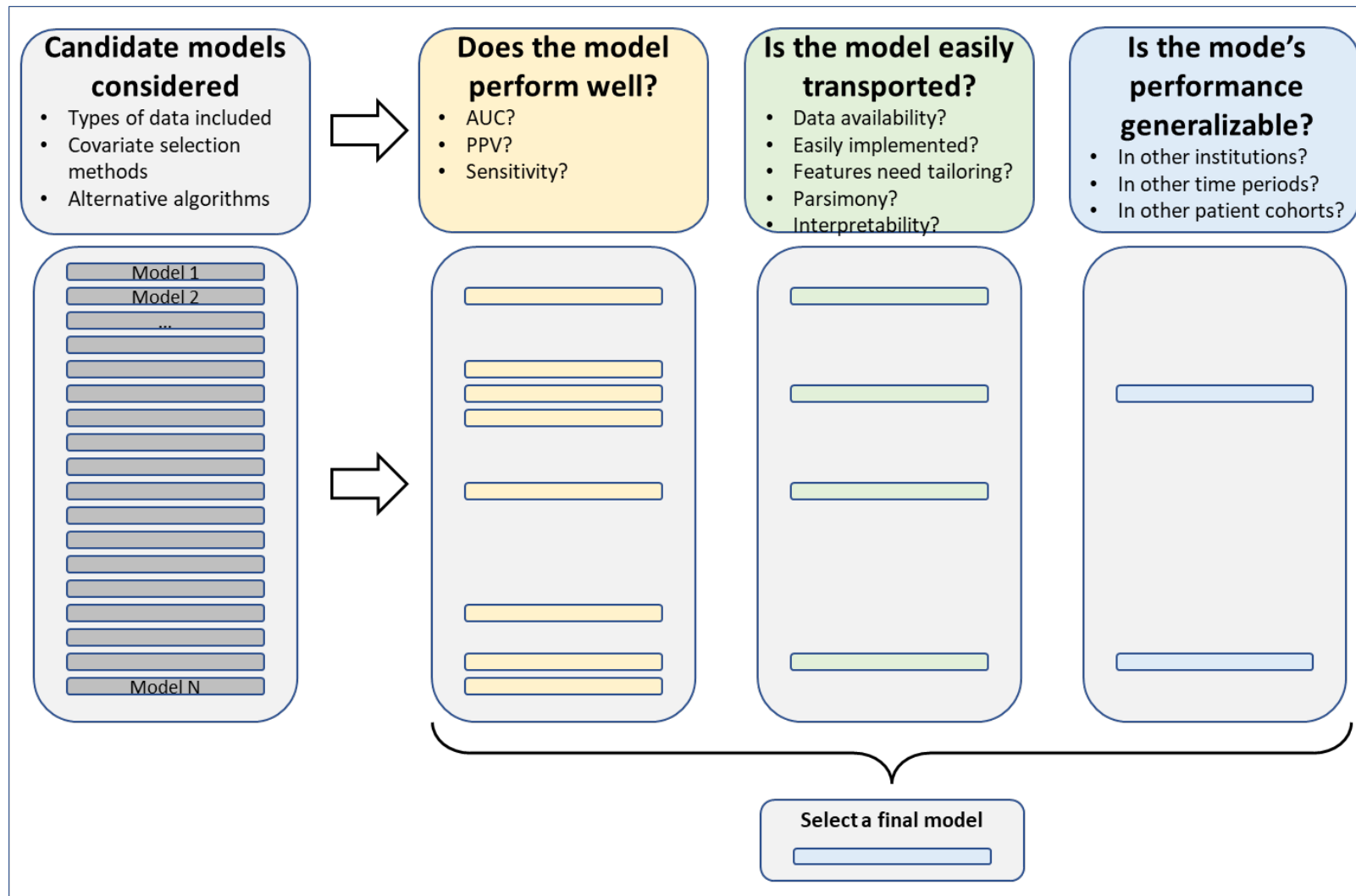
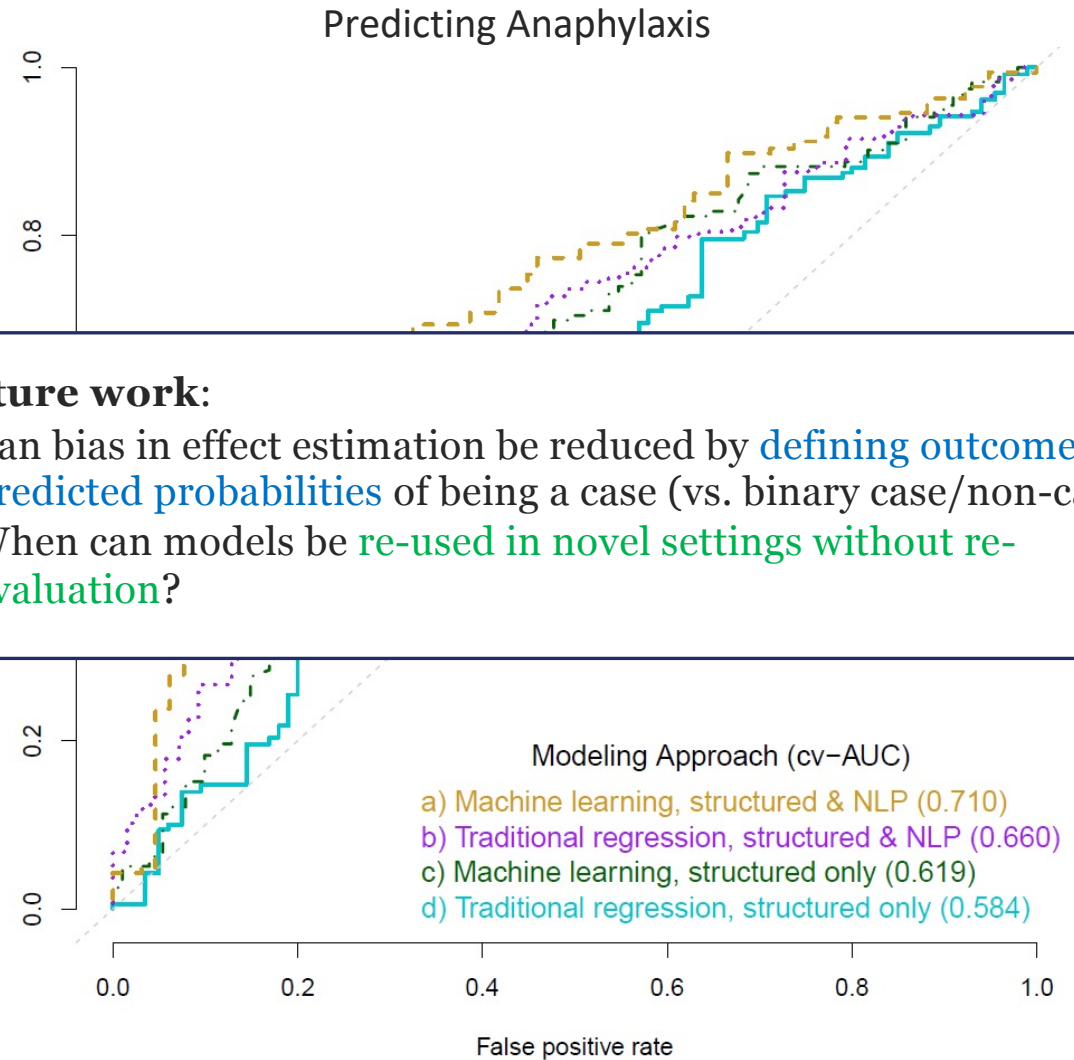


Figure 2. Selecting a final model based on considerations of model performance, model transportability, and model generalizability.

Model Evaluation and Reporting

- Goal: Understand model performance in *unseen data*
- Challenges:
 - Evaluation requires (expensive) gold-standard outcome data
 - Bias and variance of causal effect estimates depend on model sensitivity and PPV
- Best practices:
 - Consider many performance metrics
 - Use cross-validated performance metrics relevant to use case
 - For FDA safety study outcomes, choosing a cut point of predicted probability to define case status should be informed by sensitivity and PPV at alternative cut points
 - **Caution:** Narrowly focusing on high PPV may undermine power to detect non-null associations
 - Final algorithm choice guided by downstream performance, **transportability**, and **generalizability**

- **Future work:**
 - Can bias in effect estimation be reduced by **defining outcomes as predicted probabilities** of being a case (vs. binary case/non-case)?
 - When can models be **re-used in novel settings without re-evaluation**?





Implications and Next Steps

Implications and Next Steps

- NLP and machine learning have been shown to improve phenotype algorithm performance, but ...
- “General framework” principles and best practices should be considered to further enhance:
 - Efficient (“scalable”) development
 - Reusability of tools and methods
 - Generalizability to other settings
- Future methods-development work should consider:
 - NLP-assisted chart review
 - Strategic sampling of gold standard observations
 - Automated feature engineering approaches
 - Incorporating silver labels during model training
 - Probabilistic case definitions to reduce bias in effect estimation
 - When models be re-used in novel settings without re-evaluation

Motivation

- Goal: improve safety surveillance using observational data
- Active Risk Identification and Analysis (ARIA) system:

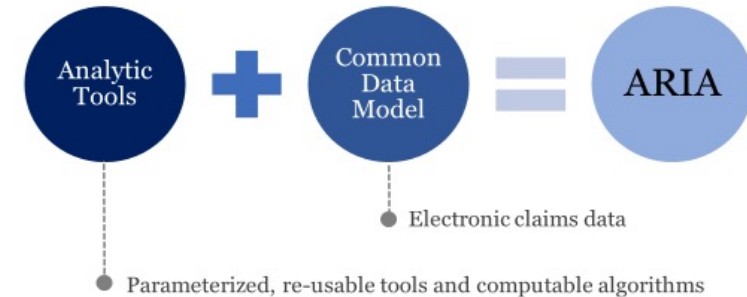


Image courtesy of Michael Nguyen

Acknowledgments

Sentinel Advanced Phenotyping Framework and Scalable Natural Language Processing (NLP) Teams

FDA

- Adebola Ajao
- Robert Ball
- Steven Bird
- Sara Karami
- Yong Ma
- Michael Nguyen
- Danijela Stojanovic
- Sanrat Wittayanukorn
- Mingfeng Zhang
- Yueqin Zhao

Harvard Pilgrim

Health Care Institute

- Adee Kennedy
- Judy Maro
- Elizabeth Messenger-Jones
- Kathleen Shattuck
- Mayura Shinde
- Darren Toh

Kaiser Washington

- Maralyssa Bann
- Will Bowers
- David Carrell
- David Cronkite
- James Floyd
- Monica Fujii
- Vina Graham
- Kara Haugen
- Eric Johnson
- Ron Johnson
- Ann Kelley
- Linda Kiel
- Jennifer Nelson
- Arvind Ramaprasan
- Mary Shea
- Brian Williamson
- Jing Zhou

Univ. of Michigan

- Xu Shi

Carelon Research

- Kevin Haynes

Duke University

- Keith Marsolo

Mass General Brigham

- Rishi Desai
- William Feldman
- Shamika More
- Shirley Wang

Univ. of Pennsylvania

- Kevin Johnson

Indiana University

- David Aronoff

Univ. of Michigan

- Xu Shi

Carelon Research

- Kevin Haynes

Duke University

- Keith Marsolo

Mass General Brigham

- Rishi Desai
- William Feldman
- Shamika More
- Shirley Wang

Univ. of Pennsylvania

- Kevin Johnson

Indiana University

- David Aronoff

Thank You

Contact: david.s.carrell@kp.org

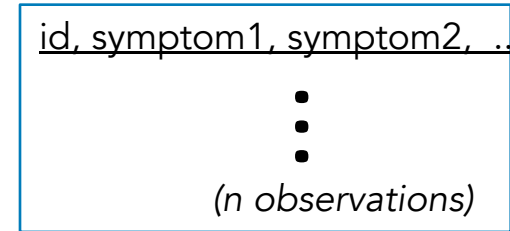
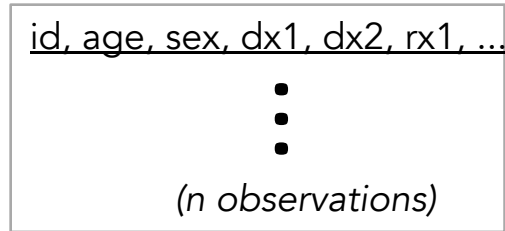


Extras

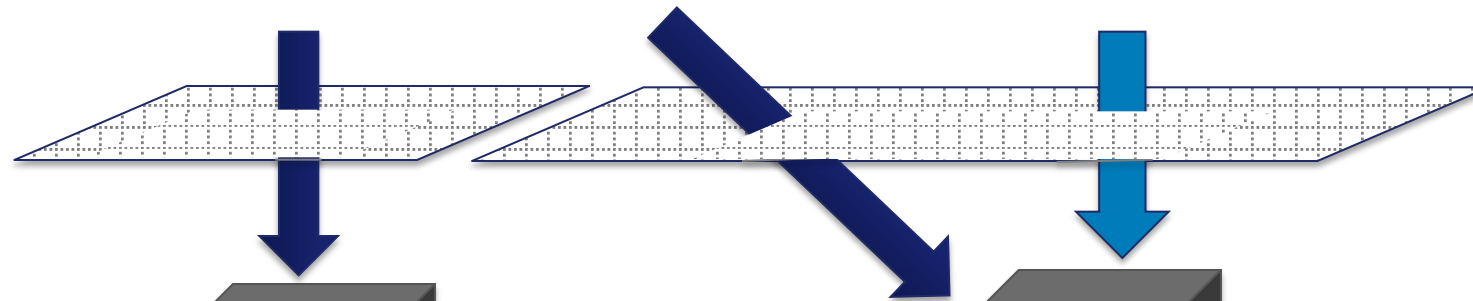
Model Development: Approach

Structured Data in Sentinel CDM + labs *EHR Text-based (NLP) covariates*

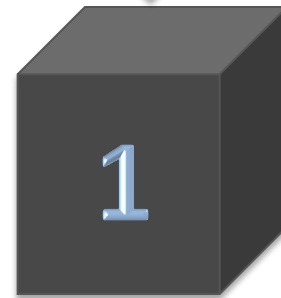
1. *Collect Data*



2. *Prescreen Covariates*



3. *Develop Model*



What's in the box?

4. *Obtain Predictions, Classifications*

0.92 CASE
0.01 CONTROL
0.84 CASE
⋮

0.97 CASE
0.02 CONTROL
0.63 CONTROL
⋮

Model Development: Approach

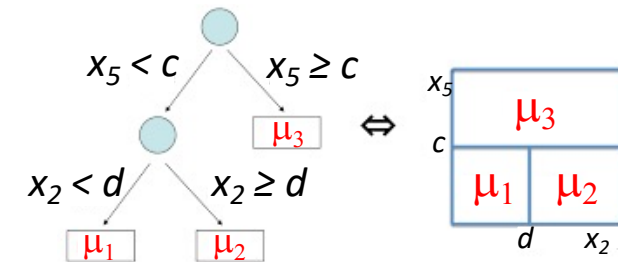
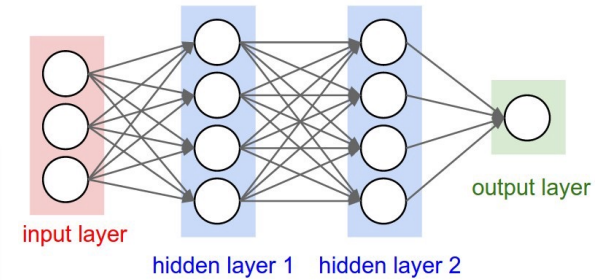
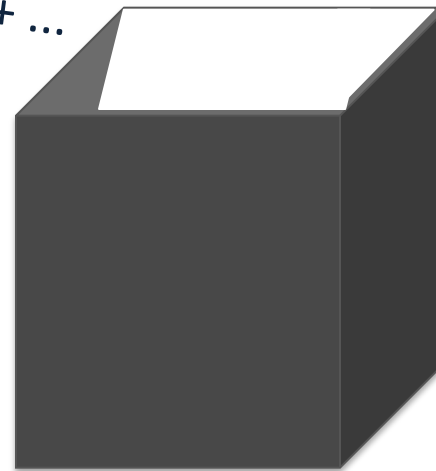
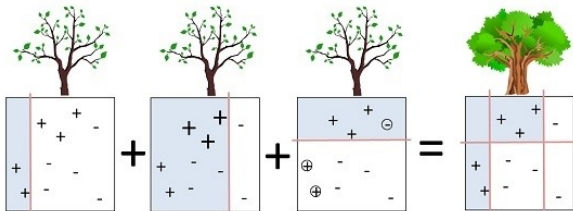
What's in the box?

- Logistic regression
- Elastic net
- Bayesian Additive Regression Trees
- Neural network
- Boosted Trees

Super Learner
(a weighted combination)

$$\beta_0 + \beta_1 * age + \beta_2 * ICD10 + \dots$$

Boosted Regression Tree is a hierarchical and supervised machine learning method that combines weak learners (binary splits) to strong prediction rules that allow a flexible partition of the feature space.



Model Evaluation: Anaphylaxis

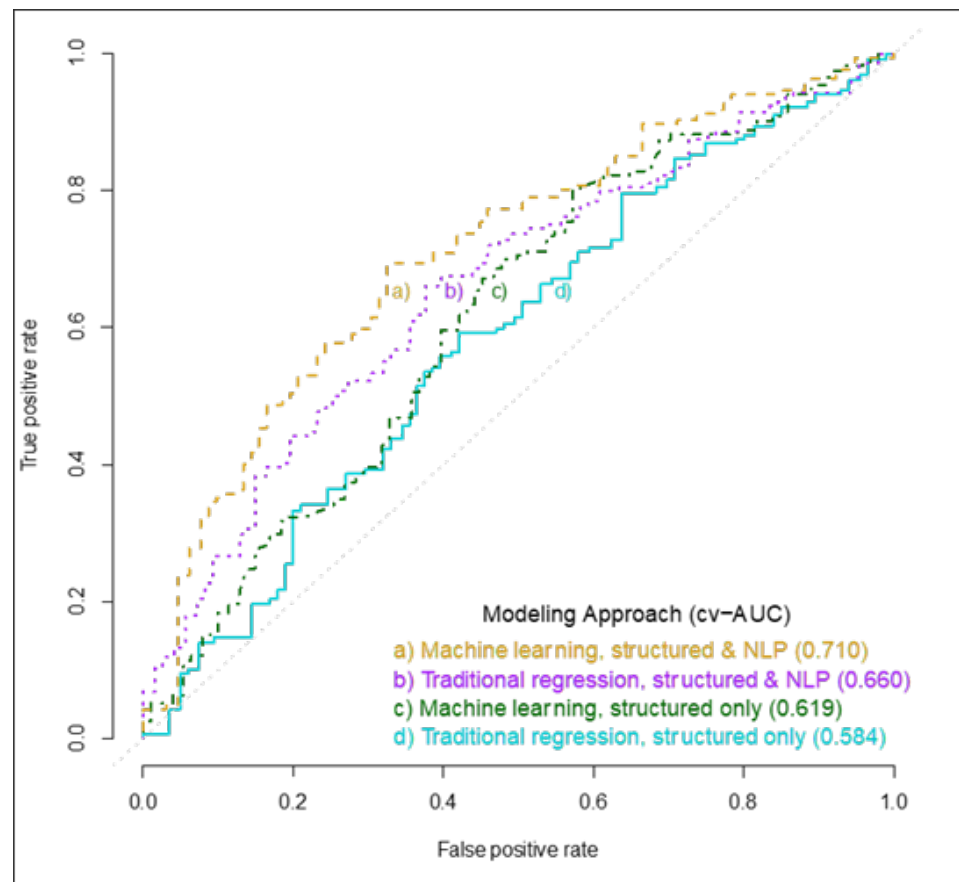
- All 25 models, **structured data** (only)
- **Performance differences** revealed in one table
- All 25 models, **structured and NLP data**

Table 2. Cross-validated weighted¹ AUC (cv-AUC) for KPWA algorithms predicting anaphylaxis case status based on A) structured data features and B) structured data and NLP features.

Feature set: Algorithm	Covariate selection strategy		
	LASSO	PAM	Retain All
A. Structured data features:			
Logistic Regression ²	0.584	0.584	0.564
Elastic Net	0.587	0.573	0.606
GBM 1	0.578	0.573	0.581
GBM 2	0.557	0.570	0.601
BART 1	0.586	0.560	0.594
BART 2	0.594	0.574	0.593
NNET 1	0.619	0.582	0.575
NNET 2	0.559	0.531	0.567
Super Learner ³	0.581 (all strategies combined)		
B. Structured data features and all NLP features:			
Logistic Regression	0.644	0.660	0.486
Elastic Net	0.664	0.650	0.649
GBM 1	0.604	0.610	0.677
GBM 2	0.604	0.621	0.672
BART 1	0.700	0.655	0.686
BART 2	0.710	0.652	0.704
NNET 1	0.572	0.617	0.579
NNET 2	0.633	0.653	0.655
Super Learner ³	0.688 (all strategies combined)		

Model Evaluation: Anaphylaxis KPWA

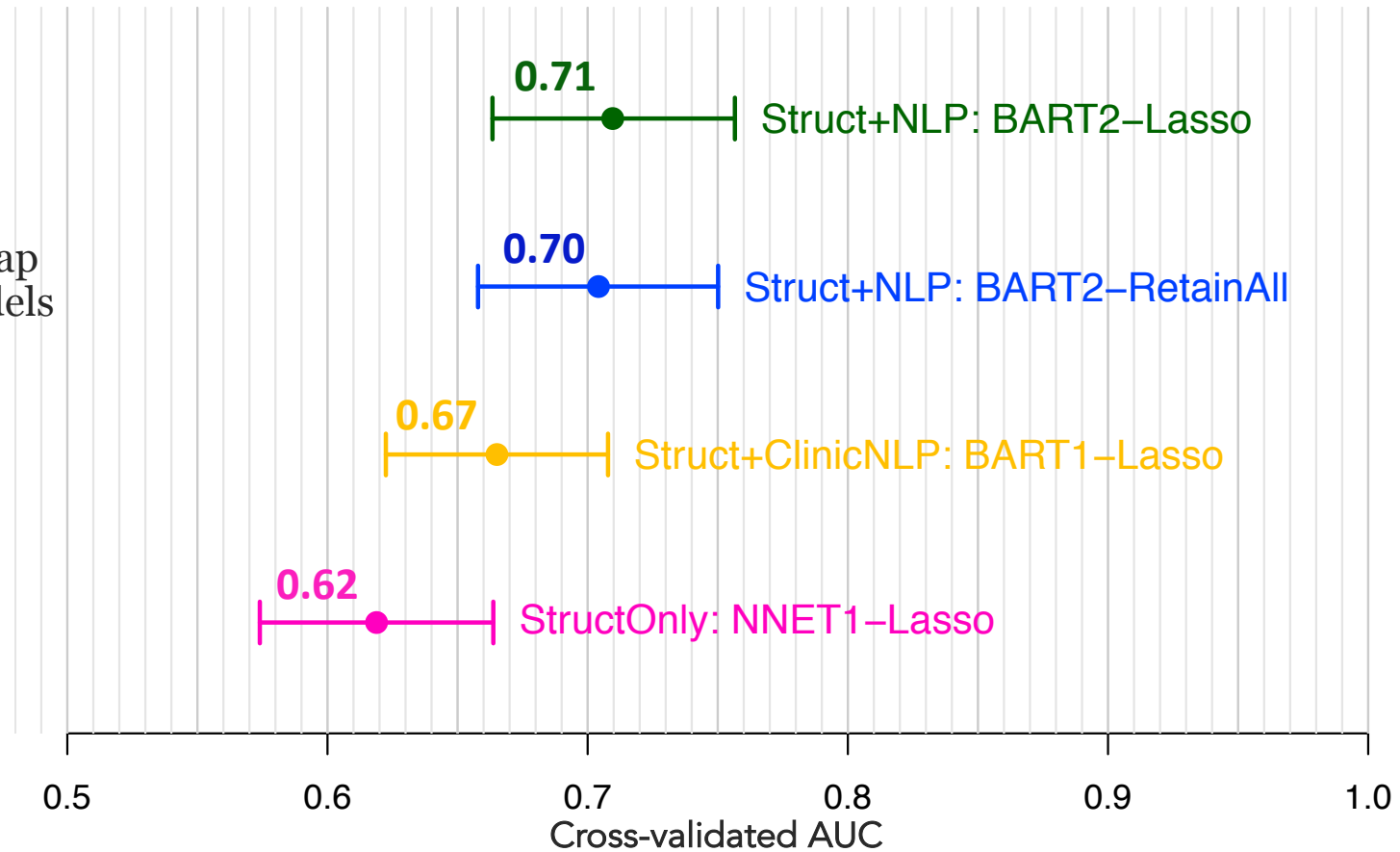
- AUC plots reveal where performance is improved



- Best of 25:
- a) Machine learning applied to structured and NLP data (0.710)
 - b) Traditional logistic regression applied to structured and NLP data (0.660)
 - c) Machine learning applied to structured data (0.619)
 - d) Traditional logistic regression applied to structured data (0.584)

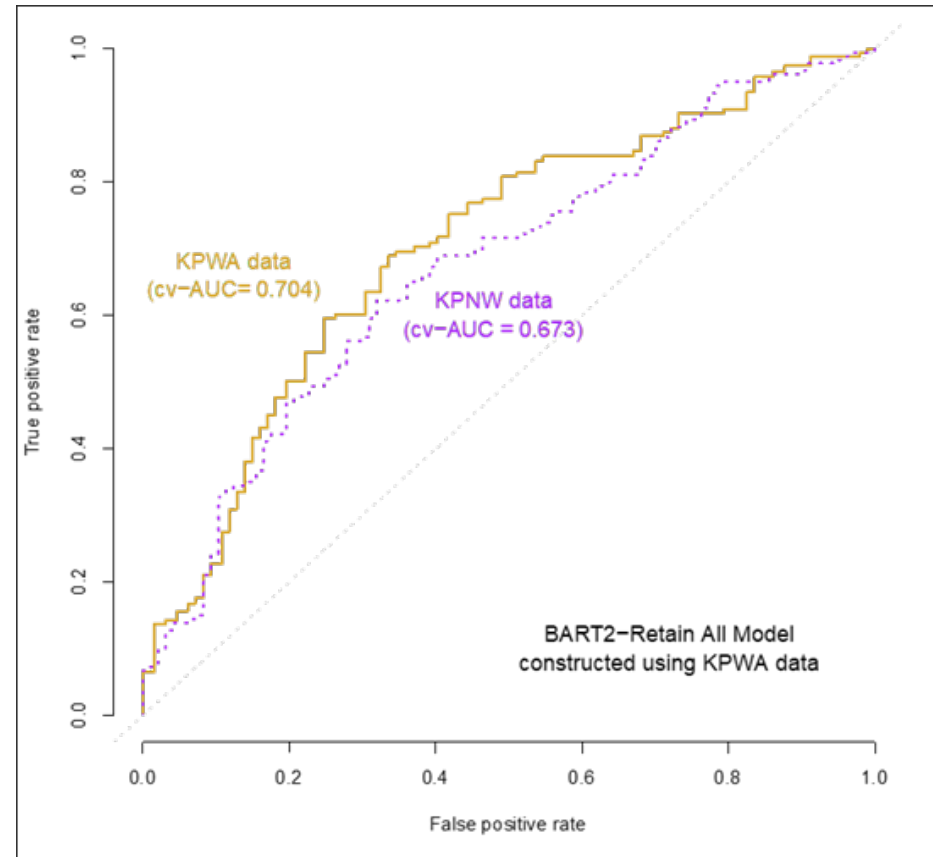
Model Evaluation: Anaphylaxis KPWA

- Error bars show overlap across models



Model Evaluation: External Validation

- AUC plots reveal change in performance at external site (KPNW)



Best anaphylaxis model developed using only KPWA data implemented and evaluated externally using KPNW data

Model Development: Challenges

1. Access to EHR is a mixed blessing. With high-dimensional data the relevant information is within reach but responding to the signal produced by key data elements remains a challenge.
2. The quantity of gold standard data available for training limits the complexity of the machine learning algorithms that can -be applied to model development.
 - High clinical complexity suggests feature-outcome associations are not straightforward; larger gold standard datasets may be needed.
3. When there is heterogeneity in the predictor-outcome associations within sub-populations larger amounts of training data are required. Even when overall performance is acceptable, performance in a minority sub-population may be poor.



Harmonizing Electronic Health Record and Claims Data Across FDA Sentinel Initiative Data Partners: Case Study and Lessons Learned

Xu Shi
University of Michigan

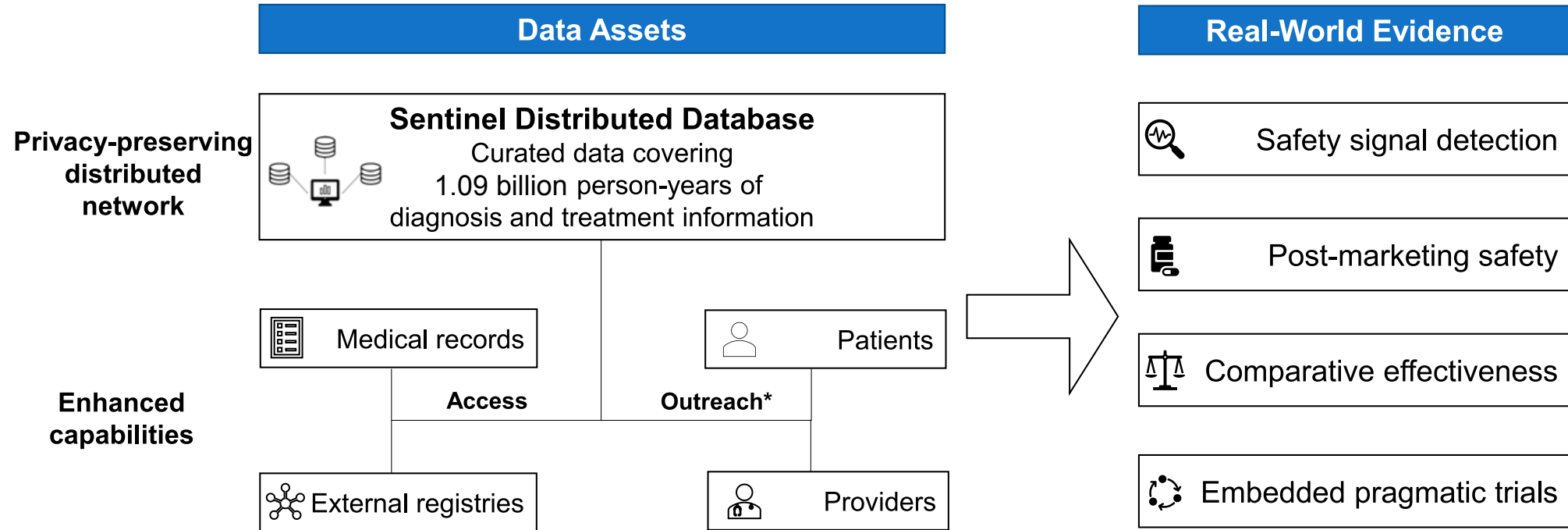
Outline

- Background
 1. FDA Sentinel's data assets to generate real-world evidence
 2. FDA Sentinel's privacy-preserving distributed network
 3. The Sentinel DATA harmonization project
- Methods
 1. Compare coding patterns between Kaiser Permanente Washington (KPWA) and Kaiser Permanente Northwest (KPNW)
 2. Automated code mapping between Kaiser Permanente Washington (KPWA) and Kaiser Permanente Northwest (KPNW)
- Results
 1. Study population, study period, and summary of coding
 2. Group- and code-level differences
 3. The “cataract” ICD-10 group
 4. Mapping of ICD-10 codes in the “cataract” group
- Validation
- Conclusion



Background

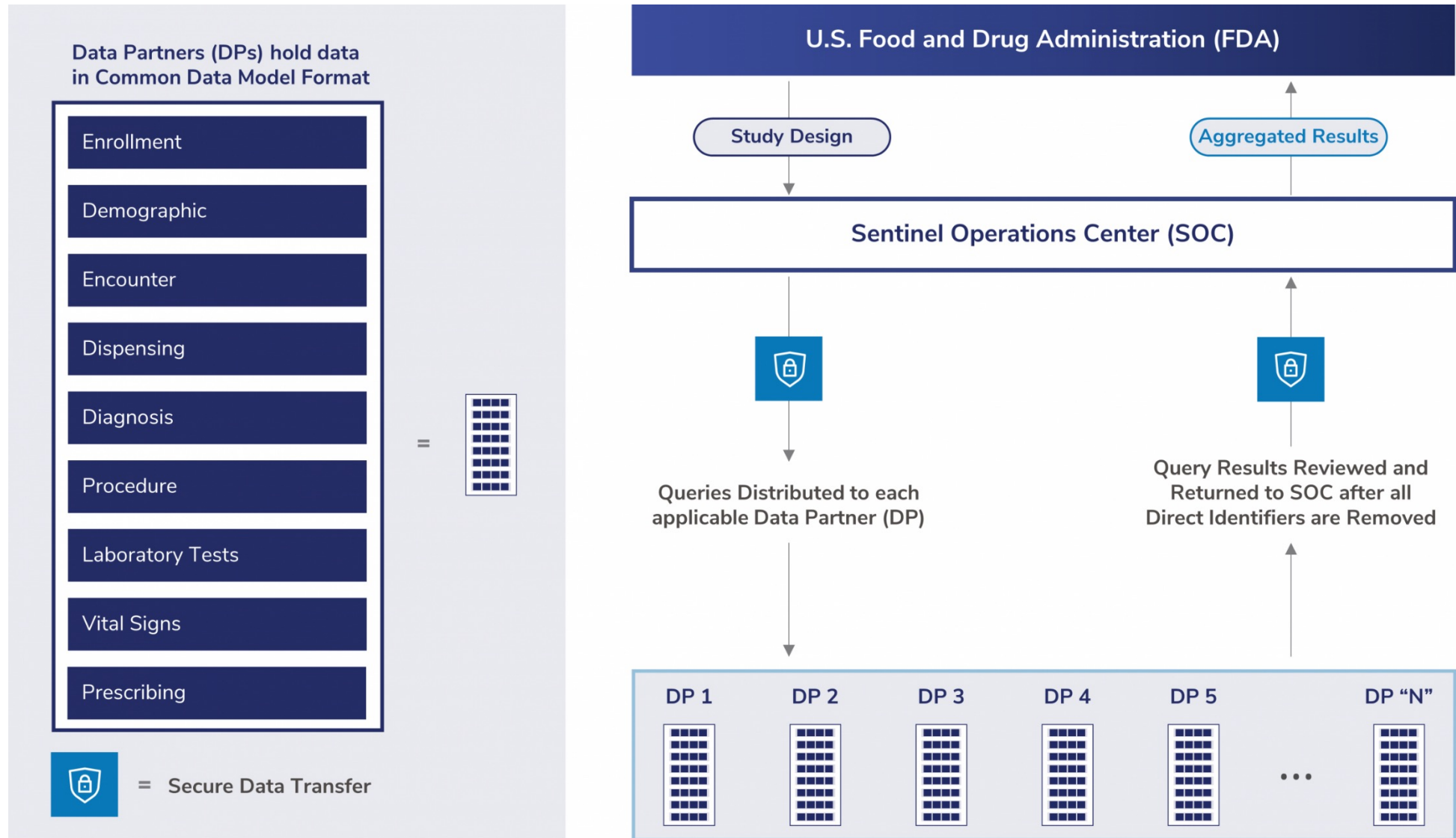
Background: Sentinel's Data Assets to Generate Real-World Evidence



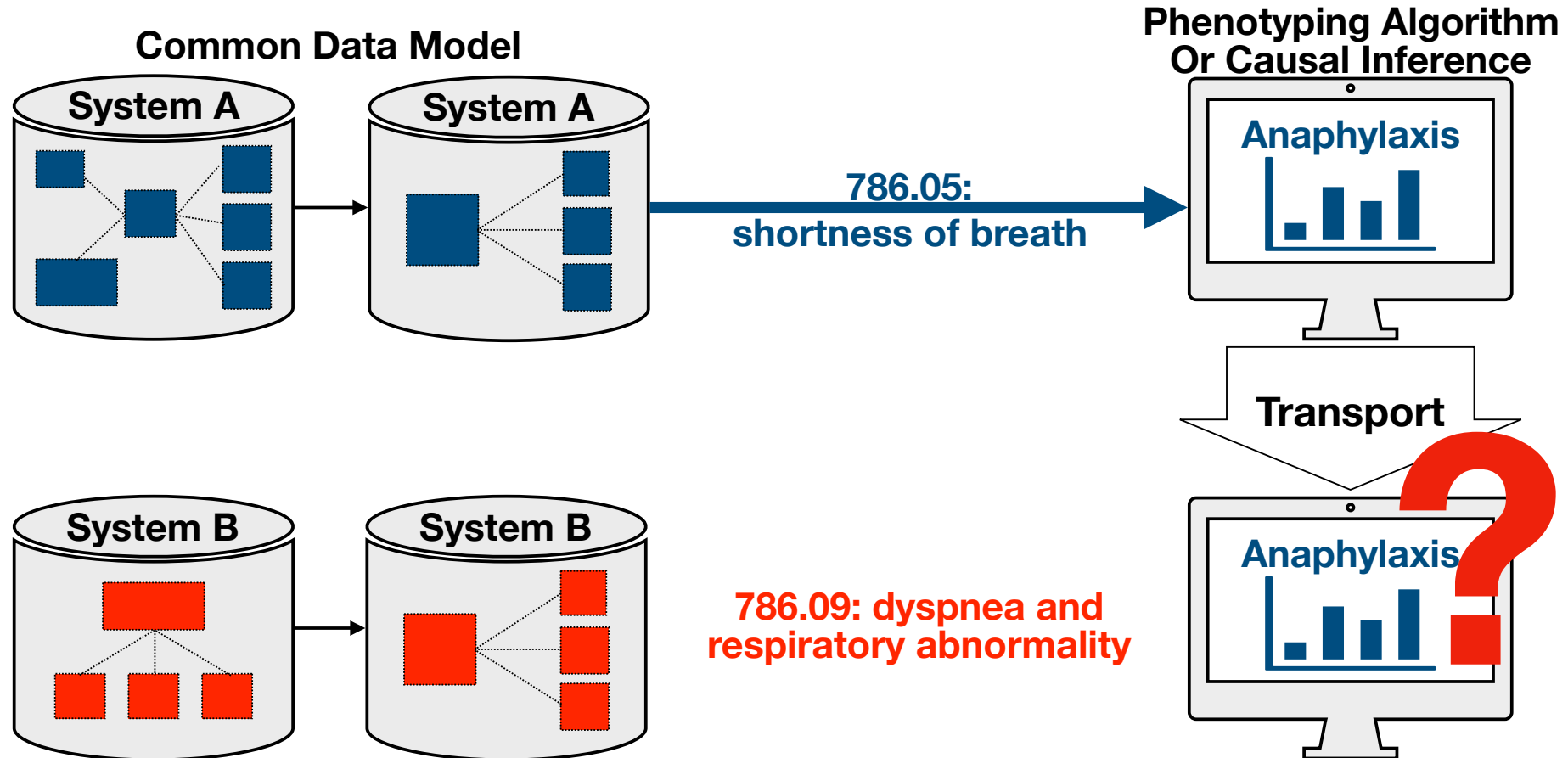
1.09 billion person-years of data from 17 data partners:

- 20.2 billion unique medical encounters
- 19.7 billion pharmacy dispensings
- 66.6 million members with at least one laboratory test result

Background: Sentinel's Privacy-Preserving Distributed Network

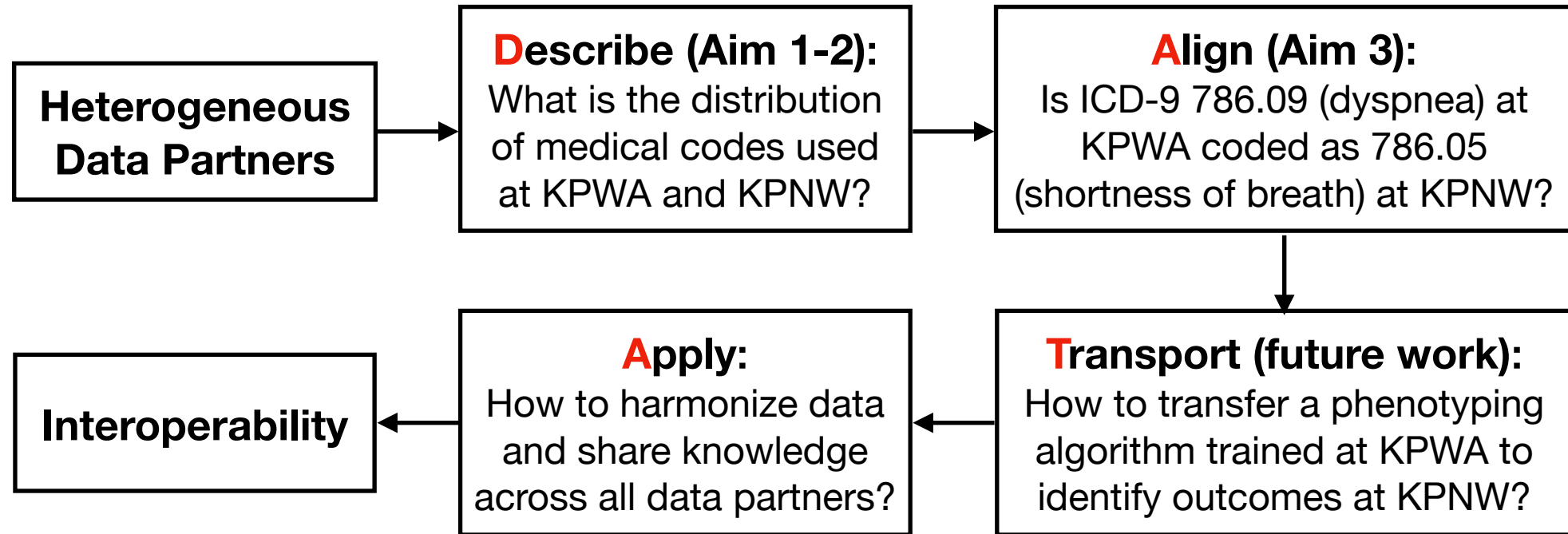


Background: A Motivating Example



Sentinel Common Data Model unifies the “vocabulary” but not the “dialect” of medical coding

Background: The Sentinel **DATA** Harmonization Project



Case study with two Sentinel Data Partners:

Kaiser Permanente Washington (KPWA) and Kaiser Permanente Northwest (KPNW)



Methods

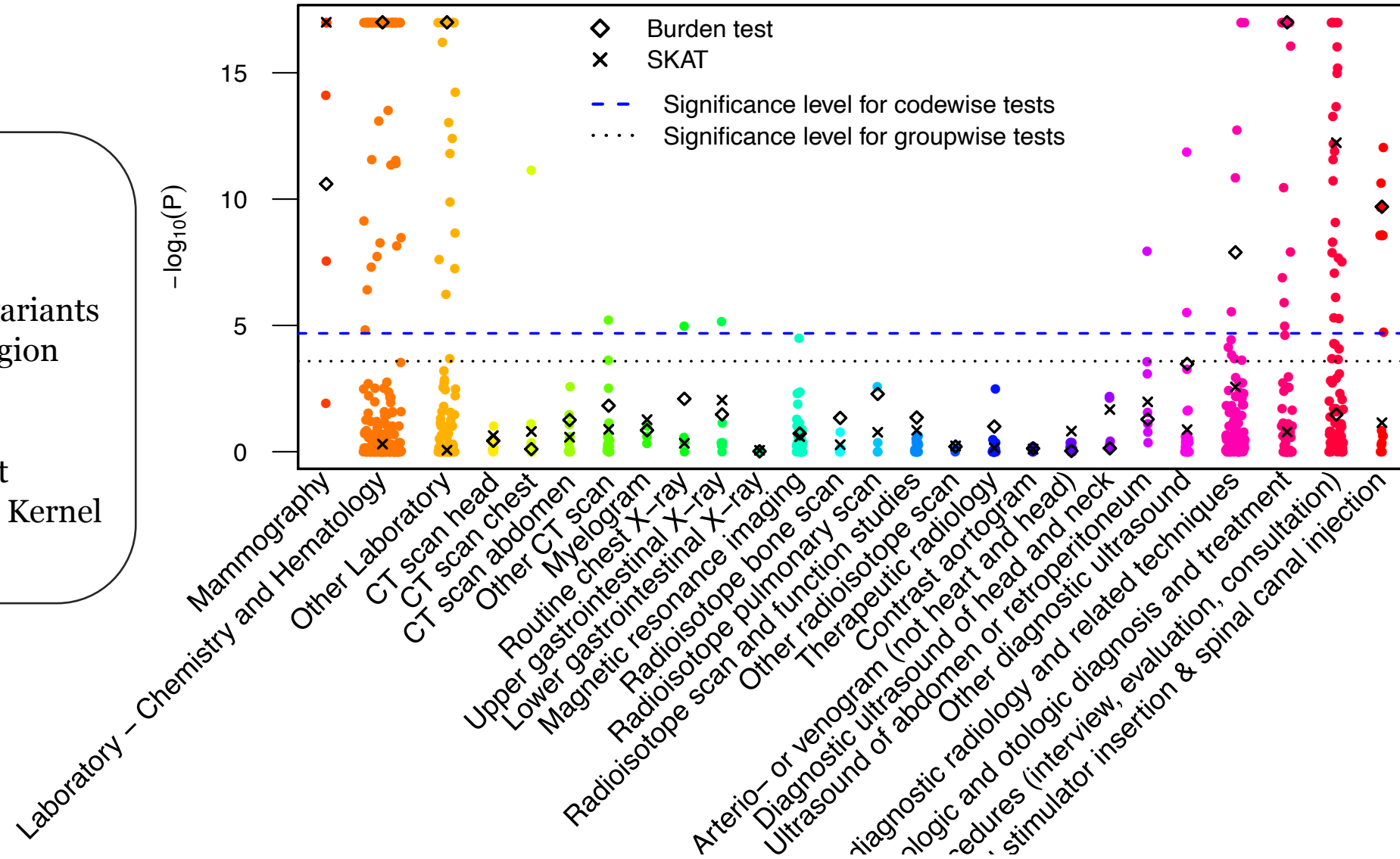
Aim 1: Compare Coding Patterns Between KPWA & KPNW

Code-wise and group-wise testing:

Medical codes in a group ↔ Genetic variants in a region

Code-level: Two sample T-test

Group-level: SKAT (Sequence Kernel Association Test) & Burden test



Aim 1 Methods: Privacy-Preserving Code/Group-Level Tests

Derive summary data needed for T-test, SKAT, and Burden adjusting for covariates

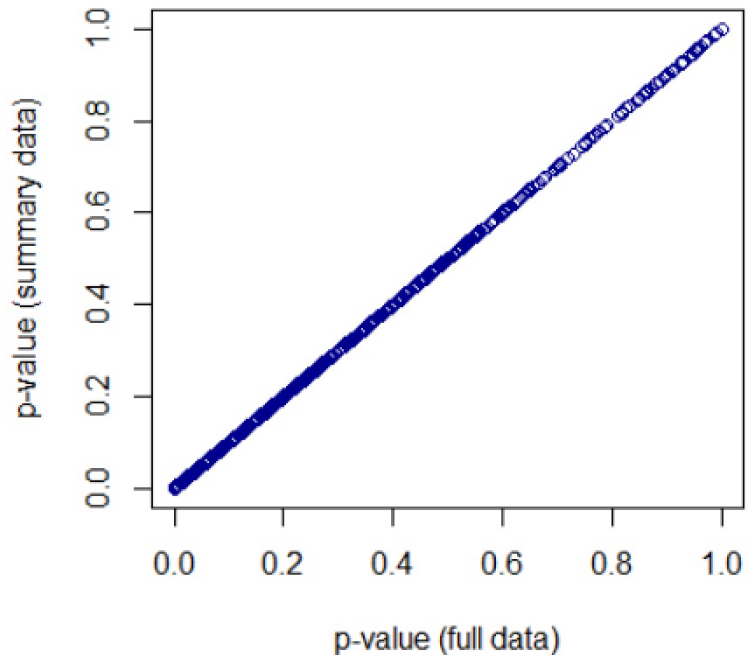
X_{ij} is the number of times that patient i got code j recorded in the specified year

Site	Year	Gender	Age group	# of patients	$\sum_i \mathbf{1}(X_{ij} > 0)$	$\sum_i X_{ij}$	$\sum_i (X_{ij})^2$	$\sum_i \mathbf{1}(X_{ij} X_{ij'} > 0)^\dagger$	$\sum_i X_{ij} X_{ij'}^\dagger$
0	2012	F	1	6000	173	562	14576	43	82
0	2012	F	2	3000
0	2013	F	1	...					
0	2013	F	2						
0	2012	M	1						
0	2012	M	2						
0	2013	M	1						
0	2013	M	2						
1	2012	F	1						
...						

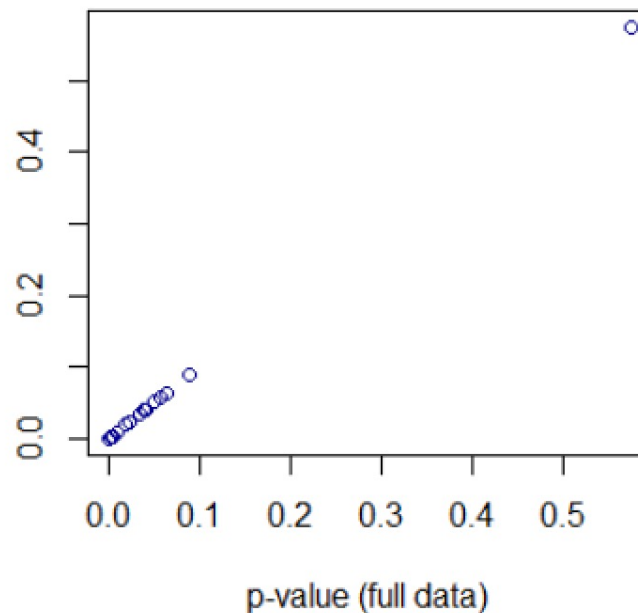
Aim 1 Methods: Privacy-Preserving Code/Group-Level Tests

Validation results (real medical records are used in the experiments below)

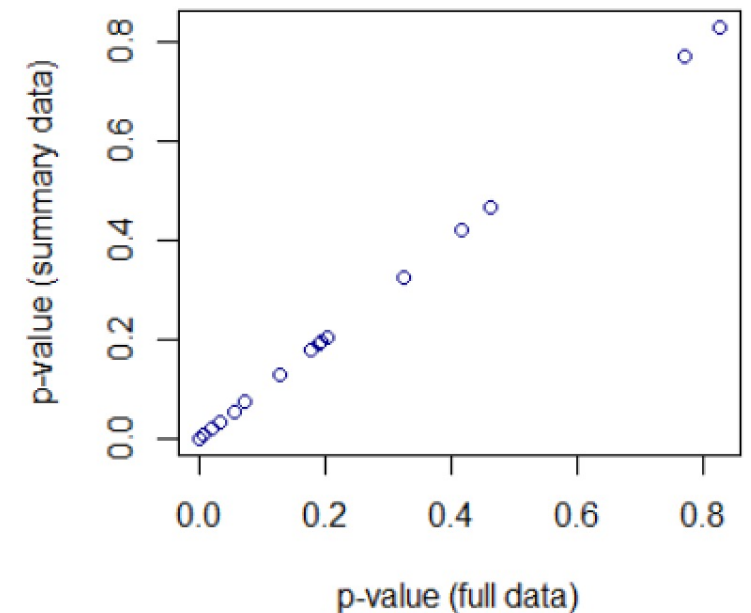
T-test for 1023 codes



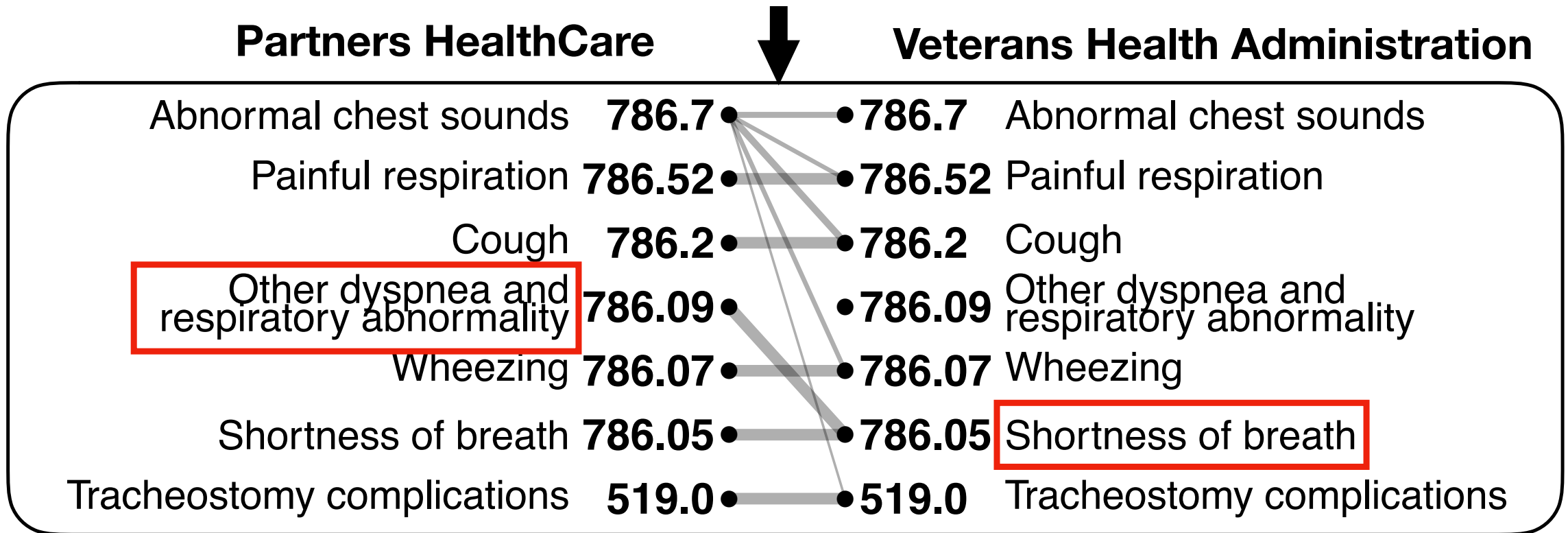
SKAT test for 30 code groups



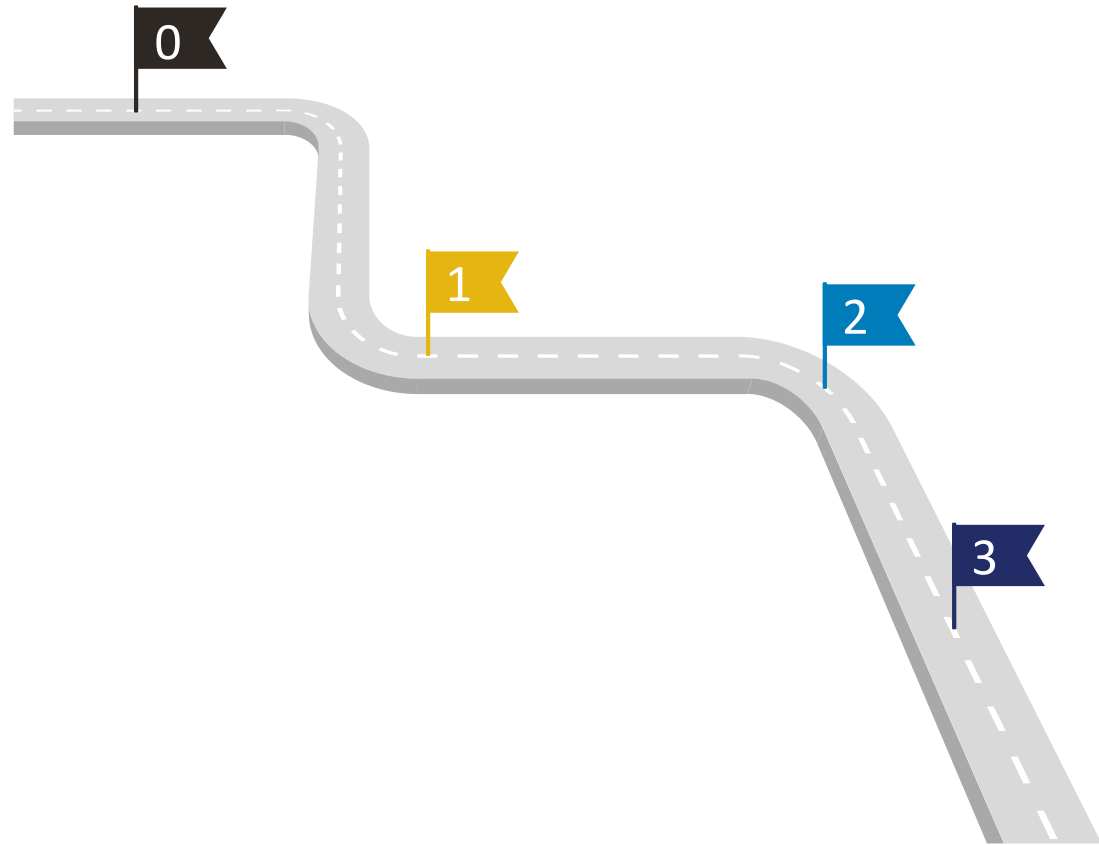
Burden test for 30 code groups



Aim 2: Automated Code Mapping Between KPWA & KPNW



Aim 2 Methods: Automated Code Mapping with Embeddings



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

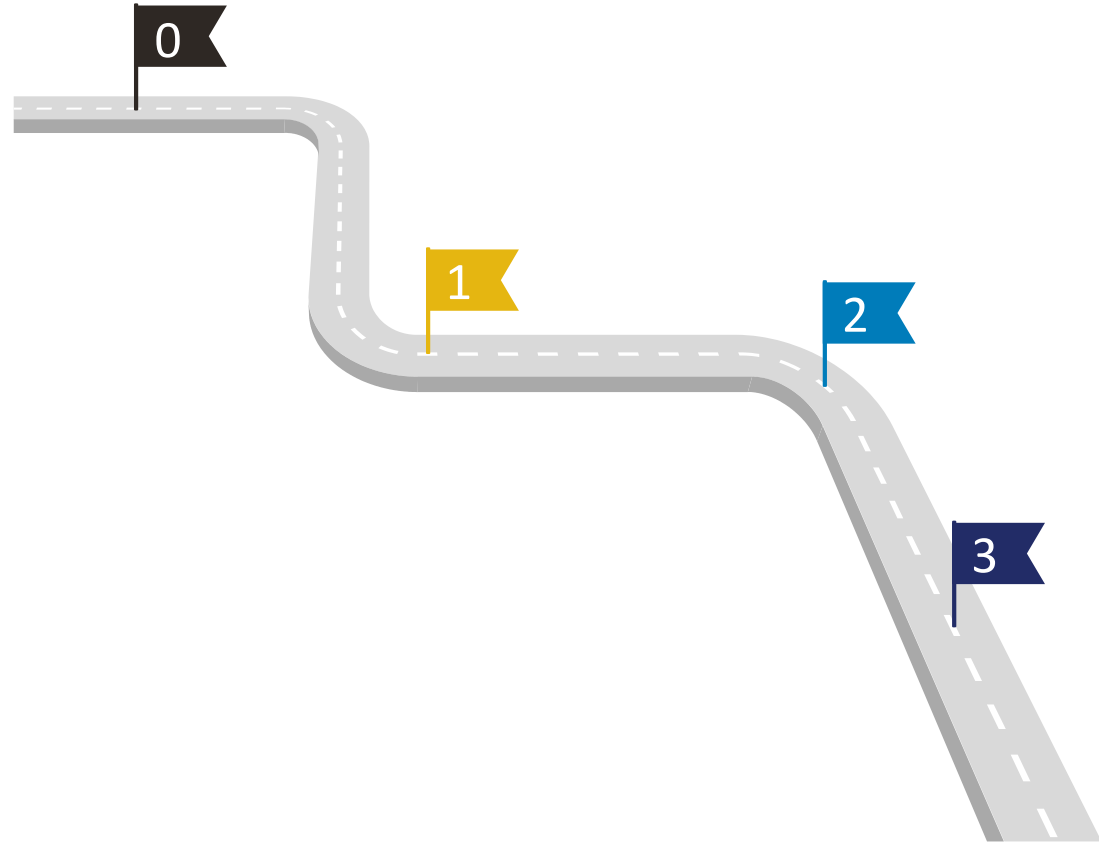
Step 2: space alignment

align two embedding spaces so we can measure distance

Step 3: code mapping

for a source code, find nearest neighbor(s) among target codes

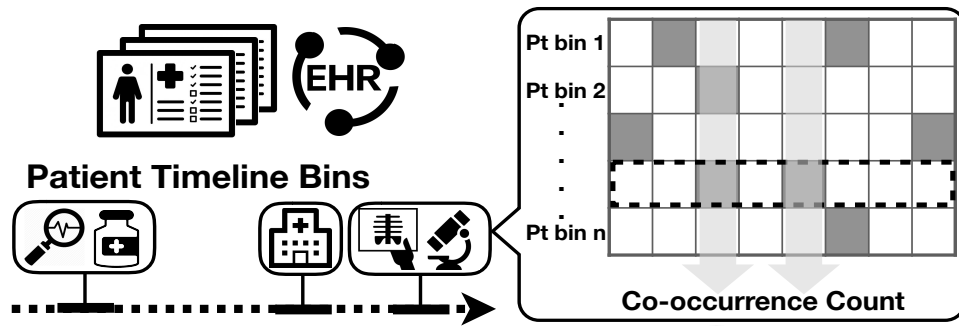
Aim 2 Methods: Automated Code Mapping with Embeddings



Step 1: code embedding

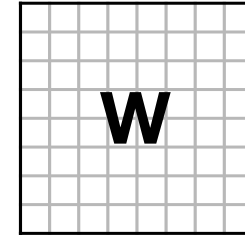
within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Aim 2 Methods: Code Embedding, Param Tuning, and Cosine Similarity

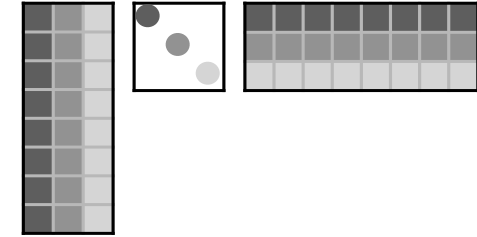


Parameter Tuning:
Selected **time window** = 1 day

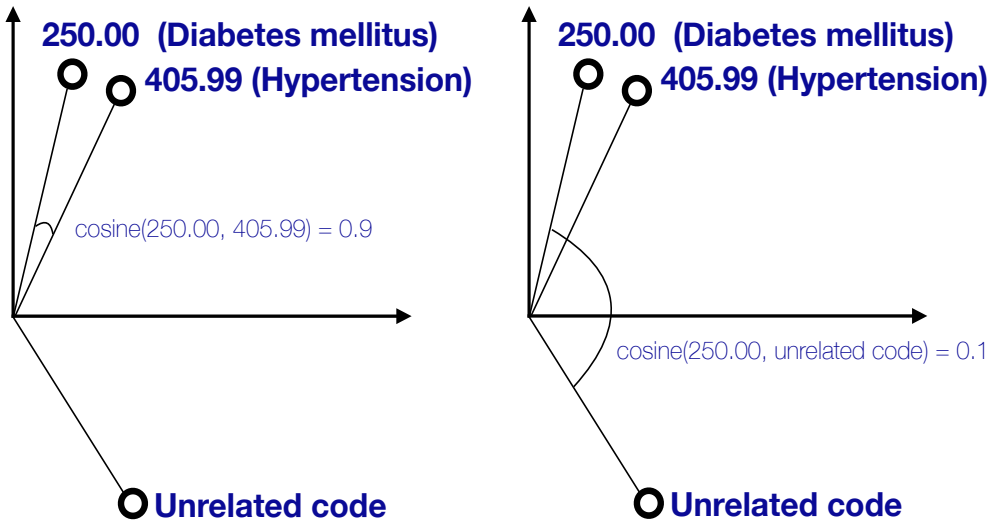
Pointwise Mutual Information Matrix



Concept Embeddings from Singular Value Decomposition

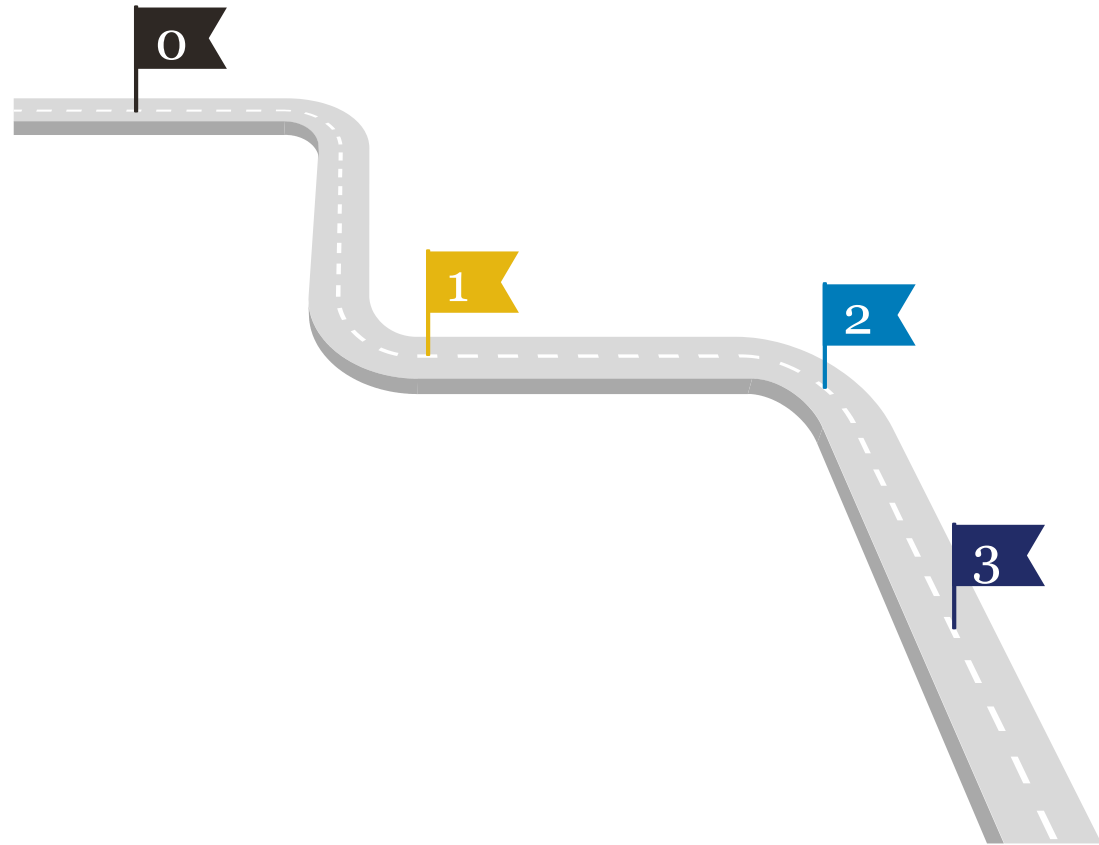


Parameter Tuning:
selected **embedding dimension** = 50



Cosine similarity ranges between $[-1, 1]$
Higher cosine = closer in embedding space = two codes are more related

Aim 2 Methods: Automated Code Mapping with Embeddings



Step 2: space alignment

align two embedding spaces so we can measure distance

Step 3: code mapping

for a source code, find nearest neighbor(s) among target codes

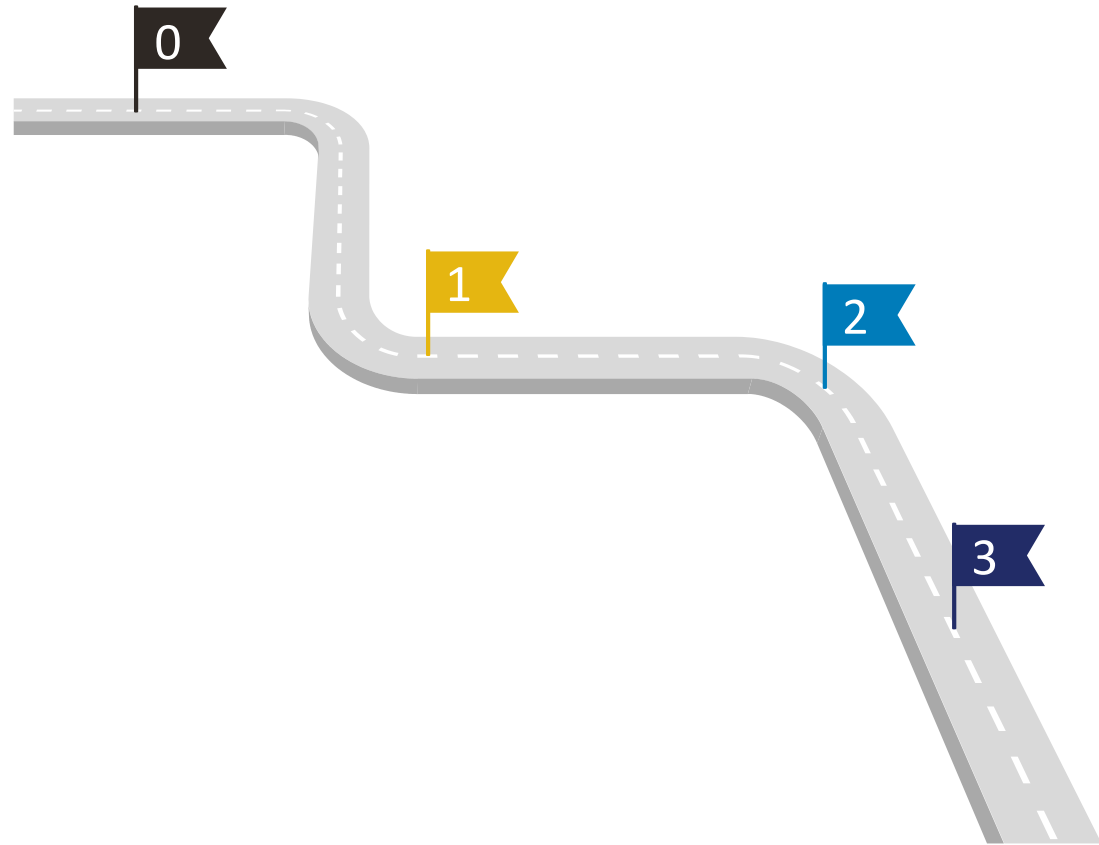
Aim 2 Methods: Automated Code Mapping with Embeddings

	PADS Method	Main Method – RADS	RARS Method
Step 0 & 1	Data preparation & Code embedding		
Step 2: space alignment (use all codes)	<u>Projection-based Alignment</u> : linear regression to project the embeddings from one system to another	<u>Rotation-based Alignment</u> : spherical regression to rotate the embeddings from one system to another	
Step 3: code mapping (within code group)	Find the largest <u>Directional-Similarities</u> (<i>unadjusted</i> association between a pair of codes)		Find the largest <u>Regression-Similarities</u> (<i>adjusted</i> association between a pair of codes)

RADS VS RARS: “Unadjusted” VS “Adjusted” Association

- Y and X are the (row-normalized) code embeddings (after space alignment) for System 1 and 2, respectively
- Mapping from System 1 (Y) to System 2 (X)
 - **RADS**: directional similarity – YX^T
 - **RARS**: regression similarity – $YX^T (XX^T)^{-1}$
- **RADS** vs **RARS** – “unadjusted” vs “adjusted”
- Note: use $YX^T (XX^T + \lambda I)^{-1}$ in the finalized RARS method

Main Method: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Step 2: space alignment

Rotation-based Alignment of all codes

Step 3: code mapping

for a source code, find nearest neighbor(s) among target codes; “nearest” defined by **largest Directional Similarity**

1. calculate the cosine-similarity matrix and then update the matrix by incorporating code frequency
2. cross-validated thresholding and additionally mark top 1-2

Main Method: Incorporate Code Frequency to Improve Mapping Matrix

- Idea: Fine tune the estimated mapping matrix $\hat{\Pi}$ such that it matches code marginal frequency between sites
 - $\hat{\Gamma} = \operatorname{argmin}_{Y=\Gamma X, \Gamma 1_n=1_n} \|\hat{\Pi} - \Gamma\|_F$, where Y and X are code frequencies in the two sites

CODE	Description	Y		X	
		freq_KPWA	freq_KPNW	frequency ratio	pValue
E08.36	Diabetes mellitus due to underlying condition with diabetic cataract	23	0	3.1	6.12E-3
E10.36	Type 1 diabetes mellitus with diabetic cataract	92	117	0.75	6.06E-2
E11.36	Type 2 diabetes mellitus with diabetic cataract	3065	2996	0.96	5.88E-1
H26.40	Unspecified secondary cataract	561	1144	0.46	<1E-6
H26.411	Soemmering's ring, right eye	11	1	1.79	1.26E-1
H26.491	Other secondary cataract, right eye	3044	771	3.67	<1E-6
H26.492	Other secondary cataract, left eye	3129	741	3.93	<1E-6
H26.493	Other secondary cataract, The 'cataract' ICD-10 group (18 codes)bilateral	3952	636	5.76	<1E-6
H26.499	Other secondary cataract, unspecified eye	70	0	7.51	<1E-6
H26.8	Other specified cataract	526	1323	0.38	<1E-6
H26.9	Unspecified cataract	16704	15786	0.99	8.53E-1
H59.021	Cataract (lens) fragments in eye following cataract surgery, right eye	47	14	2.23	1.31E-1
H59.022	Cataract (lens) fragments in eye following cataract surgery, left eye	78	10	4.13	1.03E-2
H59.029	Cataract (lens) fragments in eye following cataract surgery, unspecified eye	1	72	0.13	1.15E-6
Z96.1	Presence of intraocular lens	35888	44526	0.76	<1E-6
Z98.41	Cataract extraction status, right eye	3950	199	17.79	<1E-6
Z98.42	Cataract extraction status, left eye	3723	195	17.1	<1E-6
Z98.49	Cataract extraction status, unspecified eye	622	112	4.87	<1E-6



Results

Results: Study Population

Study population

Continuously enrolled KPWA and KPNW members aged 50+ with any diabetes

** 10-15% of population has diabetes*

Study period

2011-Jan-01 to 2020-Dec-31.

** 10 years spanning the ICD-9 and ICD-10 era*

The total number of unique codes is 65935, including 11950 ICD-9 codes, 36538 ICD-10 codes, 7749 CPT codes

	# patients	# Total code endorsements	# ICD-9 endorsements	# ICD-10 endorsements	# CPT endorsements	# unique codes
KPWA	87,178	57,938,353	9,381,687	17,223,847	20,290,686	53,004
KPNW	71,535	55,876,254	7,427,793	14,847,178	19,188,234	49,761

Results: Study Population

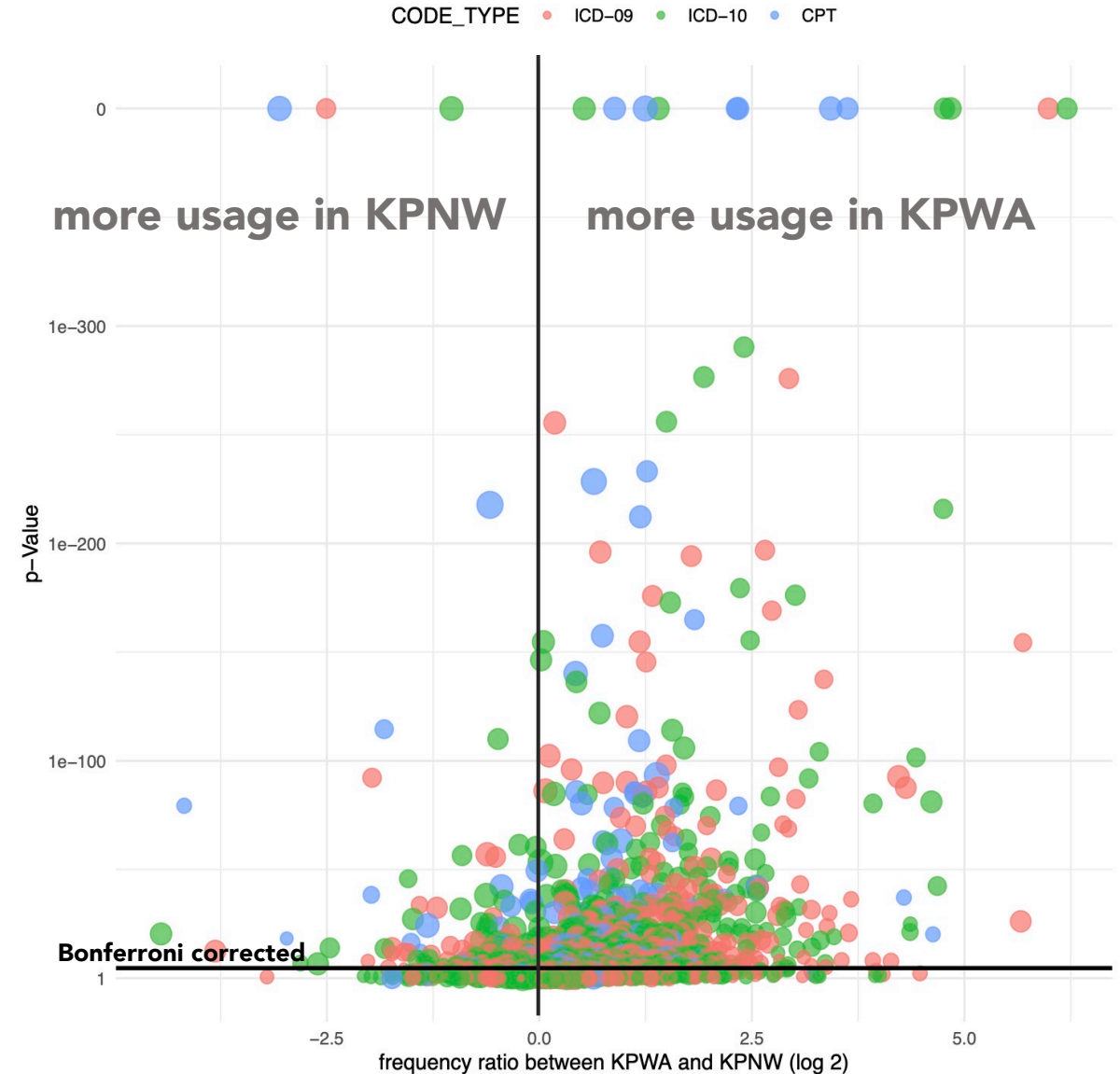
	KPWA	KPNW	Total
	(N=74475)	(N=64231)	(N=138706)
AGE			
Mean (SD)	62.8 (9.95)	62.8 (9.91)	62.8 (9.93)
Median [Min, Max]	61.0 [49.0, 102]	62.0 [49.0, 102]	62.0 [49.0, 102]
SEX			
Male	37844 (50.8%)	32770 (51.0%)	70614 (50.9%)
Female	36631 (49.2%)	31461 (49.0%)	68092 (49.1%)
INSULIN			
No	57291 (76.9%)	52024 (81.0%)	109315 (78.8%)
Yes	17184 (23.1%)	12207 (19.0%)	29391 (21.2%)
COMORBIDITY			
Mean (SD)	3.59 (2.34)	3.52 (2.28)	3.56 (2.31)
Median [Min, Max]	3.00 [0, 19.0]	3.00 [0, 17.0]	3.00 [0, 19.0]
Missing	339 (0.5%)	36 (0.1%)	375 (0.3%)

	KPWA	KPNW	Total
RACE			
Unknown	20570 (27.6%)	4168 (6.5%)	24738 (17.8%)
American Indian or Alaska Native	1300 (1.7%)	925 (1.4%)	2225 (1.6%)
Asian	5776 (7.8%)	3661 (5.7%)	9437 (6.8%)
Black or African American	3328 (4.5%)	2495 (3.9%)	5823 (4.2%)
Native Hawaiian or Other Pacific Islander	773 (1.0%)	855 (1.3%)	1628 (1.2%)
White	42728 (57.4%)	52127 (81.2%)	94855 (68.4%)
HbA1c			
Mean (SD)	7.22 (1.51)	7.17 (1.43)	7.19 (1.46)
Median [Min, Max]	6.80 [3.90, 18.6]	6.70 [4.30, 18.0]	6.80 [3.90, 18.6]
Missing	31151 (41.8%)	2916 (4.5%)	34067 (24.6%)
HOSP COUNT			
Mean (SD)	0.190 (0.598)	0.198 (0.622)	0.194 (0.609)
Median [Min, Max]	0 [0, 13.0]	0 [0, 15.0]	0 [0, 15.0]

Results: Group-Level Differences Between KPWA & KPNW

Findings from group-level comparisons:

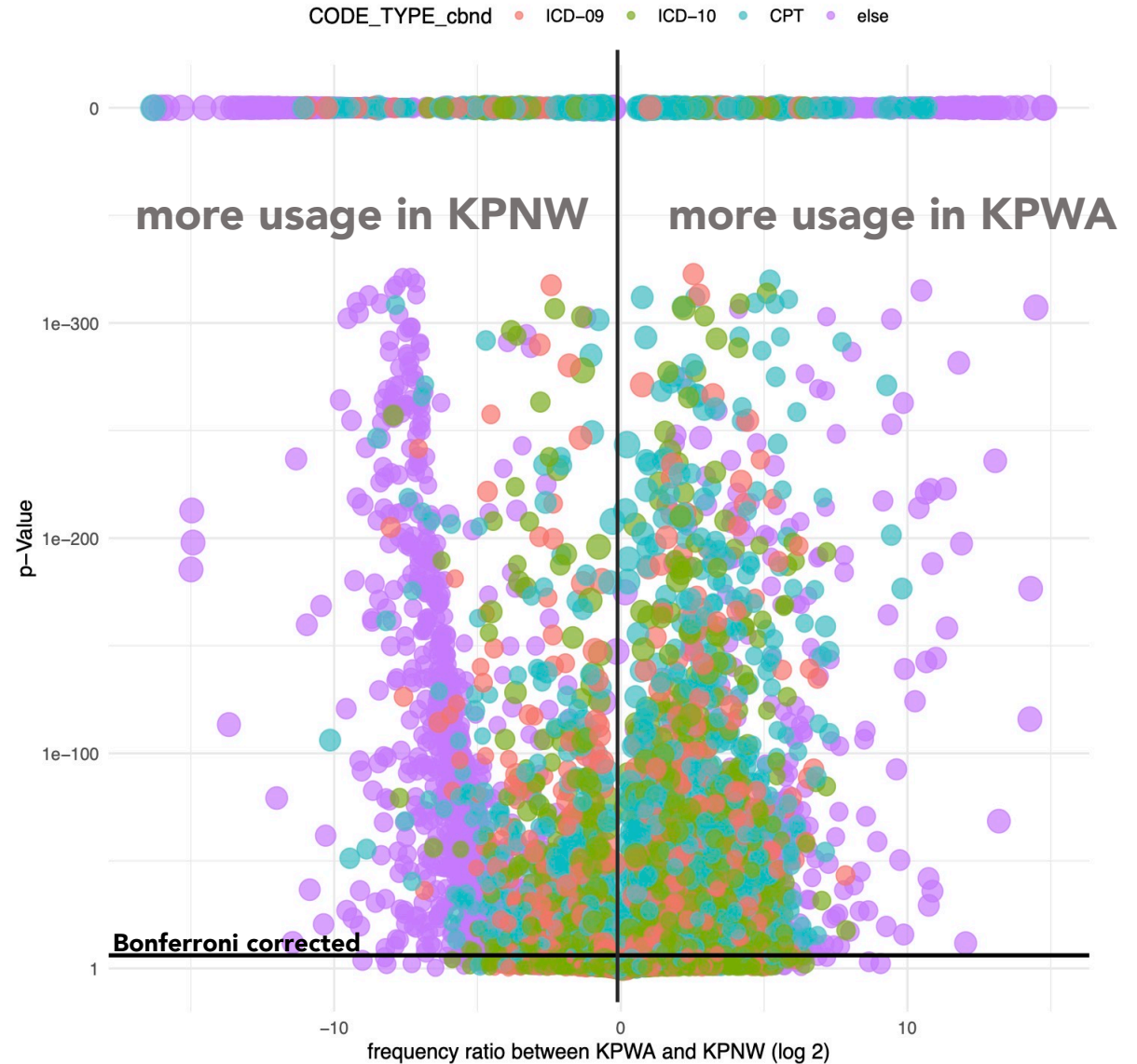
- 33% (828 out of 2523) code groups have meaningful differences between sites
- Considerable number of code groups with large magnitude of differences
 - 119 groups (4.7%) with freq ratio > 5
 - 10 groups (0.4%) with freq ratio < 1/5



Results: Code-Level Differences Between KPWA & KPNW

Findings from code-level comparisons:

- 13% (8348 out of 65935) codes have meaningful differences between sites
- Considerable number of codes with large magnitude of differences
 - 4086 codes (6.2%) with freq ratio > 5
 - 2744 codes (4.2%) with freq ratio < 1/5
- KPNW has many local codes



Results: Comparison of Coding Between Years (ICD-10 Codes)

Sentinel conducts routine data quality check with a lot of manual efforts. It is challenging to identify all abnormal changes in coding manually.

	KPWA		KPNW	
	SKAT	Burden test	SKAT	Burden test
2019 vs 2020	39 out of 1151 code groups (3.39%) have p-values <0.05/1151	36 out of 1151 code groups (3.13%) have p-values <0.05/1151	48 out of 1114 code groups (4.31%) have p-values <0.05/1114	40 out of 1114 code groups (3.59%) have p-values <0.05/1114
2018 vs 2019	24 out of 1161 code groups (2.07%) have p-values <0.05/1161	11 out of 1161 code groups (0.95%) have p-values <0.05/1161	29 out of 1108 code groups (2.62%) have p-values <0.05/1108	22 out of 1108 code groups (1.99%) have p-values <0.05/1108
2017 vs 2018	23 out of 1161 code groups (2.00%) have p-values <0.05/1161	9 out of 1161 code groups (0.78%) have p-values <0.05/1161	30 out of 1116 code groups (2.69%) have p-values <0.05/1116	18 out of 1116 code groups (1.61%) have p-values <0.05/1116
2016 vs 2017	49 out of 1157 code groups (4.24%) have p-values <0.05/1157	21 out of 1157 code groups (1.82%) have p-values <0.05/1157	46 out of 1109 code groups (4.15%) have p-values <0.05/1109	21 out of 1109 code groups (1.89%) have p-values <0.05/1109

Results: The Most Significant Code Groups Based on SKAT

SKAT detects group-wise association even if within-group differences are of different directions

CODE_TYPE	groupLabel	Description	freq KPWA	freq KPNW	freq ratio	pValue
ICD-9	216.1	Screening for malignant neoplasms of the skin	1408	7568	0.18	<1E-320
ICD-9	367.2	Astigmatism	23768	342	63.44	<1E-320
ICD-9	483	Acute bronchitis and bronchiolitis	10633	1297	7.65	<1E-320
ICD-10	208	Benign neoplasm of colon	76167	27101	2.64	<1E-320
ICD-10	216	Benign neoplasm of skin	16027	5333	2.82	<1E-320
ICD-10	216.1	Screening for malignant neoplasms of the skin	144276	277531	0.49	<1E-320
ICD-10	365.1	Open-angle glaucoma	33140	19056	1.63	<1E-320
ICD-10	366	Cataract	75535	68658	1.03	<1E-320
ICD-10	366.2	Senile cataract	80753	52593	1.44	<1E-320
ICD-10	367.1	Myopia	23479	801	27.20	<1E-320
ICD-10	367.2	Astigmatism	46150	1505	28.62	<1E-320
ICD-10	367.8	Hypermetropia	24280	299	73.83	<1E-320
ICD-10	427.6	Premature beats	14820	2617	5.30	<1E-320
ICD-10	733	Other disorders of bone and cartilage	20008	4904	3.83	<1E-320
CPT	193	Diagnostic ultrasound of heart (echocardiogram)	78787	40053	1.85	<1E-320
CPT	197	Other diagnostic ultrasound	78220	43947	1.67	<1E-320

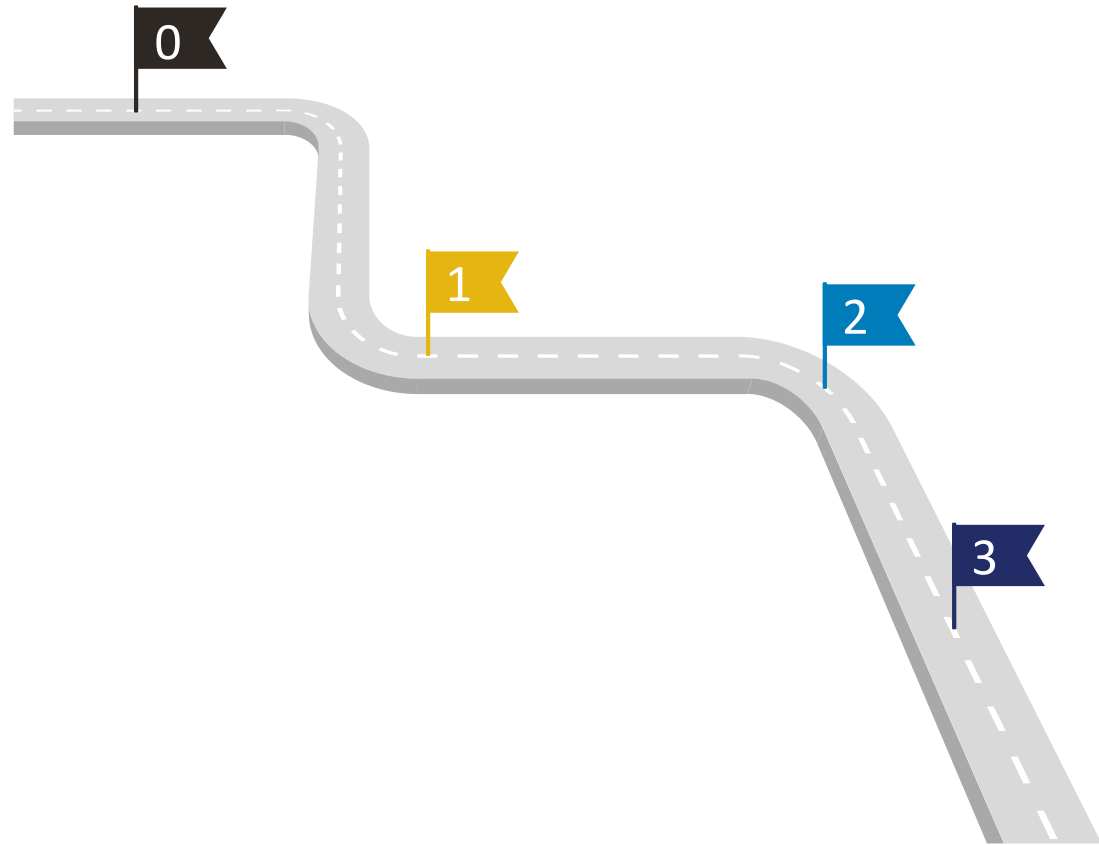
* freq ratio = $\frac{(freq\ KPWA+10)/number\ of\ patient\ years\ in\ KPWA}{(freq\ KPNW+10)/number\ of\ patient\ years\ in\ KPNW}$

Results: Within the "Cataract" ICD-10 Group (18 Codes)

CODE	Description	freq_KPWA	freq_KPNW	frequency ratio	pValue
E08.36	Diabetes mellitus due to underlying condition with diabetic cataract	23	0	3.1	6.12E-3
E10.36	Type 1 diabetes mellitus with diabetic cataract	92	117	0.75	6.06E-2
E11.36	Type 2 diabetes mellitus with diabetic cataract	3065	2996	0.96	5.88E-1
H26.40	Unspecified secondary cataract	561	1144	0.46	<1E-6
H26.411	Soemmering's ring, right eye	11	1	1.79	1.26E-1
H26.491	Other secondary cataract, right eye	3044	771	3.67	<1E-6
H26.492	Other secondary cataract, left eye	3129	741	3.93	<1E-6
H26.493	Other secondary cataract, The 'cataract' ICD-10 group (18 codes) bilateral	3952	636	5.76	<1E-6
H26.499	Other secondary cataract, unspecified eye	70	0	7.51	<1E-6
H26.8	Other specified cataract	526	1323	0.38	<1E-6
H26.9	Unspecified cataract	16704	15786	0.99	8.53E-1
H59.021	Cataract (lens) fragments in eye following cataract surgery, right eye	47	14	2.23	1.31E-1
H59.022	Cataract (lens) fragments in eye following cataract surgery, left eye	78	10	4.13	1.03E-2
H59.029	Cataract (lens) fragments in eye following cataract surgery, unspecified eye	1	72	0.13	1.15E-6
Z96.1	Presence of intraocular lens	35888	44526	0.76	<1E-6
Z98.41	Cataract extraction status, right eye	3950	199	17.79	<1E-6
Z98.42	Cataract extraction status, left eye	3723	195	17.1	<1E-6
Z98.49	Cataract extraction status, unspecified eye	622	112	4.87	<1E-6

* freq ratio = $\frac{(freq\ KPWA+10)/number\ of\ patient\ years\ in\ KPWA}{(freq\ KPNW+10)/number\ of\ patient\ years\ in\ KPNW}$

Recall: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Step 2: space alignment

Rotation-based Alignment of all codes

Step 3: code mapping

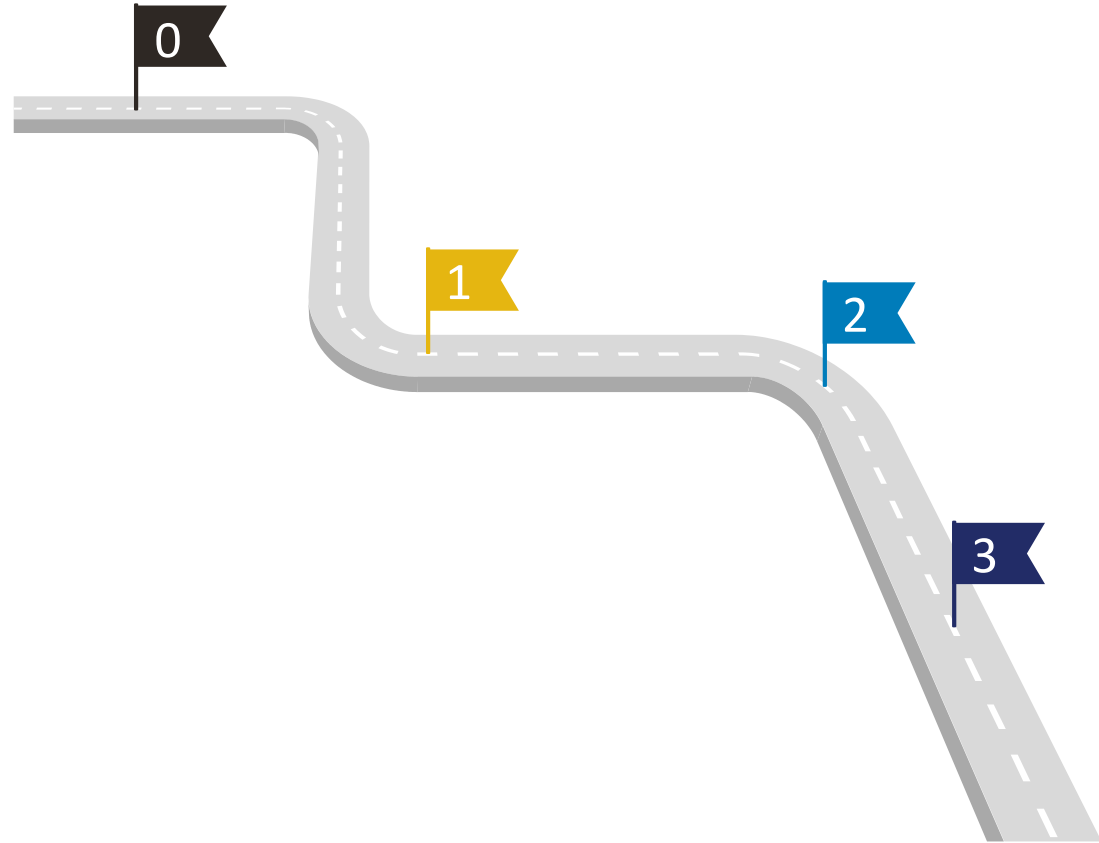
for a source code, find nearest neighbor(s) among target codes; “nearest” defined by **largest Directional Similarity**

1. calculate the cosine-similarity matrix and then update the matrix by incorporating code frequency
2. cross-validated thresholding and additionally mark top 1-2

Results: Rotation-based Alignment and Directional Similarity

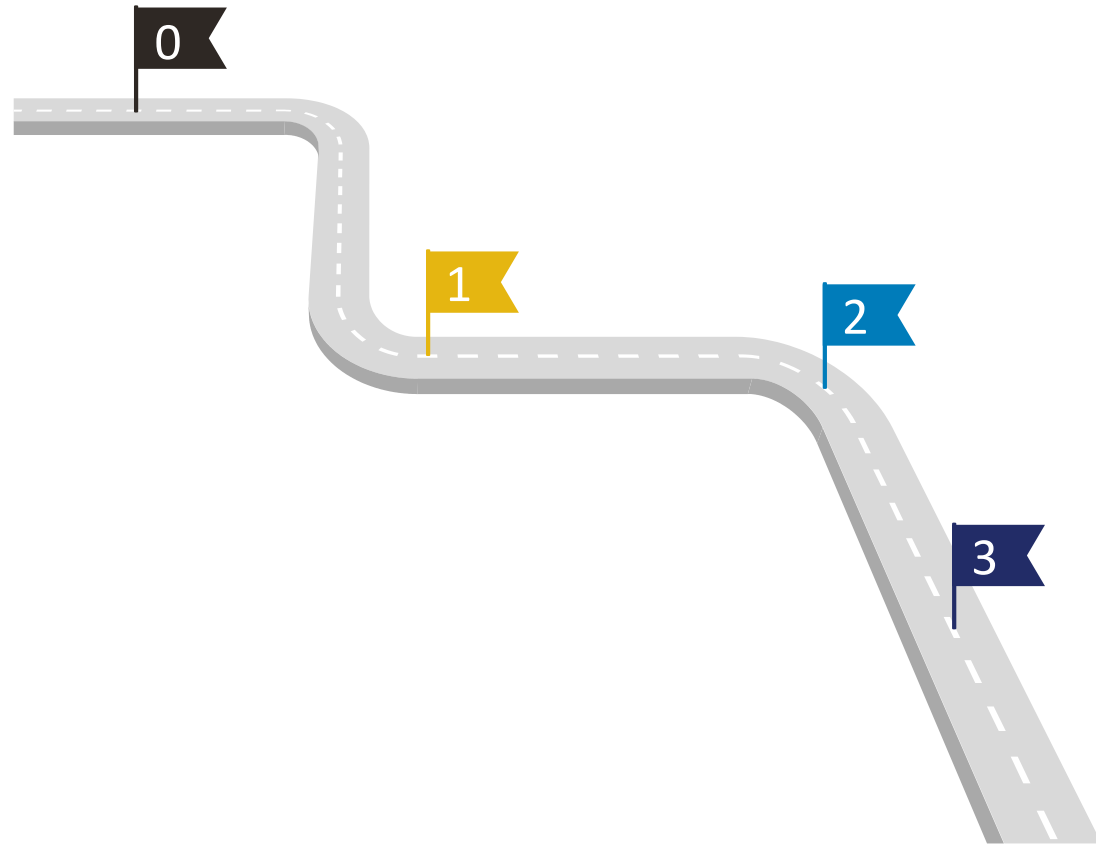
Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)



	KPWA	KPNW
Total number of codes	27,167	23,766
ICD-9	6,400	5,545
ICD-10	14,537	11,142
CPT	4,253	3,333

Results: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

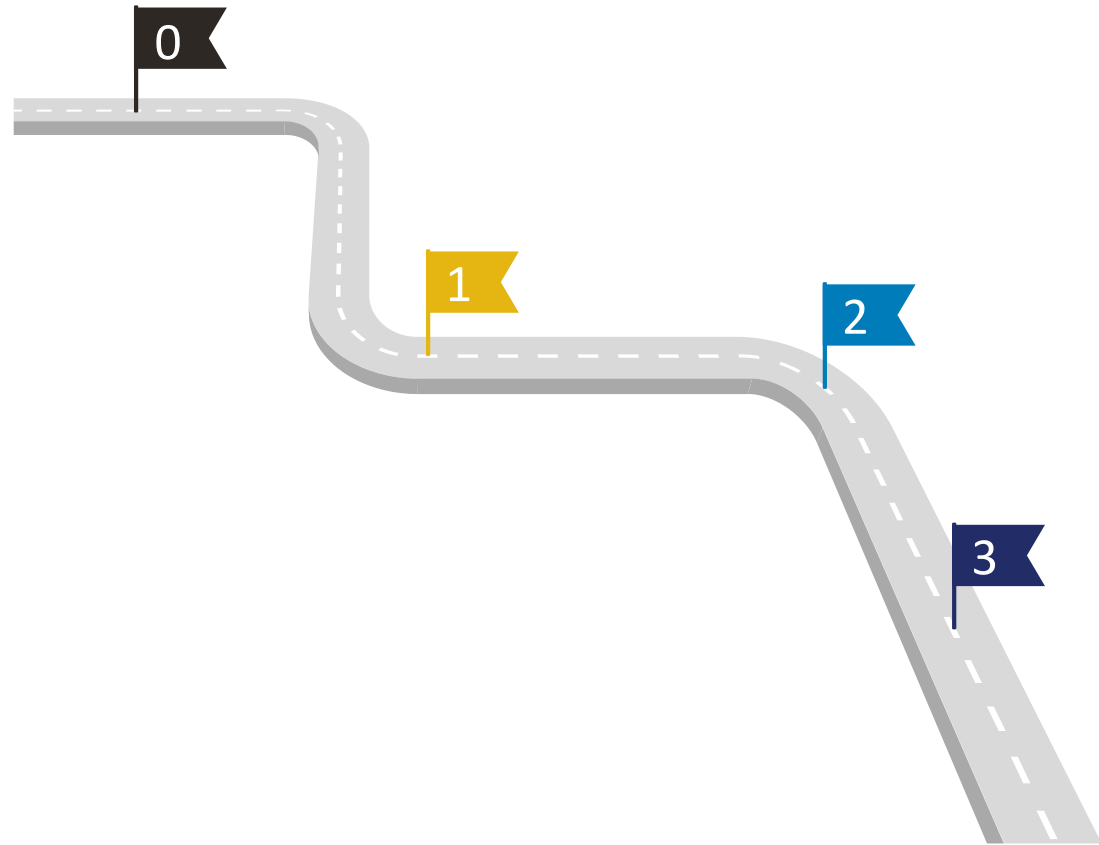
Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

	KPWA	KPNW
AUC*	0.805	0.796
Optimal time window	1 day	1 day
Optimal dimension	250	250

*Higher AUC indicates more agreement between code embedding-based clustering and human curated grouping

Results: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

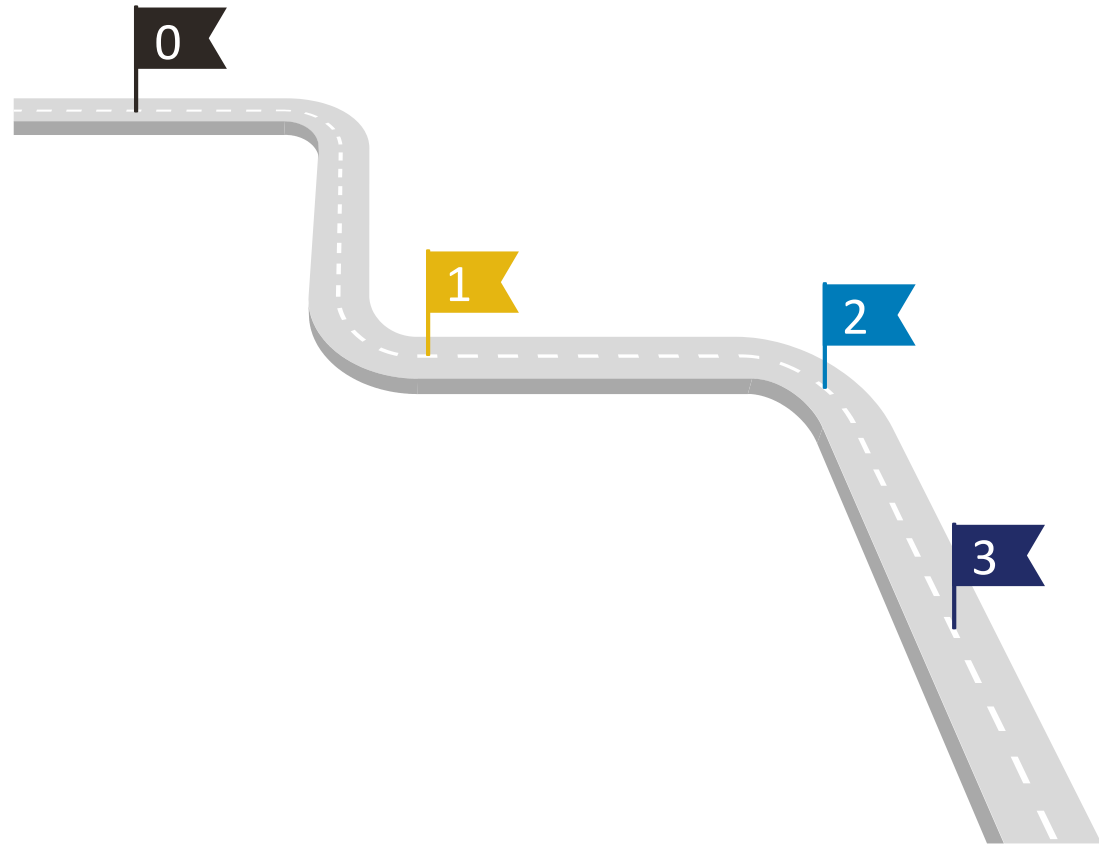
Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Step 2: space alignment

Directional similarity among all code pairs <u>w</u> ithin <u>g</u> roups	Before Alignment	After <u>R</u> otation-based <u>A</u> lignment	After <u>P</u> rojection-based <u>A</u> lignment
Mean (higher better)	0.046	0.270	0.336
Quantiles (25 th , median, 75 th)	-0.005 0.044 0.094	0.115 0.247 0.402	0.176 0.323 0.479

Results: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Step 2: space alignment

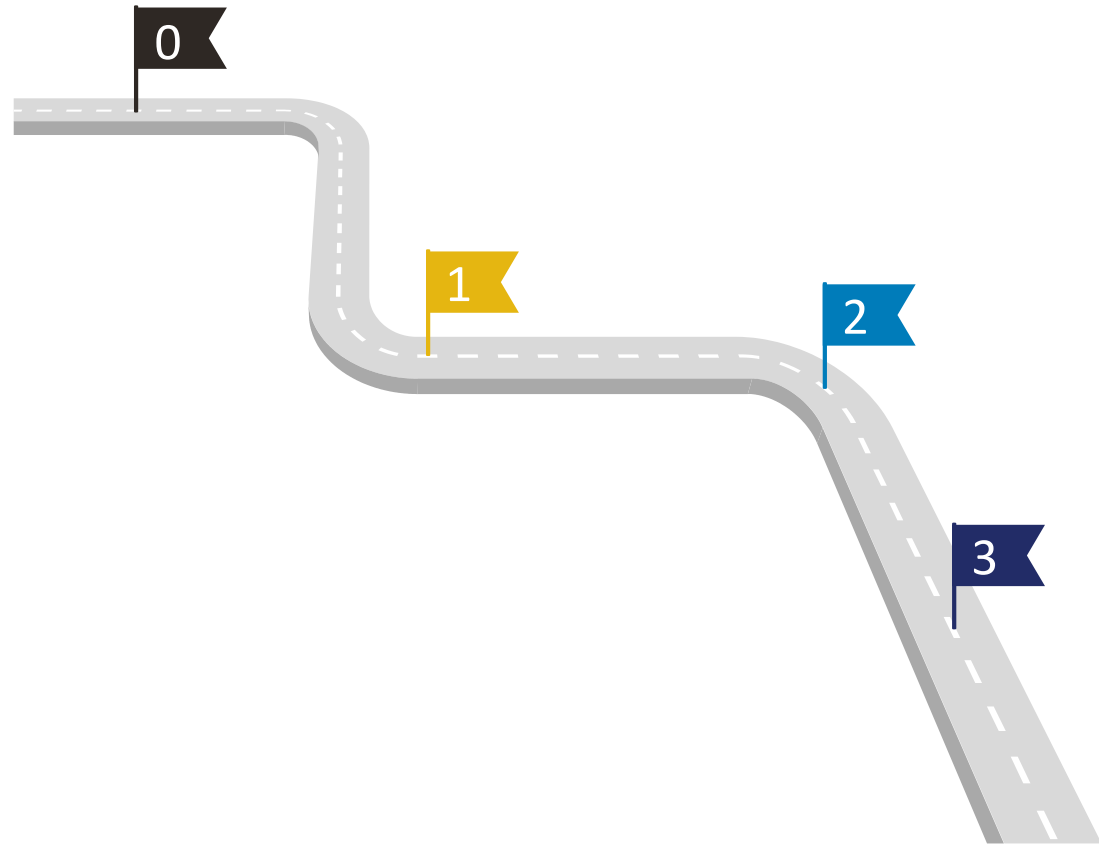
Rotation-based Alignment of all codes

Step 3: code mapping

for a source code, find nearest neighbor(s) among target codes; “nearest” defined by **largest Directional Similarity**

1. calculate the cosine-similarity matrix and then update the matrix by incorporating code frequency
2. cross-validated thresholding and additionally mark top 1-2

Results: Rotation-based Alignment and Directional Similarity



Step 0: data preparation

extract all codes, group up all rare codes (frequency <10)

Step 1: code embedding

within each KP site, compute code co-occurrences followed by dimension reduction to generate code embeddings

Step 2: space alignment

Rotation-based Alignment of all codes

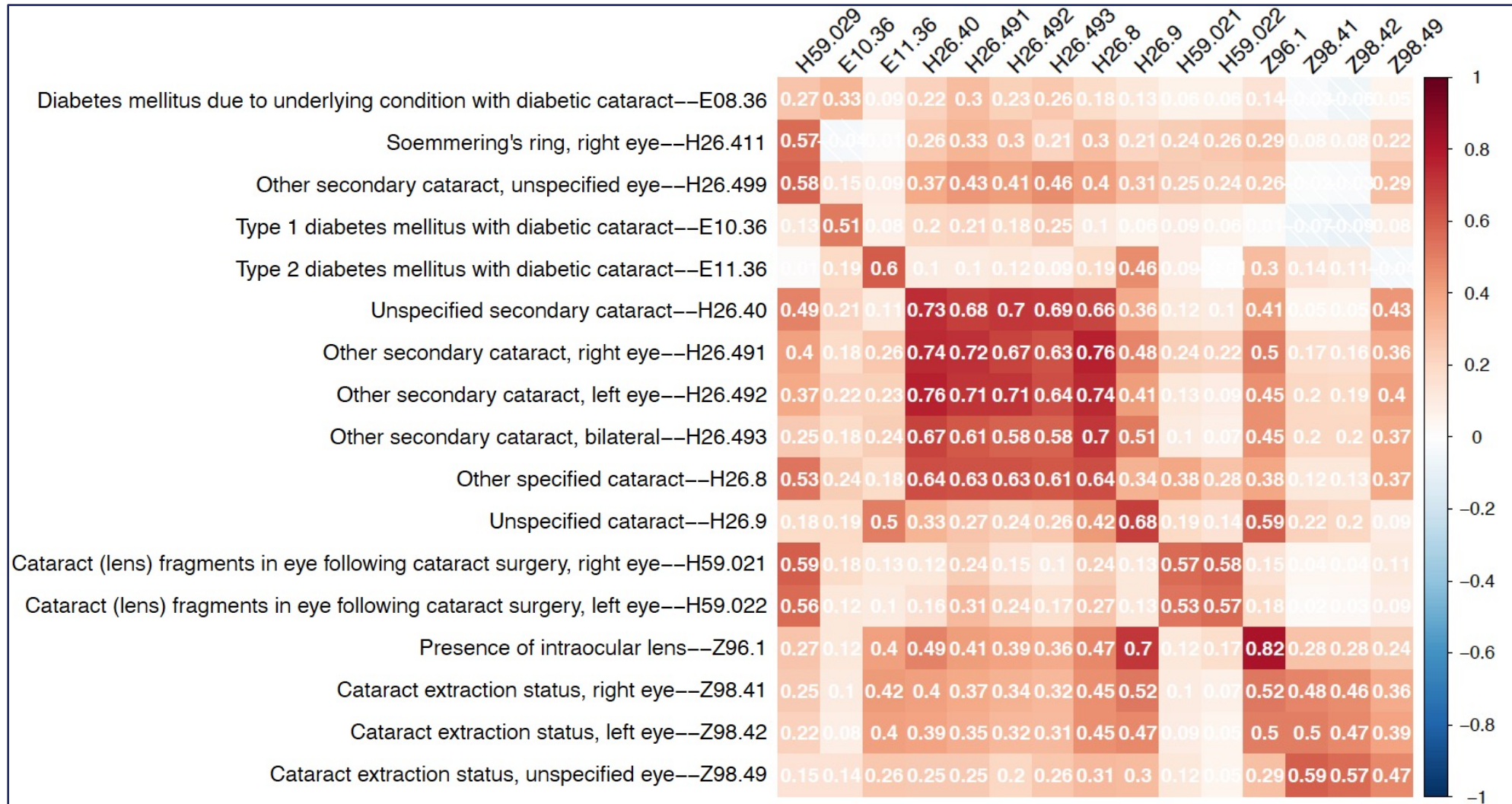
Step 3: code mapping

Cataract Group	KPWA	KPNW
Number of non-rare codes (freq > 10)	17	15

Results: Directional Similarity in the "Cataract" Group

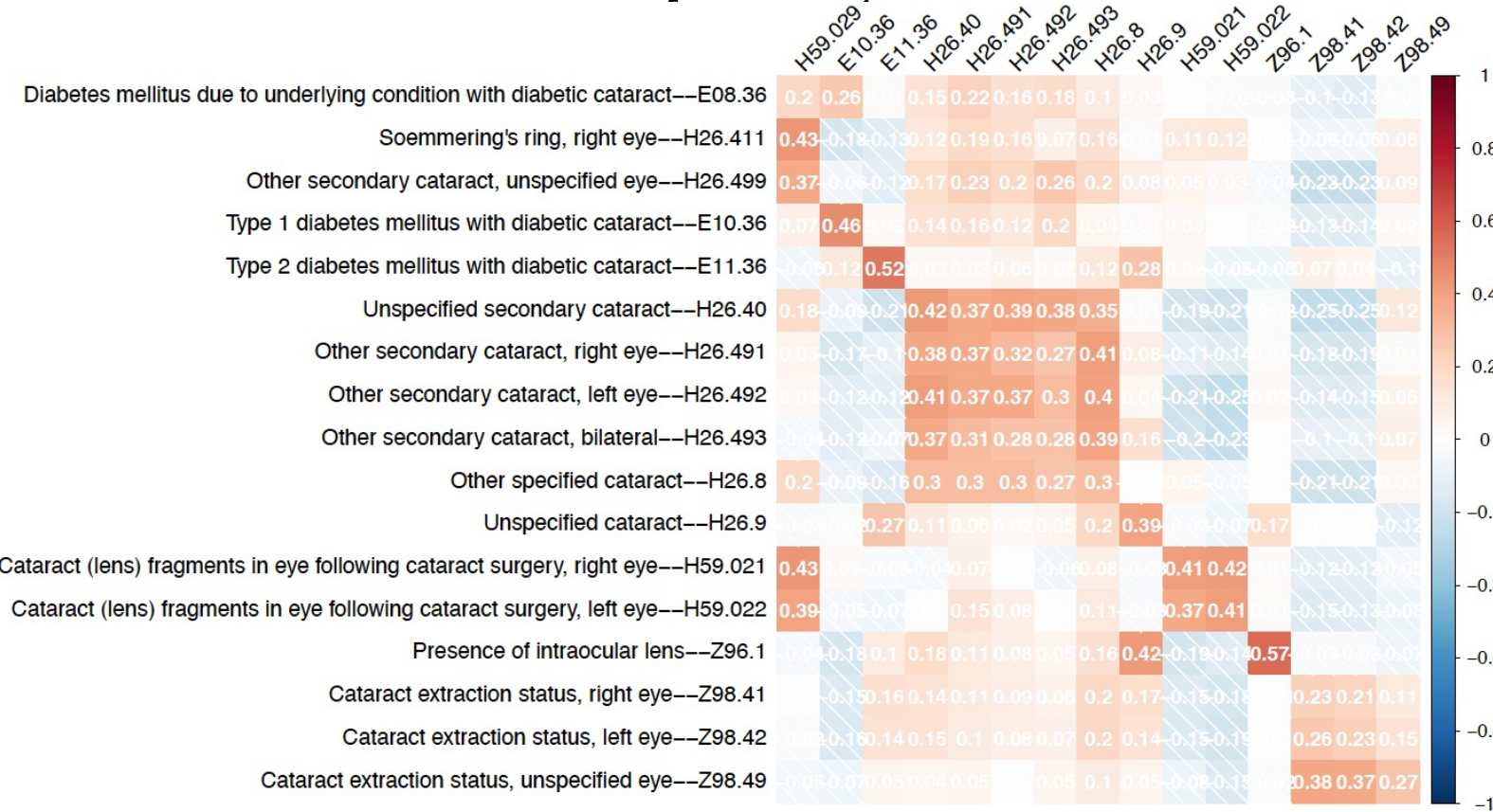
15 codes in KPNW

17 codes in KPWA

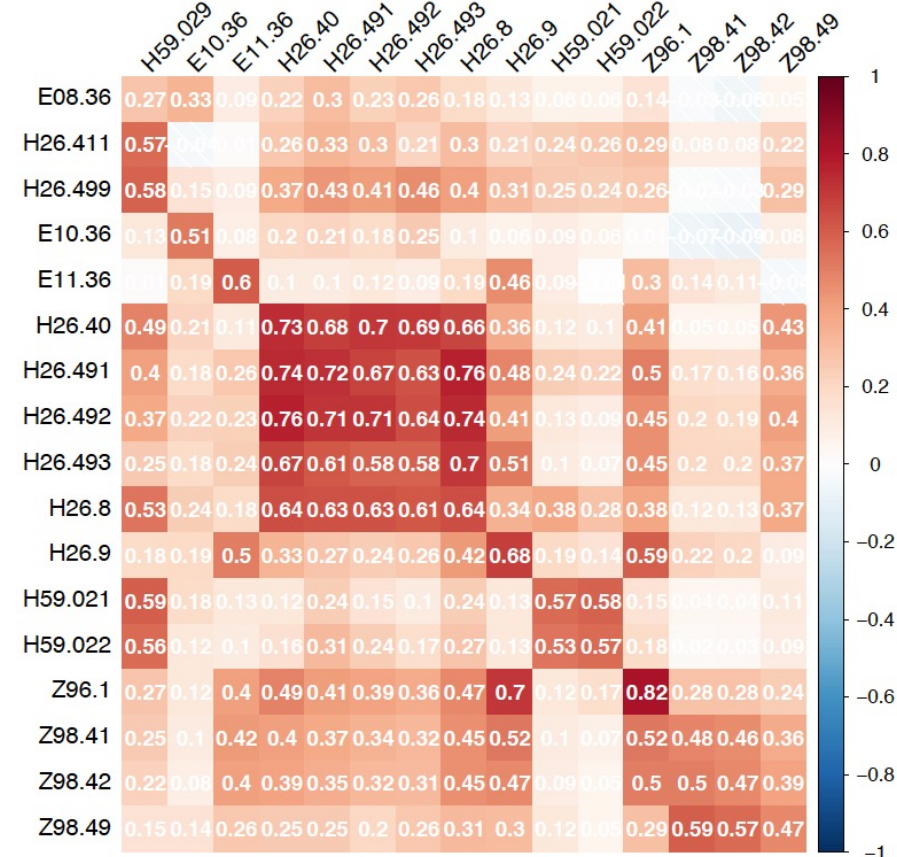


Results: Directional Similarity After Incorporating Code Frequency

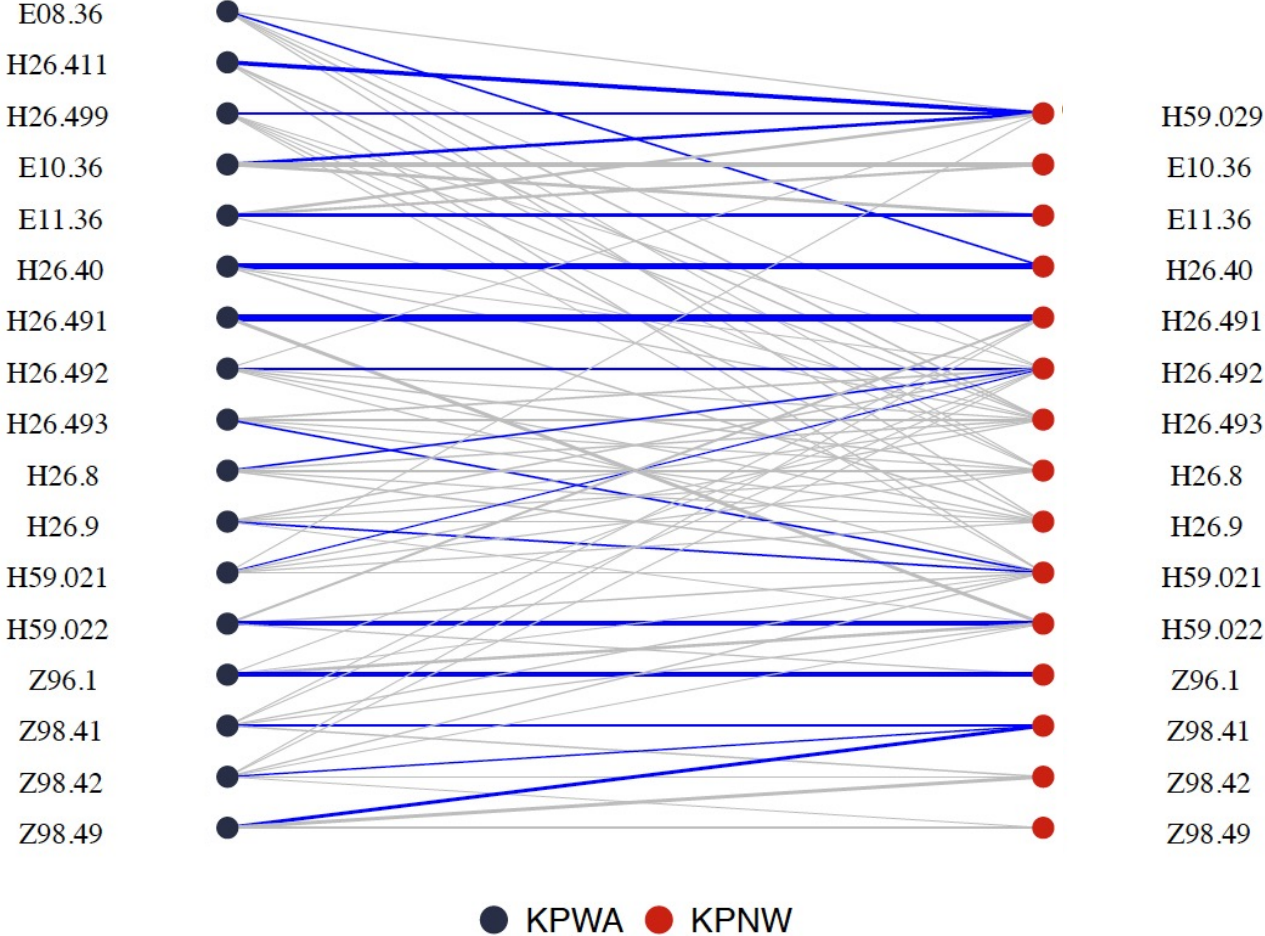
RADS - cosine similarity (after incorporating code)



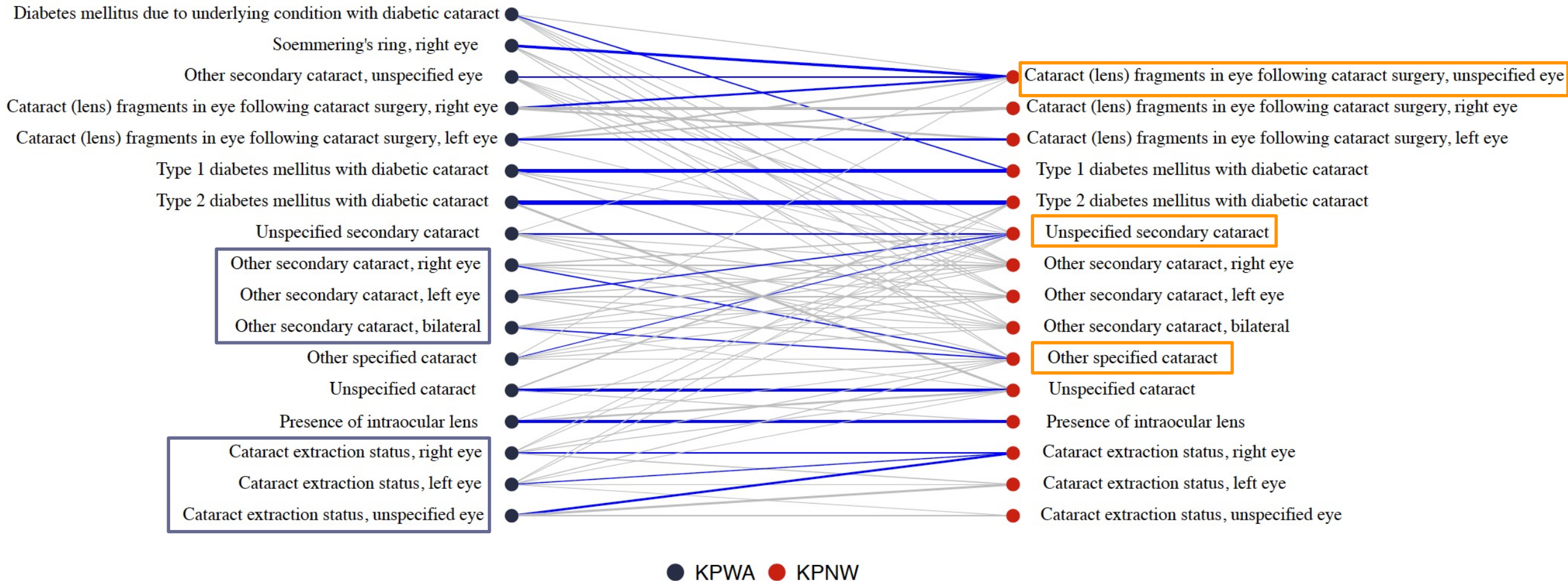
RADS - cosine similarity



Results: Mapping from KPWA (Left) to KPNW (Right)

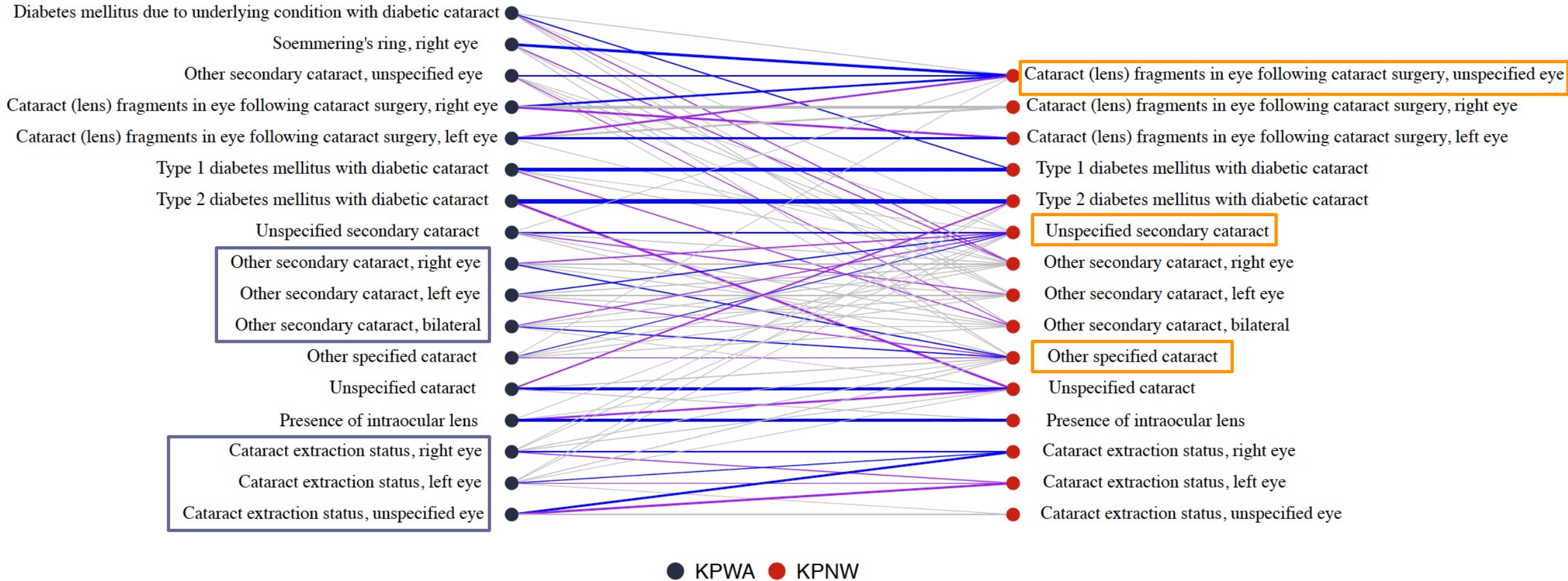


Results: Mapping from KPWA (Left) to KPNW (Right)



Main method (RADS) with cross-validated threshold (0.13)

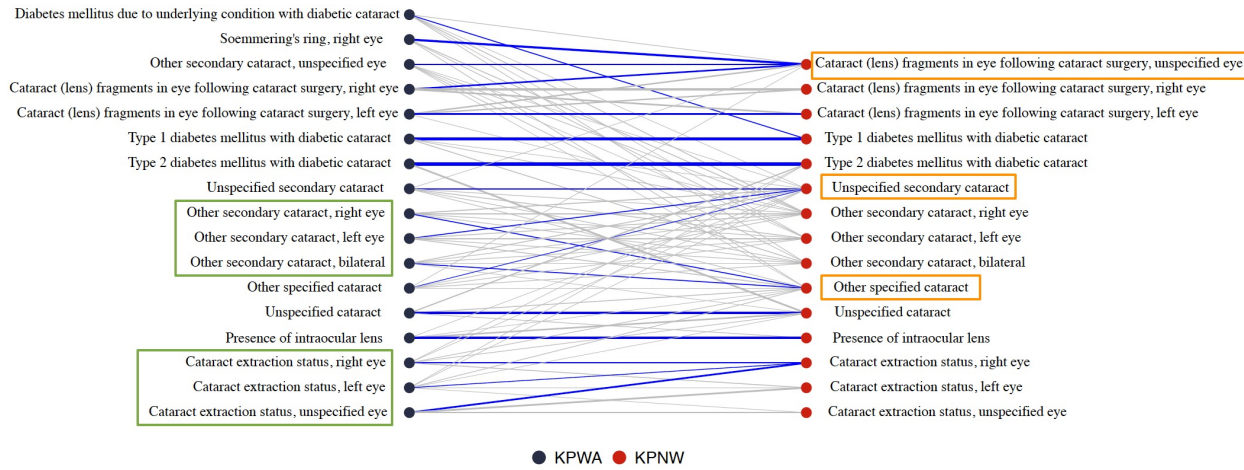
Results: Mapping from KPWA (Left) to KPNW (Right)



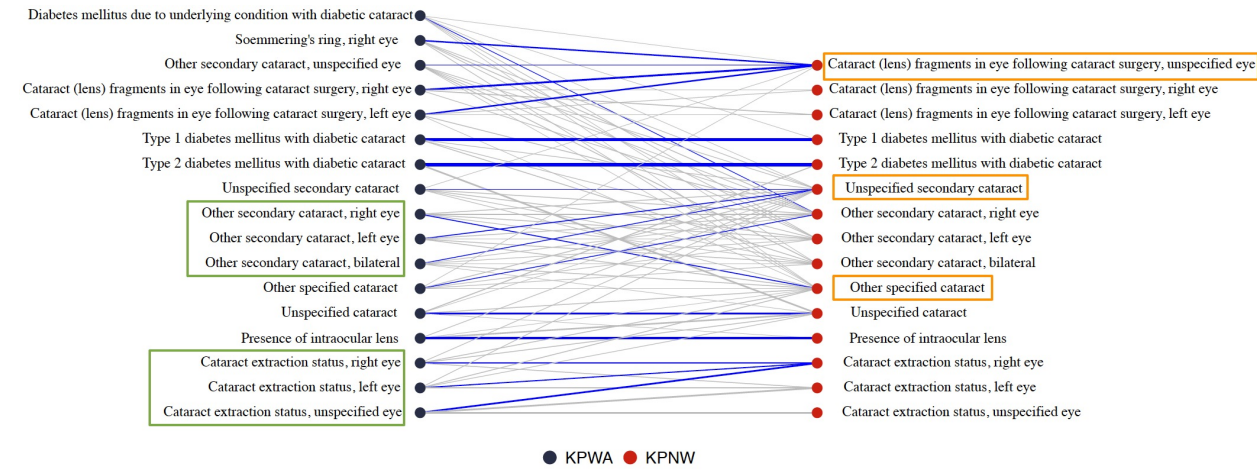
Main method (RADS) with cross-validated threshold (0.13)

Results: Sensitivity Analysis Using Projection-Based Alignment

RADS incorporating code frequency;
cross-validated threshold $\tau = 0.13$

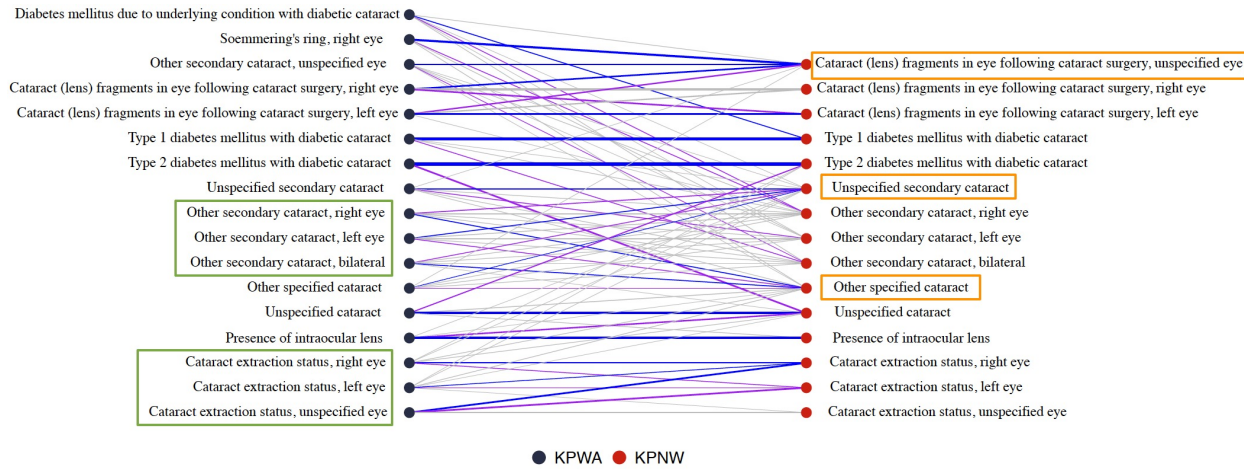


PADS incorporating code frequency;
cross-validated threshold $\tau = 0.16$

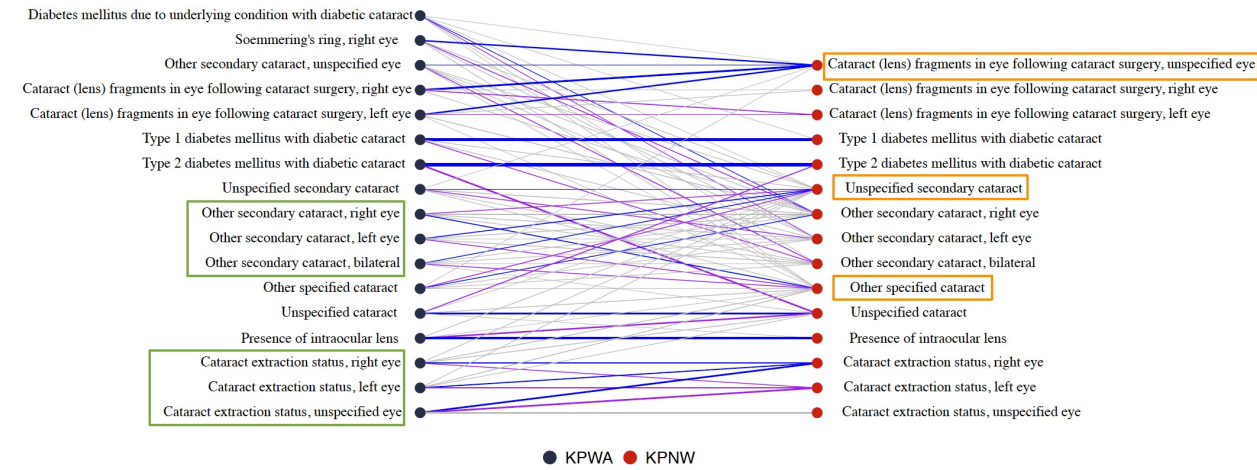


Results: Sensitivity Analysis Using Projection-Based Alignment

RADS incorporating code frequency;
cross-validated threshold $\tau = 0.13$

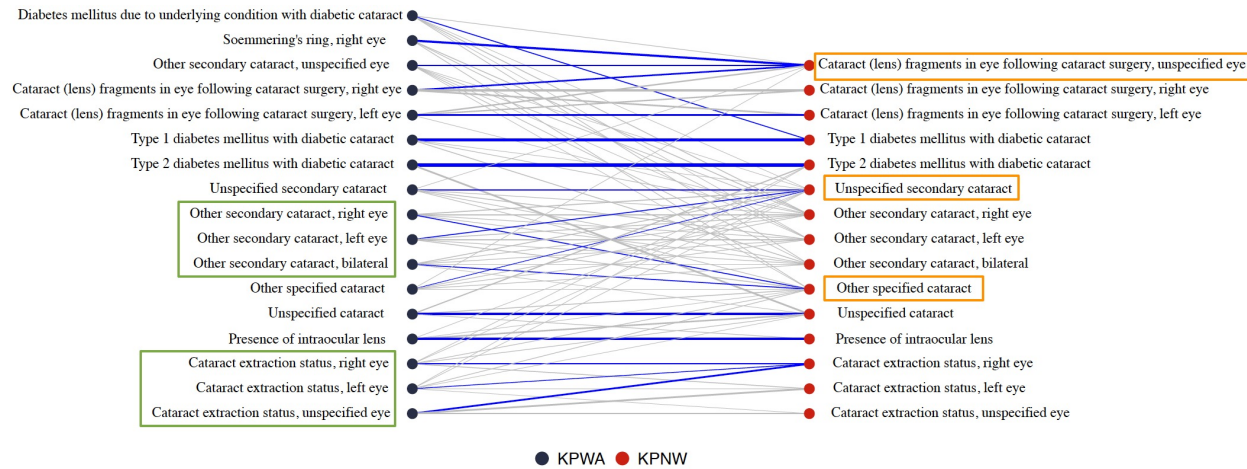


PADS incorporating code frequency;
cross-validated threshold $\tau = 0.16$

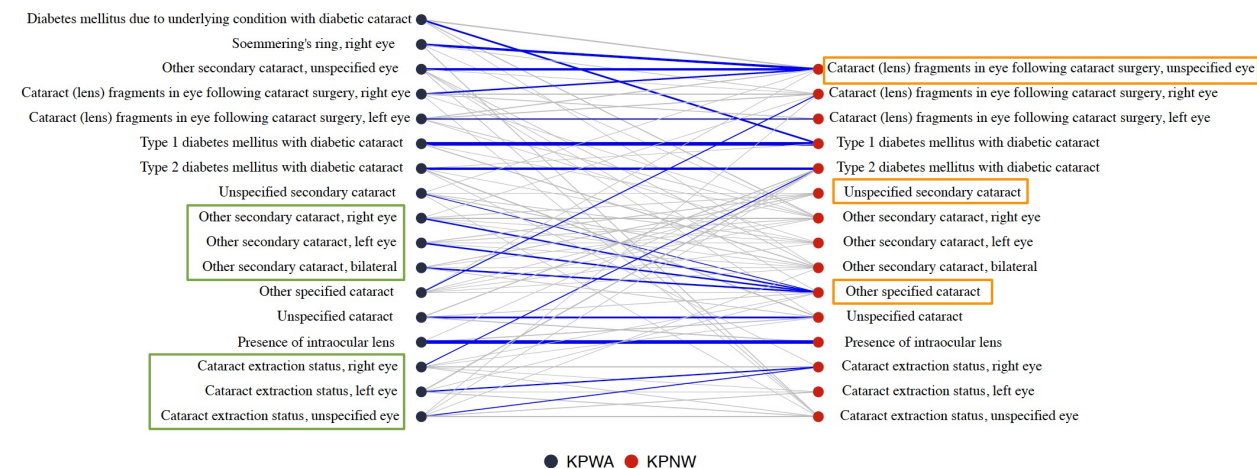


Results: Sensitivity Analysis Using Regression Similarity

RADS incorporating code frequency;
cross-validated threshold $\tau = 0.13$

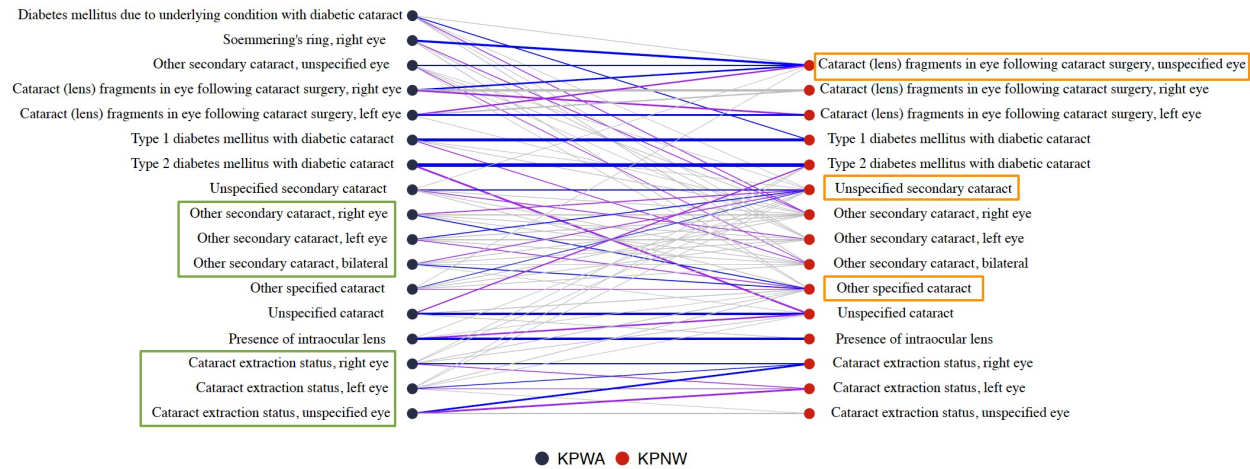


RARS with penalty $\lambda = 0.3$ incorporating code frequency;
cross-validated threshold $\tau = 0.08$

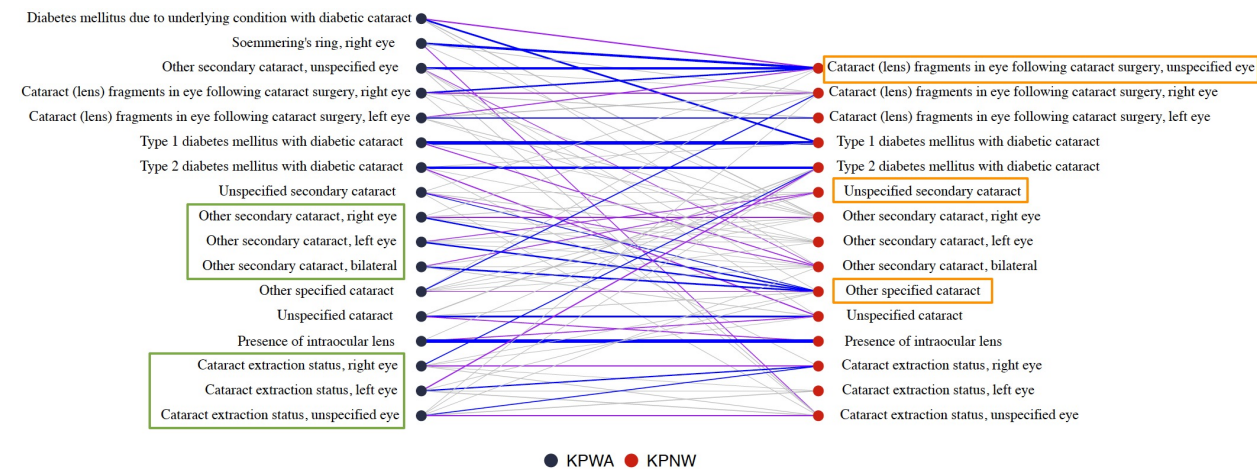


Results: Sensitivity Analysis Using Regression Similarity

RADS incorporating code frequency;
cross-validated threshold $\tau = 0.13$



RARS with penalty $\lambda = 0.3$ incorporating code frequency;
cross-validated threshold $\tau = 0.08$



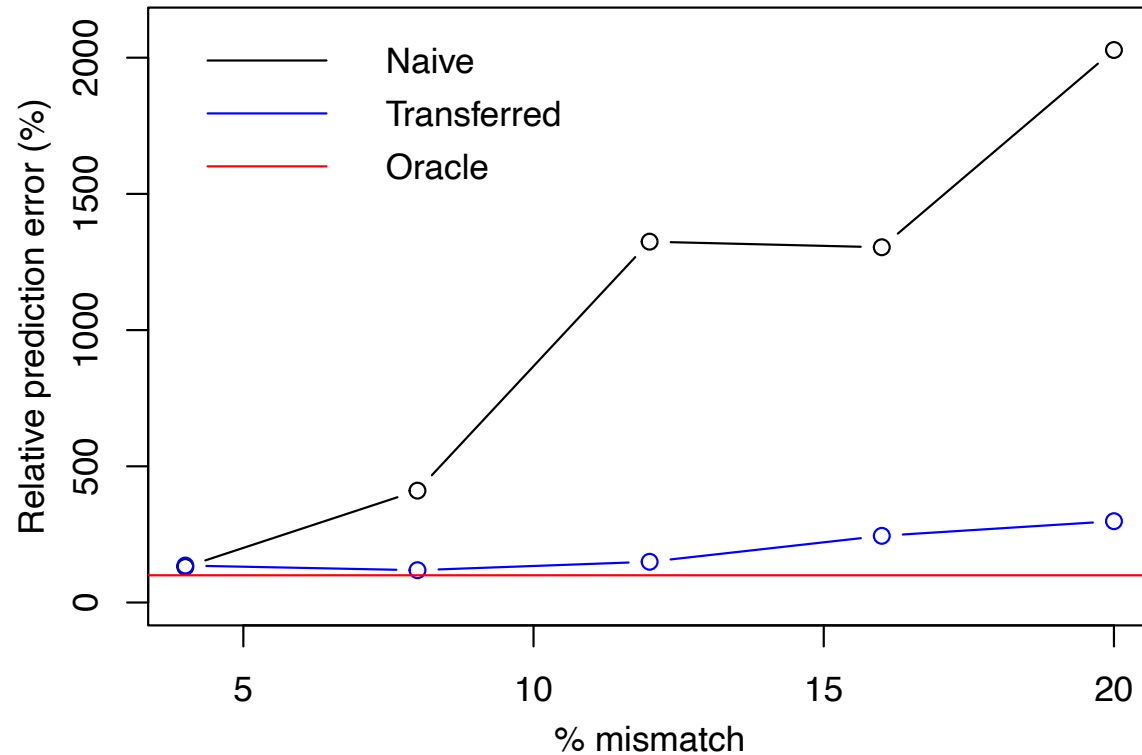


Validation

Validation 1: Why KPNW Uses More “Unspecified” Codes?

- KPNW uses “unspecified” codes, KPWA uses specific codes more frequently in the cataract group
- Specified codes are more likely used when generated externally (out of network)
 - Because external providers are using coding to bill for services so tend to use more specific codes.
 - We hypothesized that KPWA has more external coding
- Validation: Optometry and Ophthalmology Dx codes distribution
 - **KPWA** – 19.76% of Optometry Dx codes and 46.82% of Ophthalmology Dx codes are generated externally
 - **KPNW** – 2.57% of Optometry Dx codes and 0.49% of Ophthalmology Dx codes are generated externally

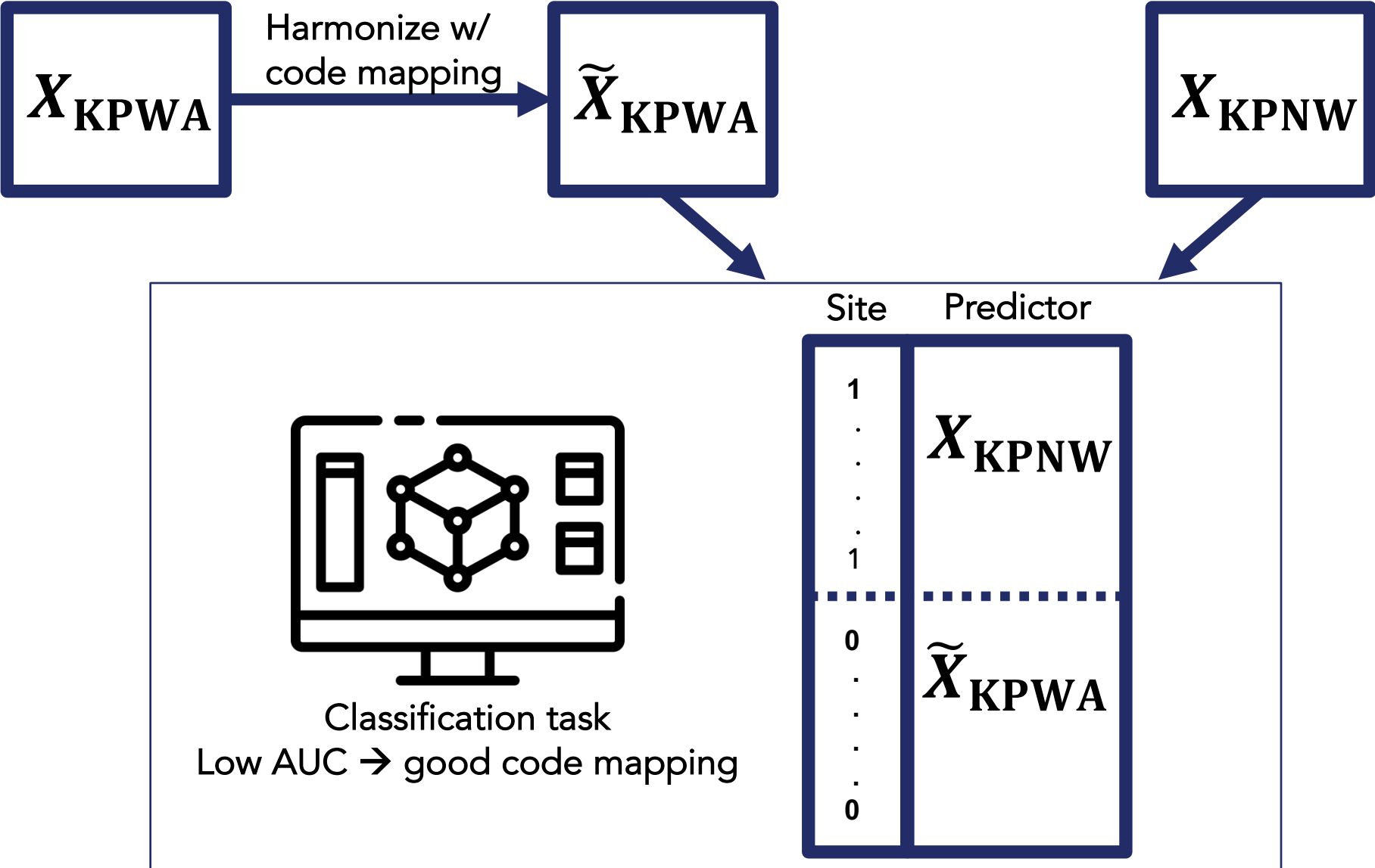
Validation 2: Impact of Data Heterogeneity on Model Transfer and Improvement from Code Mapping



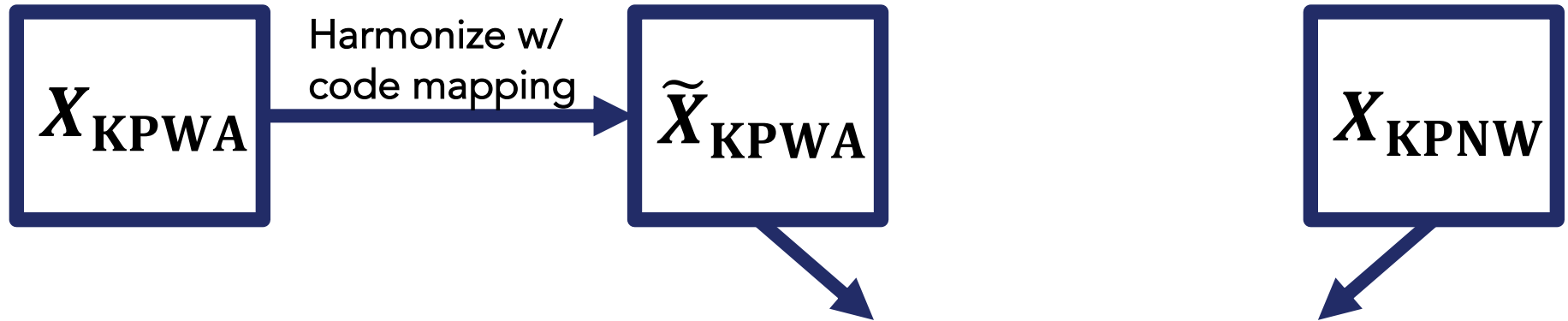
Simulation results:

- High prediction error if directly transfer
- Reduced error by incorporating code mapping

Validation 3: Can We Confuse a Site Classifier After Data Harmonization?



Validation 3: Can We Confuse a Site Classifier After Data Harmonization?



Cross-validated AUC
0.586 (0.580, 0.592)

Site	Predictor
1	X_{KPNW}
.	
.	
.	
1	
0	\tilde{X}_{KPWA}
.	
.	
.	
0	



Conclusion

Summary

Comparing coding patterns between KPWA and KPNW

- Many codes and code groups have significant differences with large magnitude
- KPNW has many local codes
- Potential code substitution in “cataract” group

Mapping codes from KPWA to KPNW

- Code mapping methods can automatically map specific codes in KPWA to “unspecified” codes in KPNW
- Data driven methods are scalable and potentially less error prone compared to human annotation
- Code mapping methods are not extremely sensitive to space alignment and distance measures

All methods are based on sharable summary data to protect patient privacy

Data harmonization is an important first step that improves model transport and multi-institutional studies

Implications for Sentinel in the future

- Develop and implement more methods for semi-automated EHR data harmonization prior to downstream analysis
- Methods studied in the project can be potentially added to the Sentinel QA package in the future to routinely detect and mitigate heterogeneity between data partners and across time

Acknowledgments

FDA

- Patricia Bright
- Jose Hernandez
- Jie Li
- Yong Ma
- Danijela Stojanovic

Kaiser Permanente Washington Health Research Institute

- David S. Carrell
- Kara L. Cushing-Haugen
- Luesa Healy
- Jennifer C. Nelson
- Brian D Williamson

University of Washington, Seattle

- James S. Floyd
- Patrick J. Heagerty

Center for Health Research, Kaiser Permanente Northwest

- Brian L. Hazlehurst
- Denis B. Nyongesa
- Daniel S. Sapp

University of Michigan

- Xianshi Yu
- Yuqi Zhai

Harvard Medical School

- Shirley V. Wang
- Jenna Wong

Harvard T.H. Chan School of Public Health

- Tianxi Cai

Duke University

- Sudha Raman

Vanderbilt University

- Sharon E. Davis

Thank You

Contact: shixu@umich.edu

Break

Innovation Day

April 12, 2023

Sentinel Innovation Center

A PROCESS guide for INferential studies using healthcare data from routine Clinical Practice to EvaLUate causal Effects of Drugs (PRINCIPLED)

Rishi Desai, MS, PhD

Mass General Brigham and Harvard Medical School



Motivation

Why Do We Need Another Framework?

Quality assessment tools

RESEARCH METHODS AND REPORTING

OPEN ACCESS

ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions



Jonathan AC Sterne,¹ Miguel A Hernán,² Barnaby C Reeves,³ Jelena Savović,^{1,4} Nancy D Berkman,⁵ Meera Viswanathan,⁶ David Henry,⁷ Douglas G Altman,⁸ Mohammed T Ansari,⁹ Isabelle Boutron,¹⁰ James R Carpenter,¹¹ An-Wen Chan,¹² Rachel Churchill,¹³ Jonathan J Deeks,¹⁴ Asbjørn Hróbjartsson,¹⁵ Jamie Kirkham,¹⁶ Peter Juni,¹⁷ Yoon K Loke,¹⁸ Theresa D Pigott,¹⁹ Craig R Ramsay,²⁰ Deborah Regidor,²¹ Hannah R Rothstein,²² Lakhbir Sandhu,²³ Pasqualina L Santaguida,²⁴ Holger J Schünemann,²⁵ Beverly Shea,²⁶ Ian Shrier,²⁷ Peter Tugwell,²⁸ Lucy Turner,²⁹ Jeffrey C Valentine,³⁰ Hugh Waddington,³¹ Elizabeth Waters,³² George A Wells,³³ Penny F Whiting,³⁴ Julian PT Higgins³⁵

RESEARCH

The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution

Nancy A. Dreyer, PhD, MPH; Priscilla Valentgas, PhD; Kimberly Westrich, MA; and Robert Dubois, MD

Reporting tools

RESEARCH METHODS AND REPORTING

OPEN ACCESS

The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE)

Check for updates

Sinéad M Langan,¹ Sigrún AJ Schmidt,² Kevin Wing,¹ Vera Ehrenstein,² Stuart G Nicholls,^{3,4} Kristian B Filion,^{5,6} Olaf Klungel,⁷ Irene Petersen,^{2,8} Henrik T Sorensen,² William G Dixon,⁹ Astrid Guttman,^{10,11} Katie Harron,¹² Lars G Hemkens,¹³ David Moher,³ Sebastian Schneeweiss,¹⁴ Liam Smeeth,¹ Miriam Sturkenboom,¹⁵ Erik von Elm,¹⁶ Shirley V Wang,¹⁴ Eric I Benchimol^{10,17,18}

RESEARCH METHODS AND REPORTING

OPEN ACCESS

STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies

Check for updates

Shirley V Wang,¹ Simone Pinheiro,² Wei Hua,² Peter Arlett,^{3,4} Yoshiaki Uyama,⁵ Jesse A Berlin,⁶ Dorothee B Bartels,⁷ Kristijan H Kahler,⁹ Lily G Besette,¹ Sebastian Schneeweiss¹

Best practices

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

Guidance for Industry

DRAFT GUIDANCE



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH



European Network of Centres for Pharmacoepidemiology and Pharmacovigilance

EMA/95098/2010 Rev.9

The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology (Revision 9)

Misc: Highly specific or focusing on parts of the process

Journal of the American Medical Informatics Association, 27(8), 2020, 1331–1337
doi: 10.1093/jamia/ocaa103



Clinical Pharmacology & Therapeutics

REVIEW | Open Access

The Structured Process to Identify Fit-for-purpose Data (SPIFD): A data feasibility assessment framework

Nicolle M Gatto, Ulka B Campbell, Emily Rubinstein, Ashley Jaksa, Pattria Mattox, Jingping Mo, Robert F Reynolds

First published: 30 October 2021 | <https://doi-org.ezp-prod1.hul.harvard.edu/10.1002/cpt.2466>

Perspective

Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND)

Martijn J. Schuemie^{1,2}, Patrick B. Ryan^{1,3}, Nicole Pratt⁴, Ruijun Chen^{3,5}, Seng Chan You⁶, Harlan M. Krumholz⁷, David Madigan⁸, George Hripcsak^{3,9}, and Marc A. Suchard^{2,10}

Why do we need another framework?

What do we have?

- Various tools exist in the literature for quality assessment, reporting, and describing best practices for pharmacoepidemiologic research

What don't we have?

- None of these tools offer a general framework to guide decision making at various steps when designing a study to answer a causal question

Vision for a framework to guide principled investigations using healthcare data

- The Sentinel Innovation Center is developing a causal inference framework proposing a stepwise process that systematically considers key choices with respect to design and analysis that influence the validity of non-interventional studies conducted with healthcare data
- A standardized process outlined in this framework will serve as a guide to inform the conduct of non-interventional studies using healthcare data for drug-outcome evaluation
- Key considerations to meet the FDA need of informing regulatory decision making based on such investigations
 - Limit variations in practice across investigators by outlining a general process
 - Focus on repeatability of the process
 - Written and endorsed by independent experts

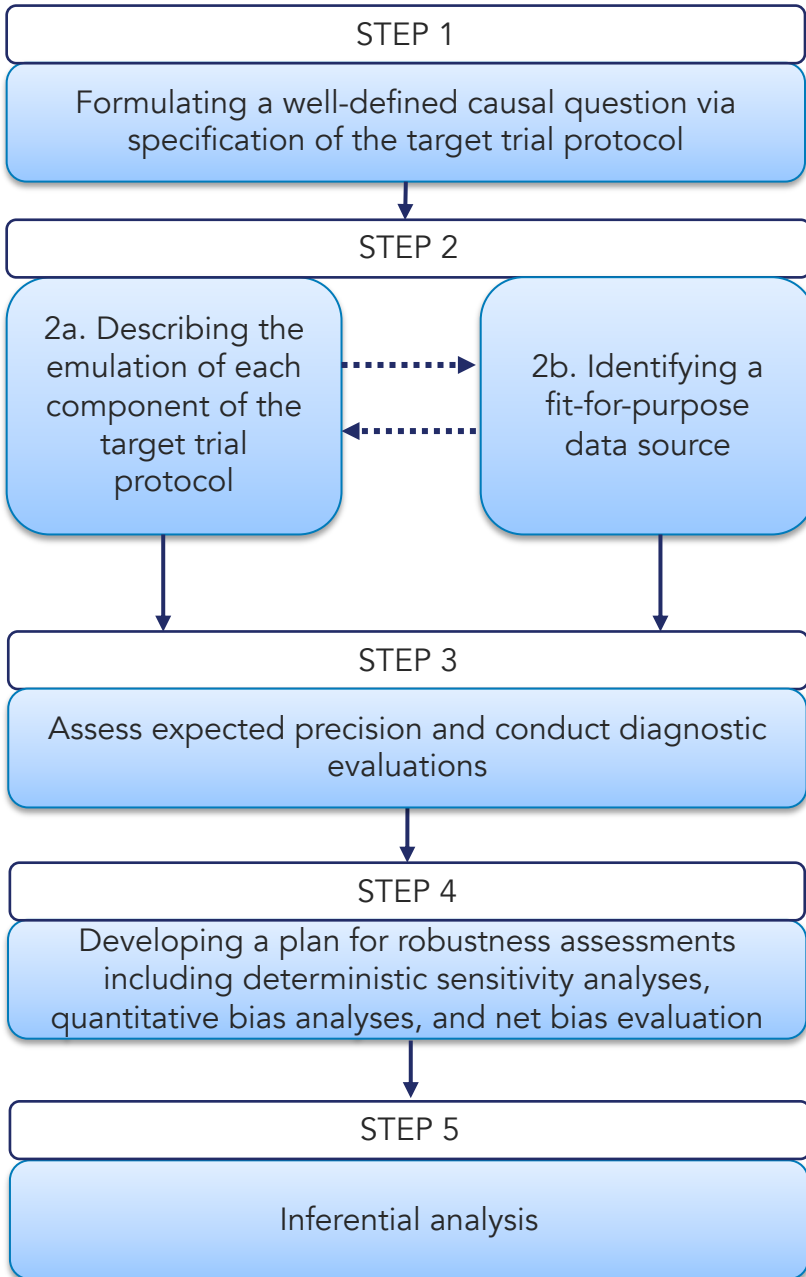


Overview of the Process

Process Overview

Study planning

Inference



See Table for Step 1

Question	Answer	Implications for study design
What is the causal question?	Define the causal question in terms of a causal diagram (DAG) and a causal model.	Identify the causal pathways and the variables that are affected by the treatment.
What is the target trial protocol?	Define the target trial protocol in terms of a causal diagram (DAG) and a causal model.	Identify the causal pathways and the variables that are affected by the treatment.
What are the data sources?	Identify the data sources that are available for the target trial protocol.	Identify the data sources that are available for the target trial protocol.
What are the data quality issues?	Identify the data quality issues that are associated with the data sources.	Identify the data quality issues that are associated with the data sources.
What are the data access issues?	Identify the data access issues that are associated with the data sources.	Identify the data access issues that are associated with the data sources.
What are the data privacy issues?	Identify the data privacy issues that are associated with the data sources.	Identify the data privacy issues that are associated with the data sources.
What are the data security issues?	Identify the data security issues that are associated with the data sources.	Identify the data security issues that are associated with the data sources.
What are the data integrity issues?	Identify the data integrity issues that are associated with the data sources.	Identify the data integrity issues that are associated with the data sources.
What are the data consistency issues?	Identify the data consistency issues that are associated with the data sources.	Identify the data consistency issues that are associated with the data sources.
What are the data completeness issues?	Identify the data completeness issues that are associated with the data sources.	Identify the data completeness issues that are associated with the data sources.
What are the data timeliness issues?	Identify the data timeliness issues that are associated with the data sources.	Identify the data timeliness issues that are associated with the data sources.
What are the data accuracy issues?	Identify the data accuracy issues that are associated with the data sources.	Identify the data accuracy issues that are associated with the data sources.
What are the data reliability issues?	Identify the data reliability issues that are associated with the data sources.	Identify the data reliability issues that are associated with the data sources.
What are the data validity issues?	Identify the data validity issues that are associated with the data sources.	Identify the data validity issues that are associated with the data sources.
What are the data usability issues?	Identify the data usability issues that are associated with the data sources.	Identify the data usability issues that are associated with the data sources.

See Figure for Step 2

See Figure for Step 3

See Figure for Step 4

Fit-for-purpose data not available for the target trial

Fit-for-purpose data available for the target trial

⊘ Reassess the research question in Step 1

✓ Consider protocol registration, Move on to step 3

Desired precision not achievable or diagnostic criteria not met

Desired precision achievable and diagnostic criteria met

⊘ Consider alternative design choices and data sources in Step 2 or reassess the research question in Step 1

✓ Move on to step 4

✓ Consider logging outcome counts and diagnostic evaluations along with pre-specified robustness assessments as amendments to the registered protocol

Step 1: Formulating a well-defined causal question via specification of the target trial protocol¹

STEP 1

Formulating a well-defined causal question via specification of the target trial protocol

See Table for Step 1

Question	Answer	Reference
Should I randomize?	Randomization is the gold standard for causal inference. It is the only way to ensure that the treatment and control groups are comparable at baseline, and that any differences in outcomes are due to the treatment itself, rather than to confounding factors.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I blind?	Blinding is important to reduce bias in the measurement of outcomes. It is particularly important in trials where the outcome is subjective, such as pain or quality of life. Blinding should be implemented for both participants and outcome assessors.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a placebo?	A placebo is a treatment that has no therapeutic effect. It is used to control for the placebo effect, which is a psychological response that can occur simply because a person believes they are receiving a treatment. Placebos should be used when the treatment being studied is subjective, and when the placebo effect is likely to be significant.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a control group?	A control group is a group of participants who do not receive the treatment being studied. It is used to compare the outcomes of the treatment group to the outcomes of the control group. A control group is essential for causal inference.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a cohort study?	A cohort study is a study in which a group of participants is followed over time to see if they develop the outcome of interest. Cohort studies are useful for studying the natural history of a disease, and for identifying risk factors for disease.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a case-control study?	A case-control study is a study in which a group of participants with the outcome of interest is compared to a group of participants without the outcome. Case-control studies are useful for studying the causes of a disease, and for identifying risk factors for disease.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a cross-sectional study?	A cross-sectional study is a study in which a group of participants is surveyed at a single point in time. Cross-sectional studies are useful for describing the prevalence of a disease, and for identifying risk factors for disease.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.
Should I use a randomized controlled trial?	A randomized controlled trial is a study in which participants are randomly assigned to either the treatment group or the control group. Randomized controlled trials are the gold standard for causal inference.	Hernan MA, Robins JM. Causal inference in epidemiology: the role of randomization. <i>Am J Epidemiol</i> . 2016;183:758-764.

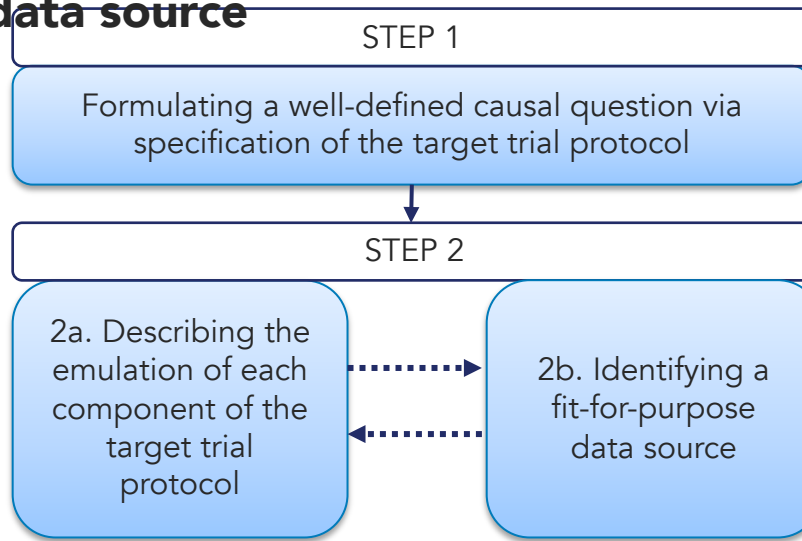
¹ Hernan and Robins. *Am J Epidemiol*. 2016;183:758-764

Step 1: Formulating a well-defined causal question via specification of the target trial protocol¹

Element	Target trial	Translation using healthcare data sources
Exposure ("treatment strategies")	Initiation of antidiabetic treatments T1: SGLT2i T2: DPP4 inhibitors	First prescription dispensing of SGLT2i (canagliflozin, dapagliflozin, empagliflozin) or DPP4 inhibitors (Alogliptin, Linagliptin, Saxagliptin, Sitagliptin) identified based on pharmacy claims
Exposure assignment	Randomized non-blinded	Non-randomized non-blinded
Eligibility criteria (Assessed before treatment start, aka "baseline")	Patients with type 2 diabetes mellitus, no use of study medications before randomization, no history of end stage renal disease (ESRD), no history of HIV	<u>Observability related</u> : continuous Medicare A, B, D enrollment for 6 months and recorded HbA1c test results in EHRs before study medication initiation <u>Treatment related</u> : No prior use of study medications prior to cohort entry <u>Indication related</u> : Diagnosis of type 2 diabetes based on ICD codes recorded pre-exposure <u>Other</u> : No history of ESRD or HIV based on ICD codes or procedure codes for dialysis pre-exposure
Follow-up start (Time 0)	At randomization	At first prescription dispensing
Follow-up end	First of administrative end of follow-up, loss to follow-up, death, treatment discontinuation, or outcome occurrence	Same as target trial
Primary outcome	Genital infections with case adjudication	Genital infections recorded in medical claims
Baseline covariates	-	Demographics, diabetes severity related variables including micro and macrovascular complications, HbA1c, comorbid conditions, comedications, markers for healthy behavior and healthcare utilization
Causal estimand	Per protocol effect (effect of receiving the treatment as stated in the protocol)	Observational analogue of per protocol effect (often referred to as "as-treated," or "on treatment")
Statistical analysis	A Cox proportional hazards model	Adjustment of baseline confounding with propensity score stratification and weighting followed by an outcome analysis using a weighted Cox proportional hazards model
Subgroup analyses	Stratified by gender, age, and baseline risk	Same as target trial

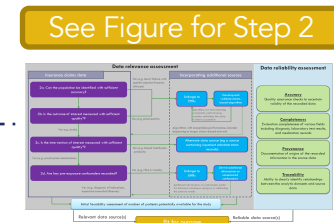
¹ Hernan and Robins. Am J Epidemiol. 2016;183:758-764.

Step 2a: Describing the emulation of each component of the target trial protocol; 2b: Identifying a fit-for-purpose data source



See Table for Step 1

Question	Response	Notes
1. Is the research question well-defined and specific?	Yes	Research question is well-defined and specific.
2. Is the research question relevant to the target population?	Yes	Research question is relevant to the target population.
3. Is the research question answerable by a trial?	Yes	Research question is answerable by a trial.
4. Is the research question important to the target population?	Yes	Research question is important to the target population.
5. Is the research question novel?	Yes	Research question is novel.
6. Is the research question ethical?	Yes	Research question is ethical.
7. Is the research question feasible?	Yes	Research question is feasible.
8. Is the research question acceptable?	Yes	Research question is acceptable.
9. Is the research question implementable?	Yes	Research question is implementable.
10. Is the research question evaluable?	Yes	Research question is evaluable.
11. Is the research question reportable?	Yes	Research question is reportable.
12. Is the research question disseminable?	Yes	Research question is disseminable.
13. Is the research question acceptable to the target population?	Yes	Research question is acceptable to the target population.
14. Is the research question acceptable to the sponsor?	Yes	Research question is acceptable to the sponsor.
15. Is the research question acceptable to the regulatory authorities?	Yes	Research question is acceptable to the regulatory authorities.
16. Is the research question acceptable to the public?	Yes	Research question is acceptable to the public.
17. Is the research question acceptable to the funding body?	Yes	Research question is acceptable to the funding body.
18. Is the research question acceptable to the media?	Yes	Research question is acceptable to the media.
19. Is the research question acceptable to the public health community?	Yes	Research question is acceptable to the public health community.
20. Is the research question acceptable to the wider community?	Yes	Research question is acceptable to the wider community.



Fit-for-purpose data not available for the target trial

Fit-for-purpose data available for the target trial



Reassess the research question in Step 1

Consider protocol registration, Move on to step 3

Step 2a: Describing the emulation of each component of the target trial protocol

A structured protocol detailing operationalization of variable definitions, including all codes and algorithms used for eligibility criteria, treatment strategies (including treatment initiation and discontinuation), outcomes, and confounders

Other considerations including statistical analysis plans for the primary analysis

Example of a template- STaRT RWE²

 OPEN ACCESS

 Check for updates

For numbered affiliations see end of the article.

Correspondence to: S V Wang, Division of Pharmacoepidemiology and Pharmacoeconomics, 1620 Tremont Street, Suite 3030, Boston, MA 02120, USA swang1@bwh.harvard.edu (ORCID 0000-0001-7761-7090)

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

Cite this as: *BMJ* 2021;372:m4856 <http://dx.doi.org/10.1136/bmj.m4856>

Accepted: 10 December 2020

STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies

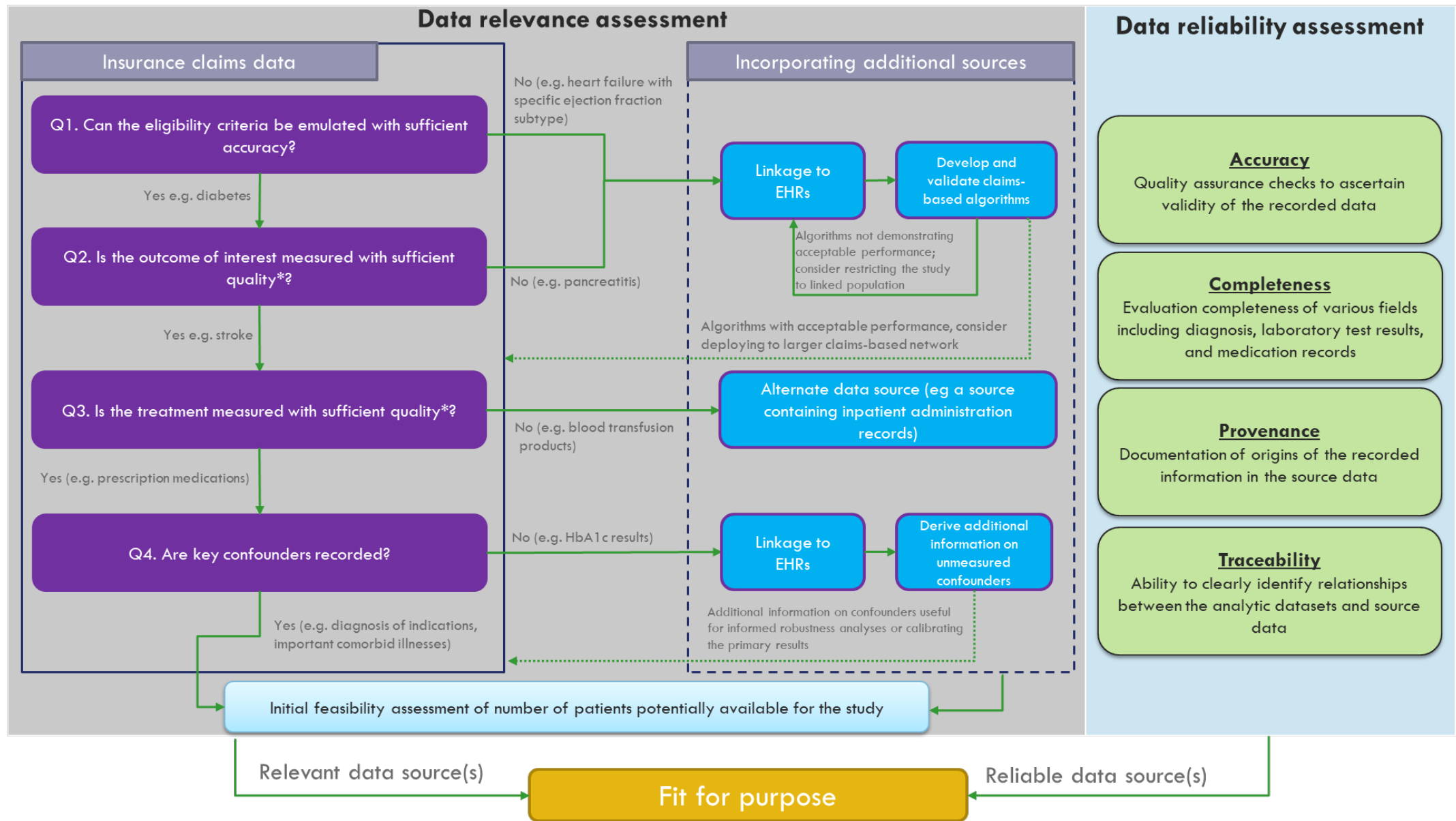
Shirley V Wang,¹ Simone Pinheiro,² Wei Hua,² Peter Arlett,^{3,4} Yoshiaki Uyama,⁵ Jesse A Berlin,⁶ Dorothee B Bartels,⁷ Kristijan H Kahler,⁹ Lily G Bessette,¹ Sebastian Schneeweiss¹

In alignment with the International Council of Harmonization's strategic goals, a public-private consortium has developed a structured template for planning and reporting on the implementation of real world evidence (RWE) studies of the safety and effectiveness of treatments. The template serves as a guiding tool for designing and conducting reproducible RWE studies; set clear expectations for transparent communication of RWE methods; reduce misinterpretation of prose that lacks specificity; allow reviewers to quickly orient and find key information; and facilitate reproducibility, validity assessment, and evidence synthesis. The template is intended for use with studies of the effectiveness and safety of medical

products and is compatible with multiple study designs, data sources, reporting guidelines, checklists, and bias assessment tools.

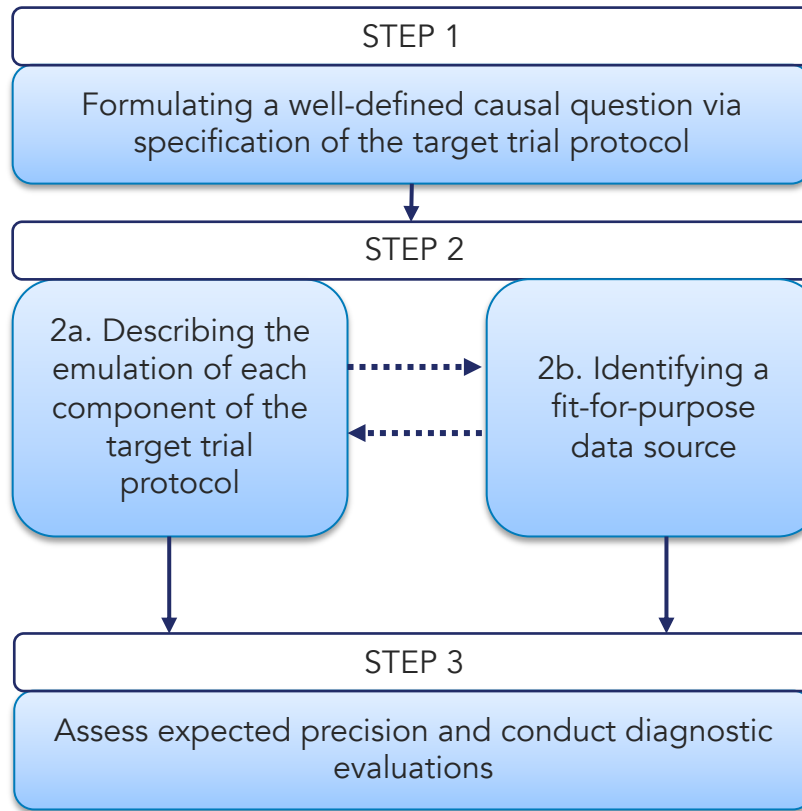
Real world evidence (RWE) generated from sources of real world data via the application of principled database epidemiology increasingly informs important decisions about the clinical effectiveness of medical products and interventions.¹⁻⁵ Unlike clinical trials, which can leverage the power of randomisation, or non-randomised studies with prospective data collection for a specific research purpose, most RWE studies make secondary use of electronic data collected as part of routine healthcare processes (eg, administrative claims and electronic health records). Generating high quality evidence when analysing data not collected for research purposes requires decision making about many complex design and analytical parameters to handle temporality, measurement, confounding, and other potential sources of bias. Compared with trials and non-experimental studies that prospectively collect data for a research question, RWE studies have greater variability in design and analysis options. Owing to the current lack of structure in study reporting, assessment of RWE studies often

Step 2b: Identifying a fit-for-purpose data source



* quality = accuracy with respect to timing and completeness for interventions; PPV, sensitivity, specificity for binary outcomes; proportion missing for continuous outcomes; accurate onset for time to event outcomes; availability of long-term follow-up data for latent outcomes

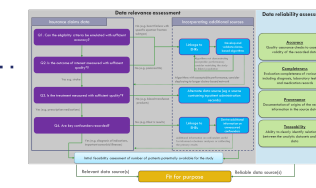
Step 3: Assess expected precision and conduct diagnostic evaluations



See Table for Step 1

Question	Response	Implications for the target trial
1. Is the causal question well-defined?	Yes	Proceed to Step 2
2. Is the causal question well-specified?	Yes	Proceed to Step 2
3. Is the causal question well-posed?	Yes	Proceed to Step 2
4. Is the causal question well-answered?	Yes	Proceed to Step 2
5. Is the causal question well-answered?	Yes	Proceed to Step 2
6. Is the causal question well-answered?	Yes	Proceed to Step 2
7. Is the causal question well-answered?	Yes	Proceed to Step 2
8. Is the causal question well-answered?	Yes	Proceed to Step 2
9. Is the causal question well-answered?	Yes	Proceed to Step 2
10. Is the causal question well-answered?	Yes	Proceed to Step 2

See Figure for Step 2



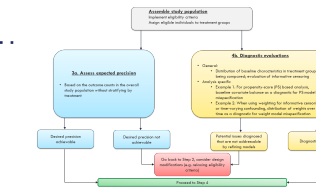
Fit-for-purpose data not available for the target trial

Reassess the research question in Step 1

Fit-for-purpose data available for the target trial

Consider protocol registration,
Move on to step 3

See Figure for Step 3



Desired precision not achievable or diagnostic criteria not met

Consider alternative design choices and data sources in Step 2 or reassess the research question in Step 1

Desired precision achievable and diagnostic criteria met

Move on to step 4

Step 3: Assess expected precision and conduct diagnostic evaluations

Assemble study population
Implement eligibility criteria
Assign eligible individuals to treatment groups

3a. Assess expected precision⁴

- Based on the outcome counts in the overall study population without stratifying by treatment

3b. Diagnostic evaluations

- General:
 - Distribution of baseline characteristics in treatment groups being compared; evaluation of informative censoring
- Analysis specific
 - Example 1: For propensity-score (PS) based analysis, baseline covariate balance as a diagnostic for PS model misspecification
 - Example 2: When using weighting for informative censoring or time-varying confounding, distribution of weights over time as a diagnostic for weight model misspecification

Desired precision achievable

Desired precision not achievable

Potential issues diagnosed that are not addressable by refining models

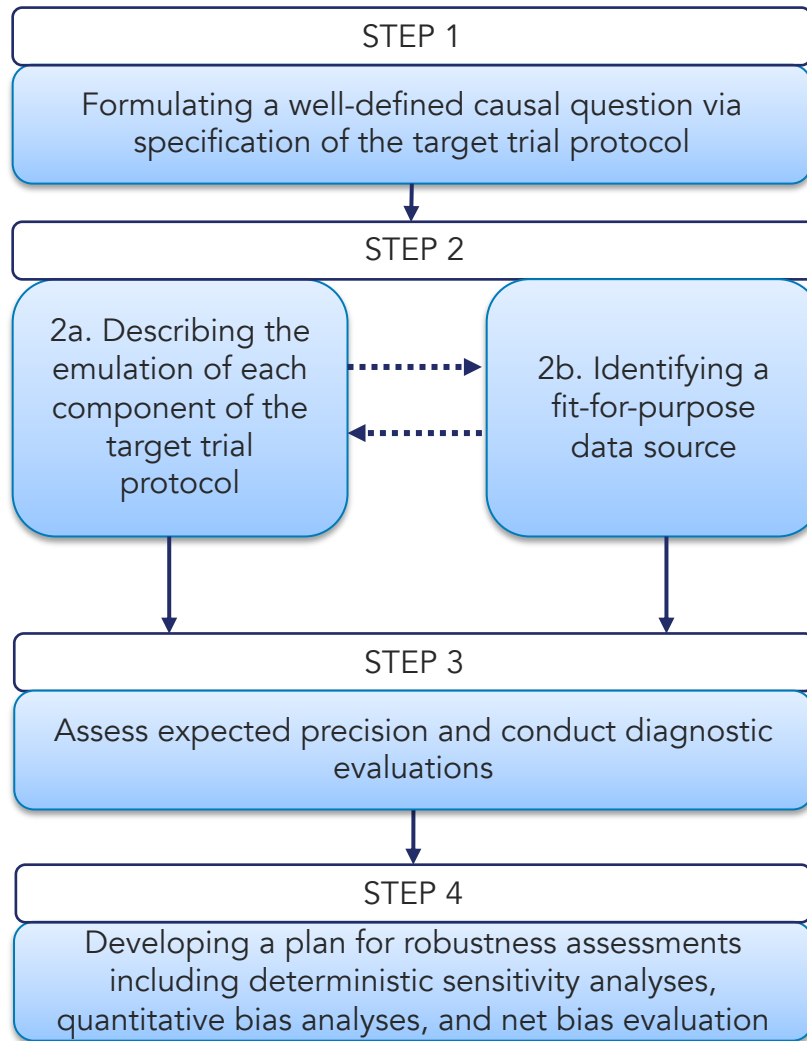
Diagnostics passed

Go back to Step 2, consider design modifications (e.g. relaxing eligibility criteria)

Proceed to Step 4

⁴ Rothman and Greenland, *Epidemiology* 2018;29: 599–603

Step 4: Developing a plan for robustness assessments including deterministic sensitivity analyses, quantitative bias analyses, and net bias evaluation



See Table for Step 1

Question	Answer	Implications for the target trial
1.1. What is the research question?
1.2. What is the target population?
1.3. What is the intervention?
1.4. What is the comparator?
1.5. What is the outcome?
1.6. What is the time horizon?
1.7. What is the setting?
1.8. What is the perspective?
1.9. What is the discount rate?
1.10. What is the cost of the intervention?
1.11. What is the cost of the comparator?
1.12. What is the cost of the outcome?
1.13. What is the cost of the time horizon?
1.14. What is the cost of the setting?
1.15. What is the cost of the perspective?
1.16. What is the cost of the discount rate?
1.17. What is the cost of the cost of the intervention?
1.18. What is the cost of the cost of the comparator?
1.19. What is the cost of the cost of the outcome?
1.20. What is the cost of the cost of the time horizon?
1.21. What is the cost of the cost of the setting?
1.22. What is the cost of the cost of the perspective?
1.23. What is the cost of the cost of the discount rate?
1.24. What is the cost of the cost of the cost of the intervention?
1.25. What is the cost of the cost of the cost of the comparator?
1.26. What is the cost of the cost of the cost of the outcome?
1.27. What is the cost of the cost of the cost of the time horizon?
1.28. What is the cost of the cost of the cost of the setting?
1.29. What is the cost of the cost of the cost of the perspective?
1.30. What is the cost of the cost of the cost of the discount rate?

See Figure for Step 2

Fit-for-purpose data not available for the target trial Reassess the research question in Step 1

Fit-for-purpose data available for the target trial Consider protocol registration, Move on to step 3

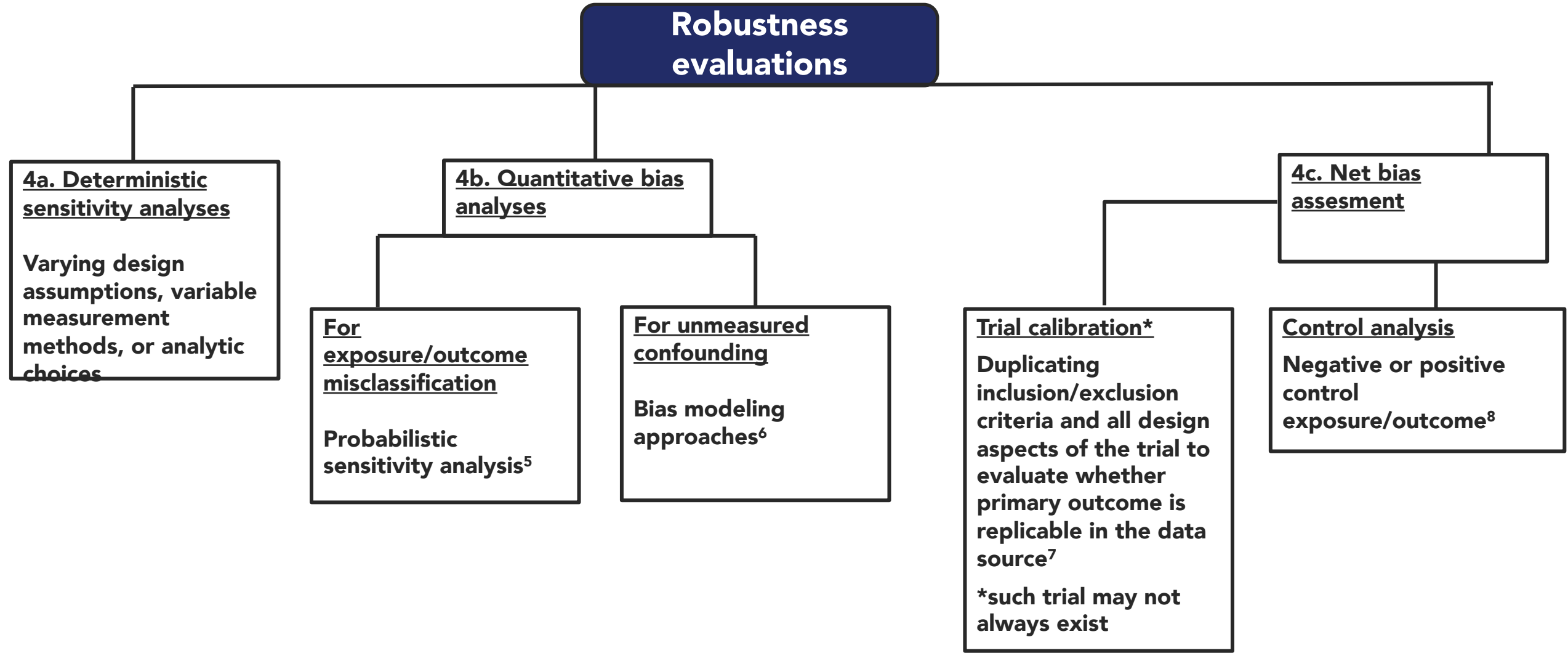
See Figure for Step 3

Desired precision not achievable or diagnostic criteria not met Consider alternative design choices and data sources in Step 2 or reassess the research question in Step 1

Desired precision achievable and diagnostic criteria met Move on to step 4

See Figure for Step 4

Consider logging outcome counts and diagnostic evaluations along with pre-specified robustness assessments as amendments to the registered protocol

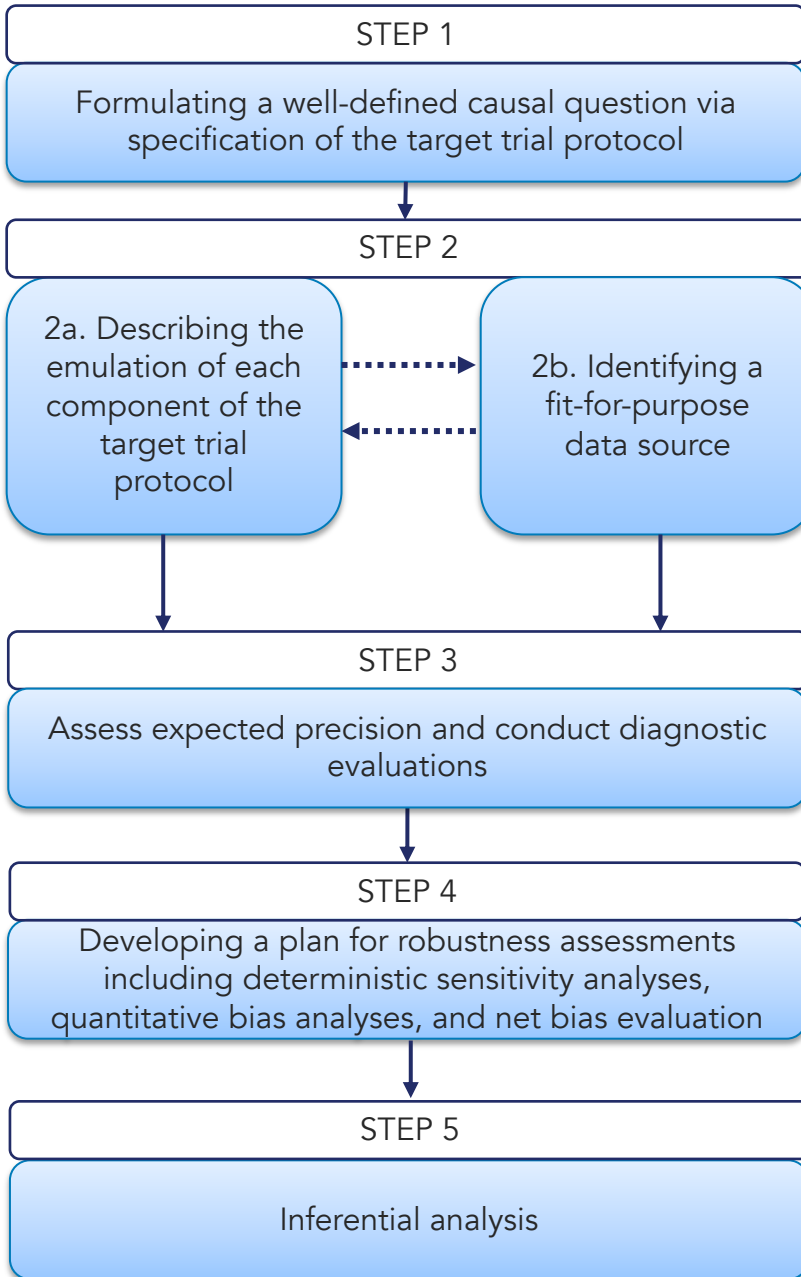


⁵ Fox et al. International Journal of Epidemiology 2005;34:1370–1376
⁶ Schneeweiss. Pharmacoepidemiology Drug Saf 2006; 15: 291–303
⁷ Khosrow-Khavar et al. Annals Rheum Dis. 2022 Jun;81(6):798-804.
⁸ Lipsitch et al. Epidemiology 2010;21: 383–388

Step 5: Inferential analysis

Study planning

Inference



See Table for Step 1

Question	Answer	Implications for study design
What is the research question?
What is the target population?
What is the target intervention?
What is the target comparator?
What is the target outcome?
What is the target time point?
What is the target setting?
What is the target data source?
What is the target data type?
What is the target data format?
What is the target data access?
What is the target data security?
What is the target data privacy?
What is the target data integrity?
What is the target data availability?
What is the target data completeness?
What is the target data accuracy?
What is the target data consistency?
What is the target data timeliness?
What is the target data reliability?
What is the target data validity?
What is the target data veracity?
What is the target data trustworthiness?

See Figure for Step 2

Fit-for-purpose data not available for the target trial
 Fit-for-purpose data available for the target trial

⊘ Reassess the research question in Step 1

✓ Consider protocol registration, Move on to step 3

See Figure for Step 3

Desired precision not achievable or diagnostic criteria not met

Desired precision achievable and diagnostic criteria met

⊘ Consider alternative design choices and data sources in Step 2 or reassess the research question in Step 1

✓ Move on to step 4

See Figure for Step 4

✓ Consider logging outcome counts and diagnostic evaluations along with pre-specified robustness assessments as amendments to the registered protocol

Acknowledgements

Mass General Brigham

- Rishi J. Desai
- Shirley V. Wang
- Sushama Kattinakere Sreedhara
- Luke Zobotka
- Farzin Khosrow-Khavar
- Richard Wyss
- Elisabetta Patorno
- Sebastian Schneeweiss

Kaiser Washington

- Jennifer C. Nelson

Univ. of Michigan

- Xu Shi

Harvard Pilgrim Health Care Institute

- Darren Toh

FDA

- Sarah Dutcher
- Jie Li
- Christina Greene
- Hana Lee
- Robert Ball
- Gerald Dal Pan

John Hopkins Univ.

- Jodi B. Segal

McGill University

- Samy Suissa

Research Triangle Institute/Boston Univ.

- Kenneth J. Rothman

UCLA

- Sander Greenland

Harvard Univ.

- Miguel Hernan

Univ. of Washington

- Patrick J. Heagerty

Thank You

Contact:
rdesai@bwh.harvard.edu

Approaches to Handling Partially Observed Confounder Data from Electronic Health Records (EHR) in Non-randomized Studies of Medication Outcomes

Janick Weberpals, RPh, PhD

Division of Pharmacoepidemiology and Pharmacoeconomics,
Mass General Brigham and Harvard Medical School

Disclosures





Janick Weberpals reports prior employment by Hoffmann-La Roche and previously held shares in Hoffmann-La Roche.



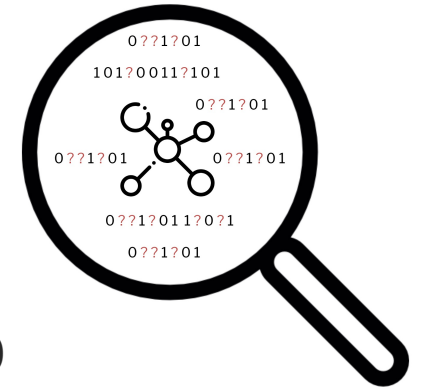
Background



Background

- Administrative claims databases are increasingly linked to electronic health records (EHR) to **improve confounding adjustment** for variables which cannot be measured in claims
- Examples:
 - Labs (HbA1c, LDL, etc.) 
 - Vitals (Blood pressure, BMI, etc.) 
 - Disease-specific data (cancer stage, biomarkers, etc.) 
 - Lifestyle factors (smoking, alcohol, etc.) 
- These covariates are often just partially observed for various reasons
 - Physician did not perform/order a certain test
 - Certain measurements are just collected for particularly sick patients
 - Information is ‘hiding’ in unstructured records, e.g. clinical notes

Knowledge Gaps and Objectives



Established missing data taxonomies

Mechanisms: Missing completely at random (**MCAR**), at random (**MAR**) and not at random (**MNAR**)

Patterns: Monotone, Non-monotone

Unresolved challenges for causal inference

Textbook '**MNAR**' definition often not helpful for causal inference: MNAR = anything that is not MCAR or MAR

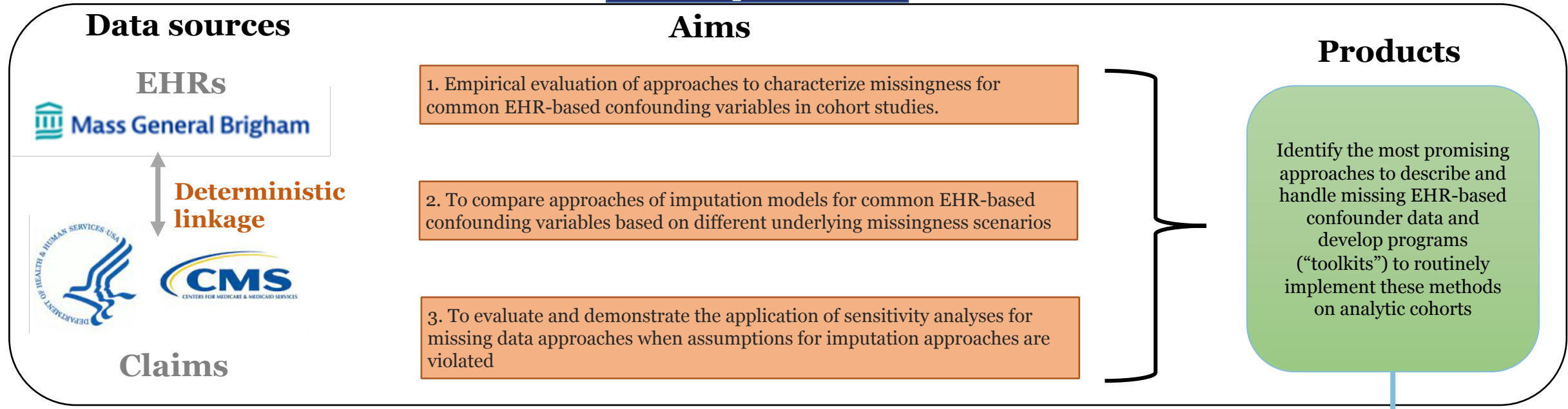
How do any of these mechanisms relate to **bias**, given the strength of correlations of between exposure, covariates and outcomes in high-dimensional database settings (e.g., database linkages)

Many siloed approaches have been proposed, but not much guidance on **systematic end-to-end approaches**

Aims of this workstream

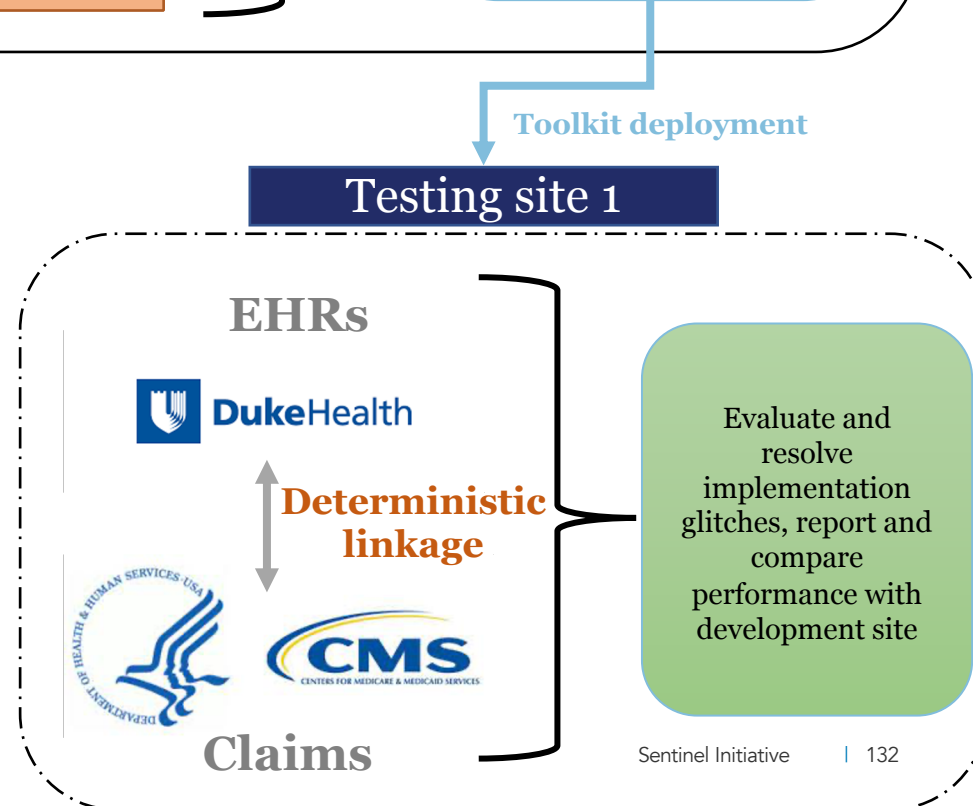
- Integrate Rubin's framework with multivariate missing data under causal exposure-outcome relationships:
Structural missing data assumptions
- Establish routine diagnostics for structural missingness based on causal diagrams/M-graphs which provide a more natural way to understand the assumptions regarding missing data for a given research question
- Provide context which analytical decisions (e.g., as part of primary analysis and sensitivity analyses) may be best suited under resulting structural missingness assumptions

- Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581-592. doi:10.2307/2335739
- Mitra, R., McGough, S.F., Chakraborti, T. et al. Learning from data with structured missingness. *Nat Mach Intell* 5, 13–23 (2023)
- Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Curran Associates Inc.; 2013:1277-1285.



Phases of this workflow

1. Test assumptions and ability to differentiate underlying missingness mechanisms under best possible realistic simulated scenarios
2. Assess which analytical approach works best under which scenario
3. Based on theoretical results from 1 & 2: funnel insights into an implementable and operationalizable toolkit which will be developed in our group and tested/validated at Duke





Plasmode Simulation Study

Cohort Entry Date (Day 0)
 [Dispensation of Exposure Group (SGLT2 inhibitor)] OR
 [Reference Group (DPP4 inhibitor)]

Washout Window
 (No Use of an SGLT-2 Inhibitor or a DPP4 Inhibitor)
 Days [-180, -1]

Part A, B, and D Enrollment with EHR activity
 Days [-180, 0]

INCL: Type 2 DM & **COPD, HbA1c, BMI, Smoking** results available
 Days [-180, 0]

EXCL: Concurrent use of both study
 drugs Days [0]

EXCL: The Following Conditions:
 Type 1 DM, Gestational diabetes, Secondary diabetes, ESRD, HIV
 Days [-180, 0]

EXCL: <65 years of age
 Days [-180,0]

EXCL: Missing age or gender
 Days [-180,0]

Covariate Assessment Window
 Days [-180, 0]



OUTCOME(S):
 Composite cardiovascular endpoint (MACE + HHF + Death)

Follow Up Window: Days [1, follow-up end^a]

Day 0

Time

^a Follow-up end (ITT – 1 through 5, as treated – 1 through 7)

1. Occurrence of outcome
2. Disenrollment
3. Death
4. 365 days after CED
5. Calendar time reached (Dec. 31st, 2019)
6. As treated: Treatment arm switch
7. As treated: Discontinuation of study drug (30-day grace and exposure window)

Operationalization of Plasmode

Complex COI Datasets

Patient ID	Treatment	Age	Gender	HbA1c	...
1	SGLT2i	72	M	6	...
2	DPP4i	68	M	7	...
3	DPP4i	66	F	7.5	...
4	DPP4i	78	M	9	...
5	SGLT2i	87	M	8.8	...
6	SGLT2i	77	F	7.2	...
...
1,498	DPP4i	69	M	7.7	...

Model empirical associations

1. Time to outcome (MACE) $\sim \alpha_Y * Treatment + \theta_Y * COI + \sum_{k=1}^n \beta_{Yk} x_k$
 2. Time to censoring $\sim \alpha_{(Y-1)} * Treatment + \theta_{Y-1} * COI + \sum_{k=1}^n \beta_{(Y-1)k} x_k$

Use parameter estimates to write event-free survival and censoring-free survival functions; true **null treatment effect** is introduced

Stratified resampling with replacement (100 datasets per COI)

1. $S_Y(t) = h_{0Y}(t) * e^{0 * Treatment + \widehat{\theta}_Y * HbA1c_i + \widehat{\beta}_Y * x_i}$
 2. $S_{Y-1}(t) = h_{0(Y-1)}(t) * e^{0 * Treatment + \widehat{\theta}_{Y-1} * HbA1c_i + \widehat{\beta}_{Y-1} * x_i}$

Simulate outcome

Plasmode simulated dataset (1-100)

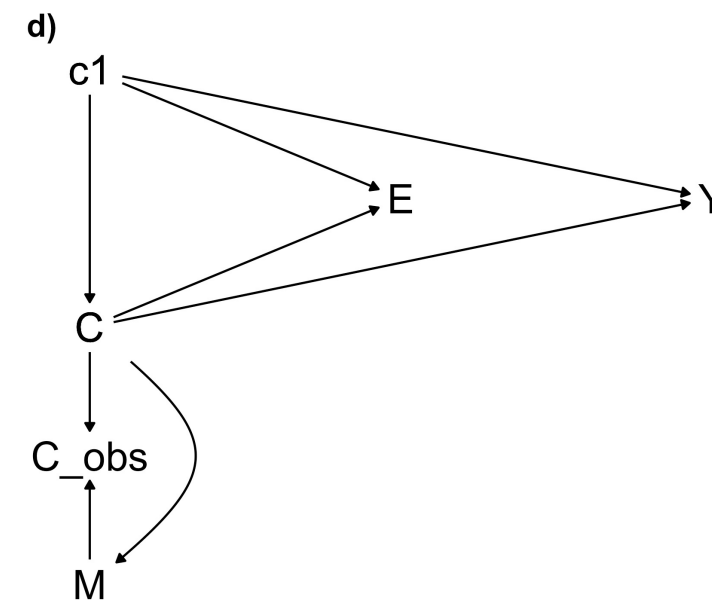
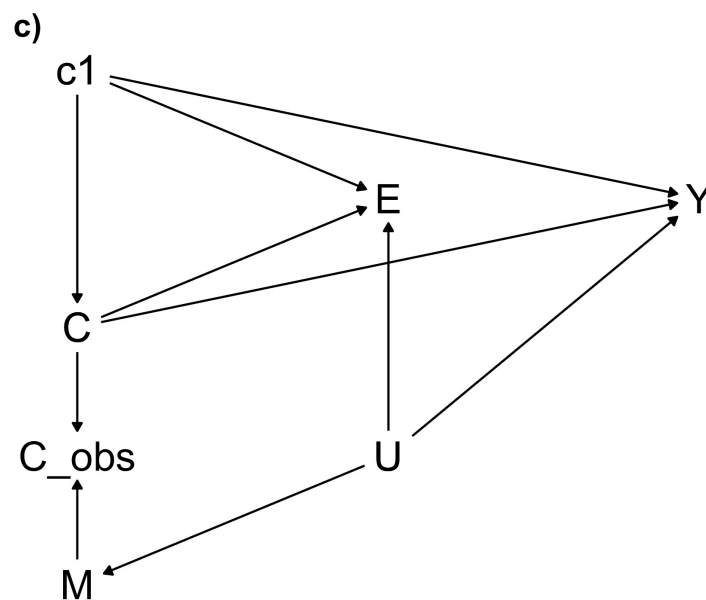
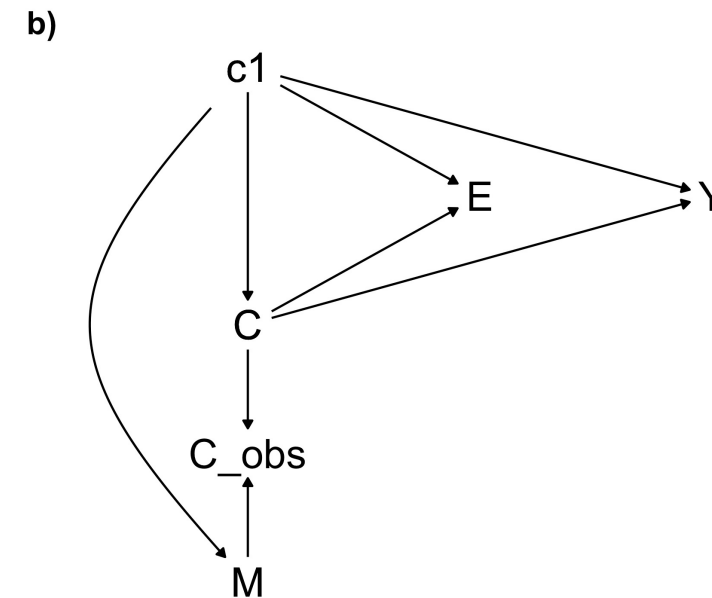
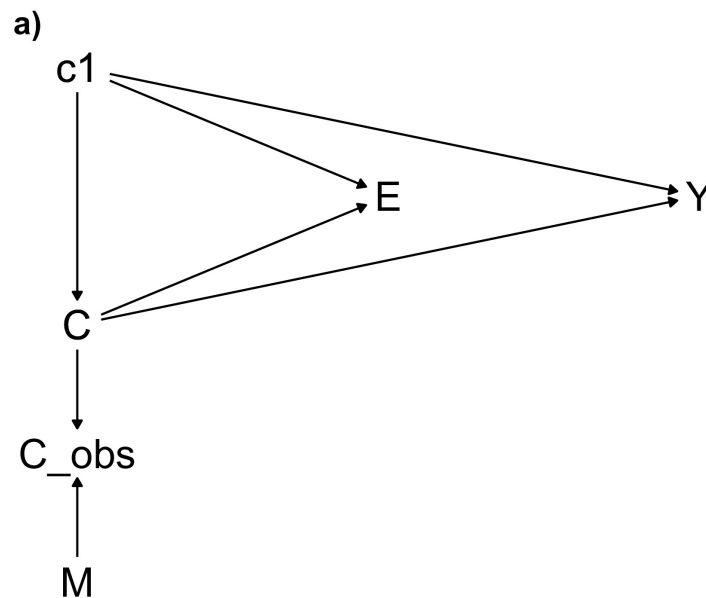
Patient ID	Treatment	Age	Gender	HbA1c	...	Event time	Censoring time	Outcome
1	SGLT2i	72	M	6
2	DPP4i	68	M	7
2	DPP4i	68	M	7
4	DPP4i	78	M	9
5	SGLT2i	87	M	8.8
5	SGLT2i	87	M	8.8
...
1,498	DPP4i	69	M	7.7

COI = confounder of interest (HbA1c, BMI, smoking)

Assumed causal missingness structures

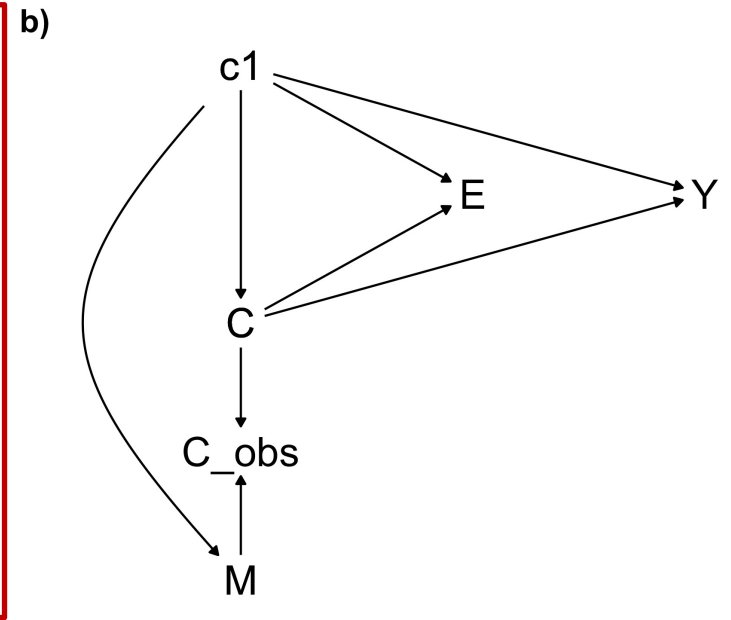
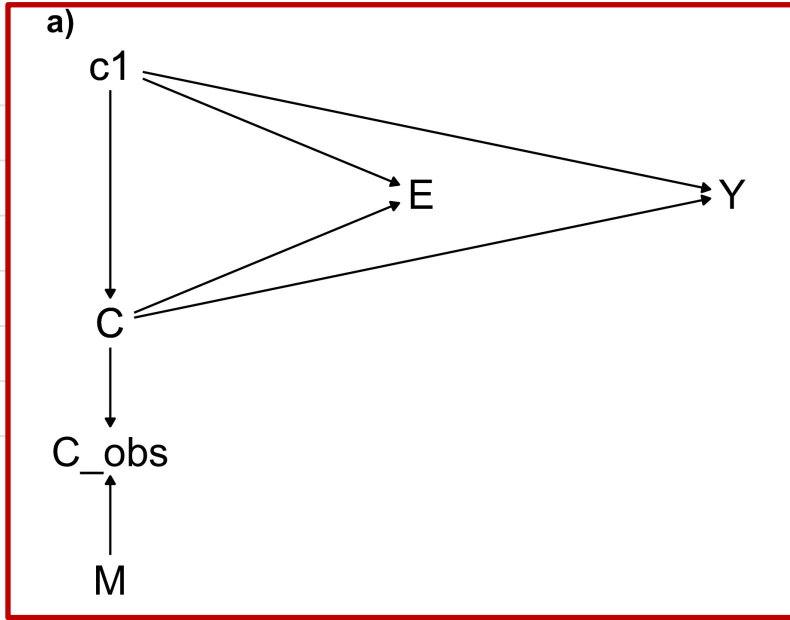
Causal diagrams/M-graphs provide a more natural way to understand the assumptions regarding missing (**confounder**) data for a given research question

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder



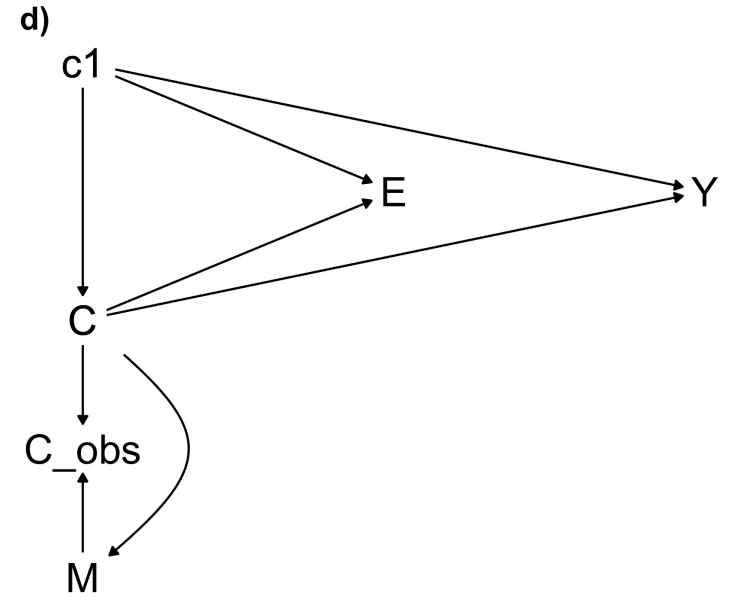
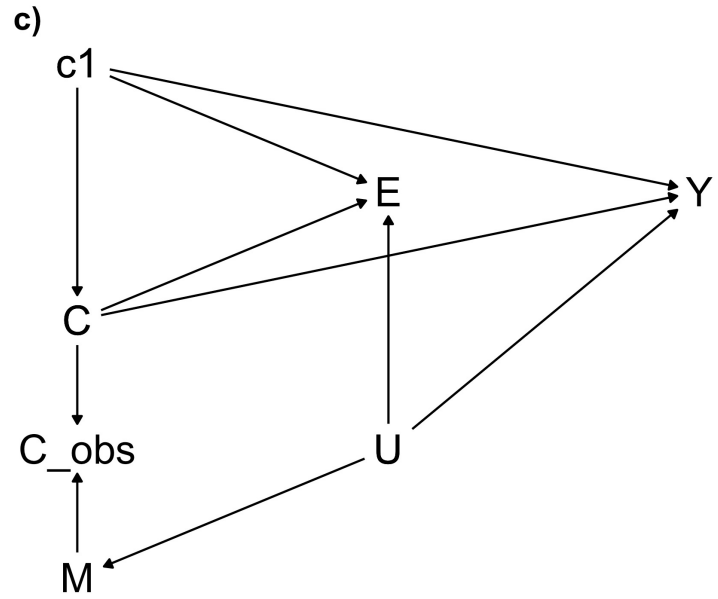
a) MCAR, b) MAR, c) MNAR (unmeasured), d) MNAR (value)

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder



MCAR:

- Confounder is randomly set to missing



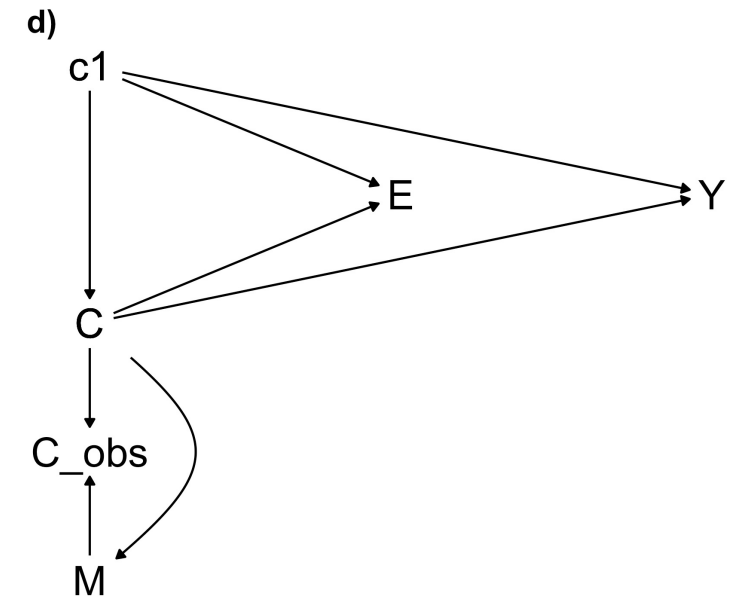
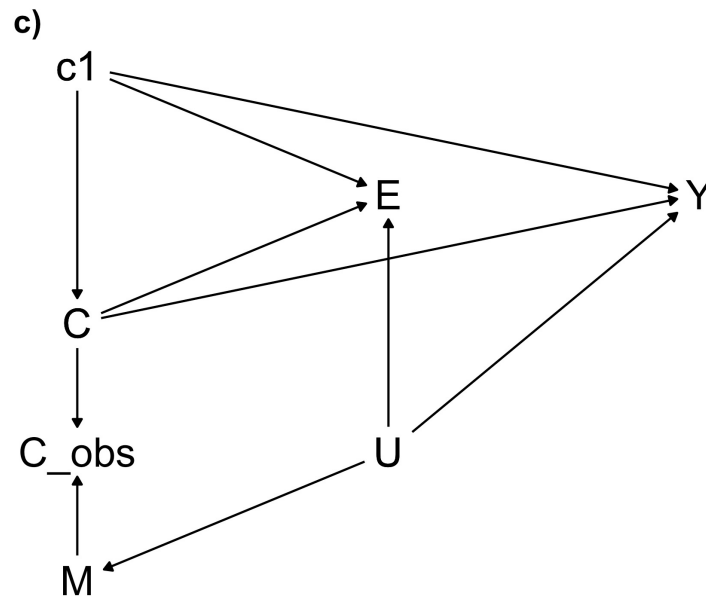
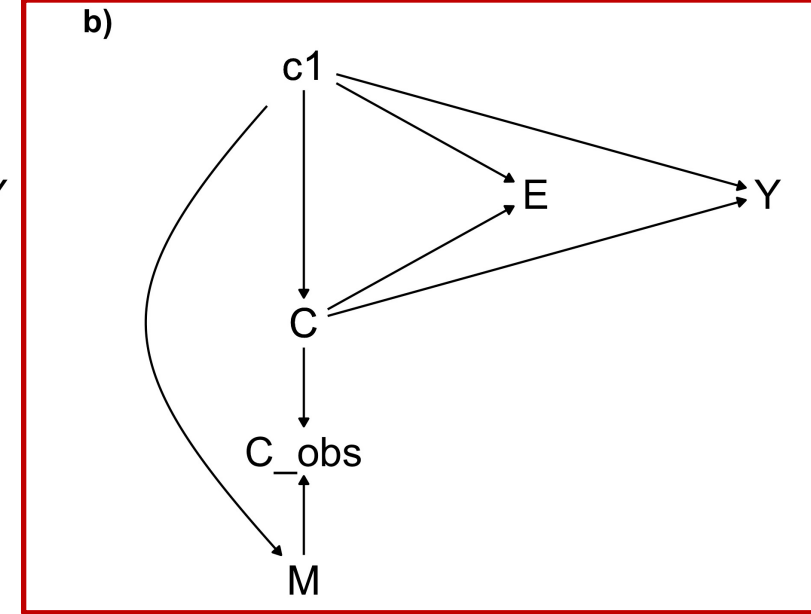
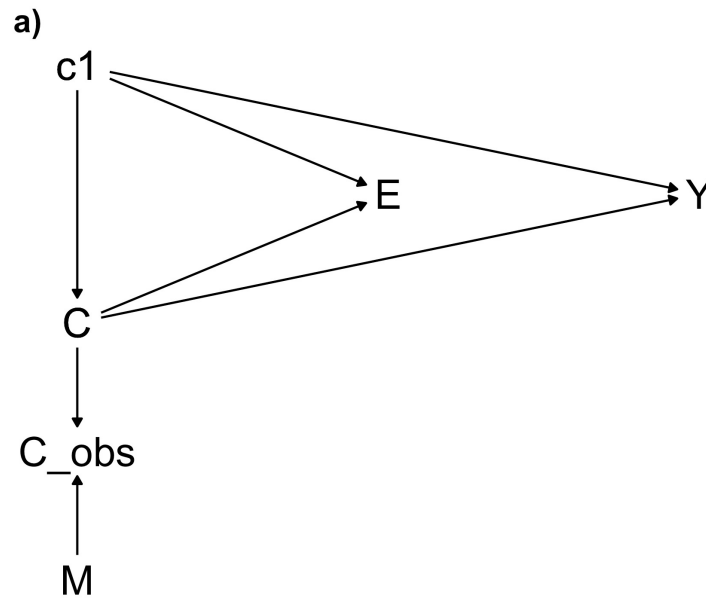
a) MCAR, b) MAR, c) MNAR (unmeasured), d) MNAR (value)

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder

MAR:

- Probability of confounder of interest to being set missing depends on weights/weighted sum scores (wss)
- Patients with high wss will have a larger probability of becoming NA
- Wss is determined by coefficients of C_1 covariates using a linear regression model
- In this simulation: every C_1 covariate has the same influence

* C_1 covariates were used in plasmode outcome generating model = true confounders

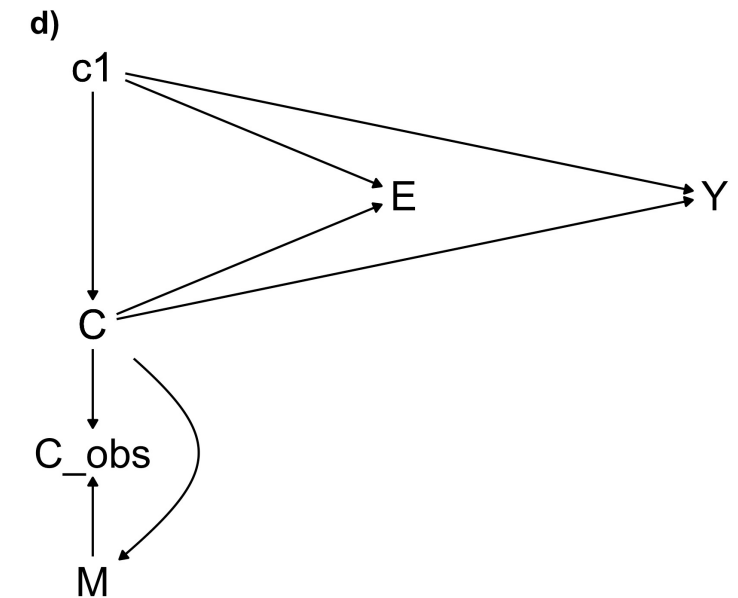
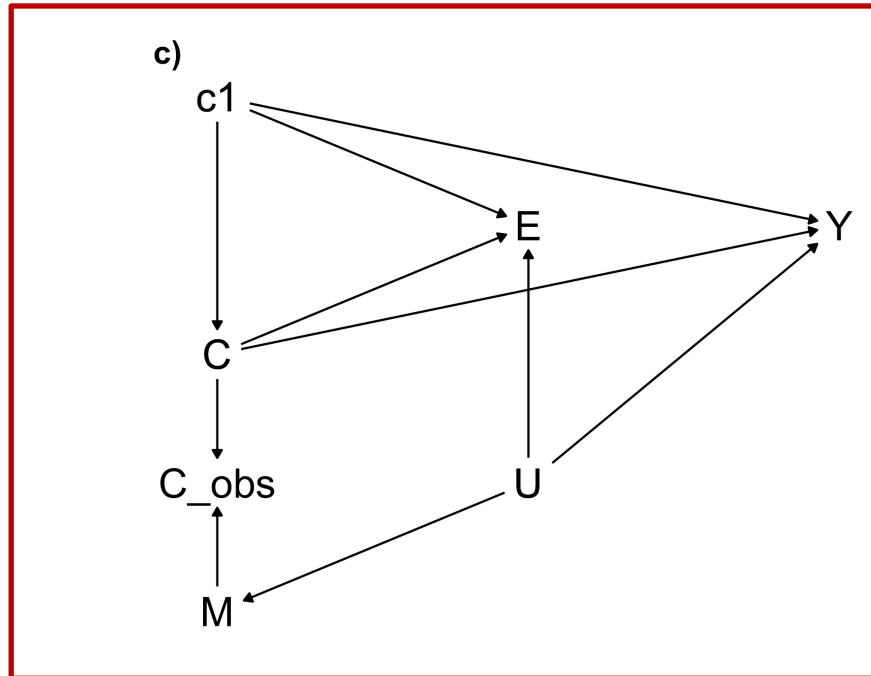
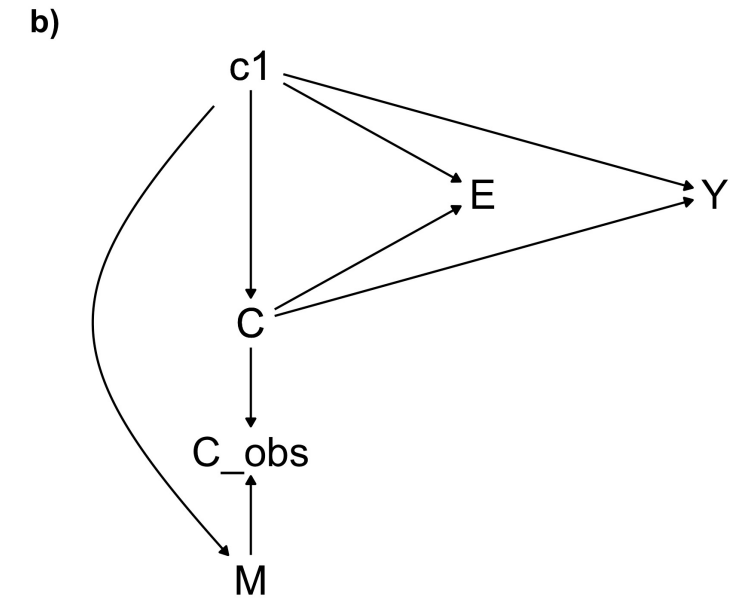
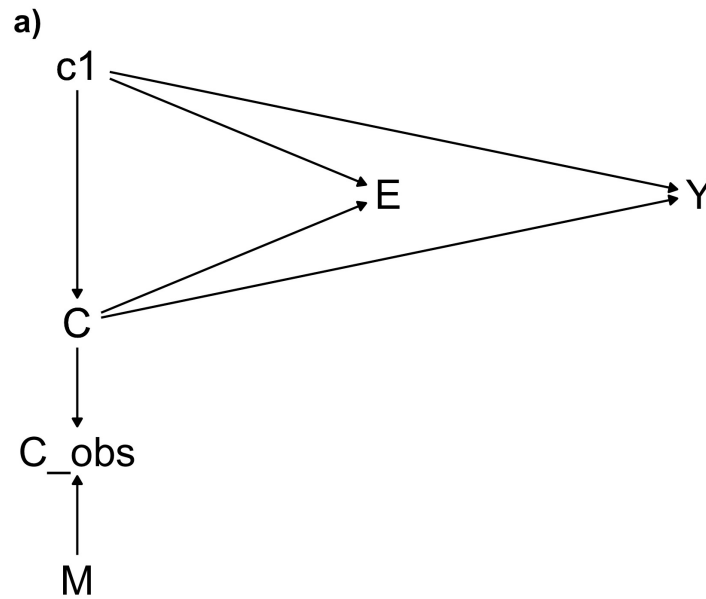


a) MCAR, b) MAR, c) MNAR (unmeasured), d) MNAR (value)

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder

MNAR (unmeasured):

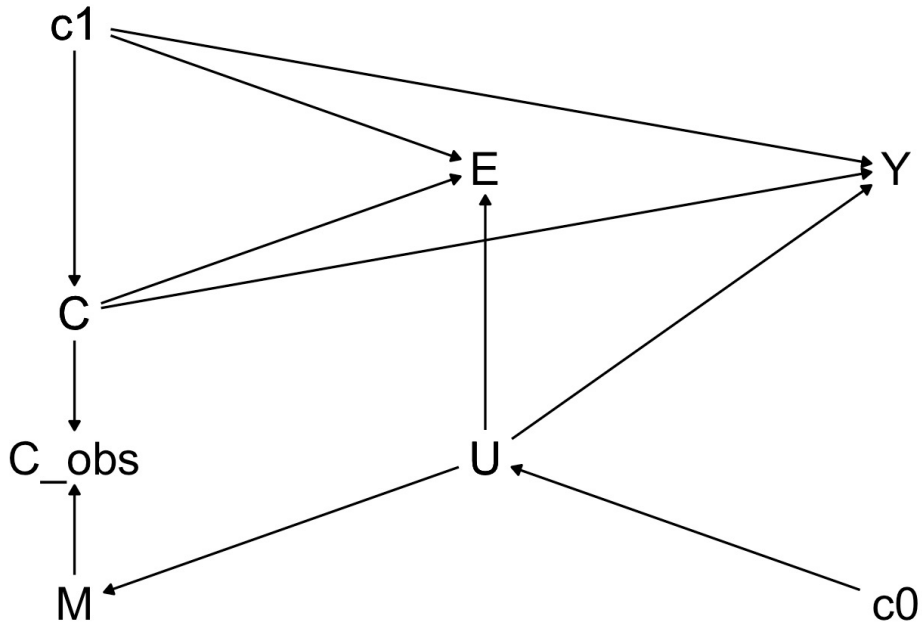
- The value of a true confounder U (age) is highly correlated with the probability for being observed with the confounder of interest {HbA1c, BMI, smoking}
- Age is used as a linear predictor to introduce missingness and subsequently dropped for all diagnostic and imputation approaches
- As age is a true confounder, the resulting missingness is not at random and very likely differential



a) MCAR, b) MAR, c) MNAR (unmeasured), d) MNAR (value)

Auxiliary covariate vector (C_0 vector)

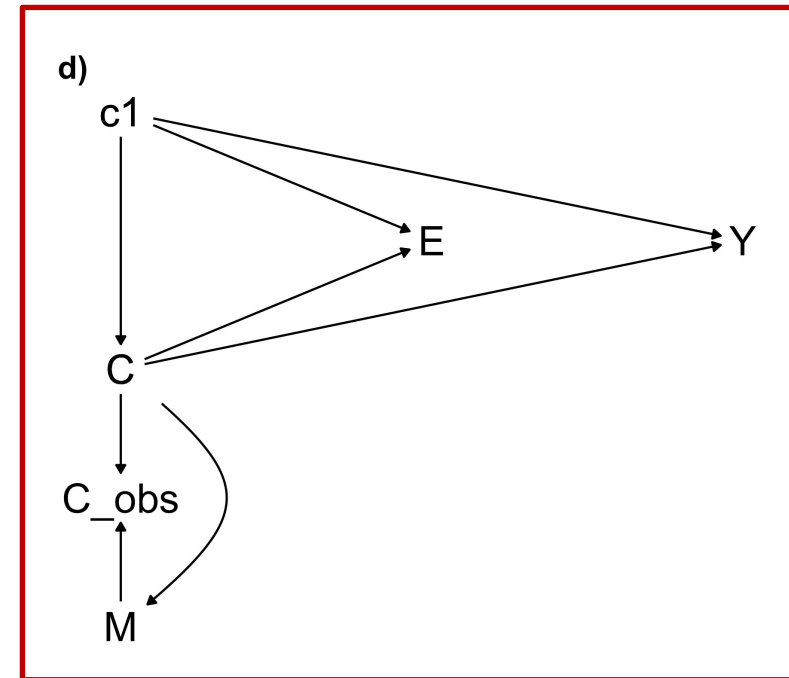
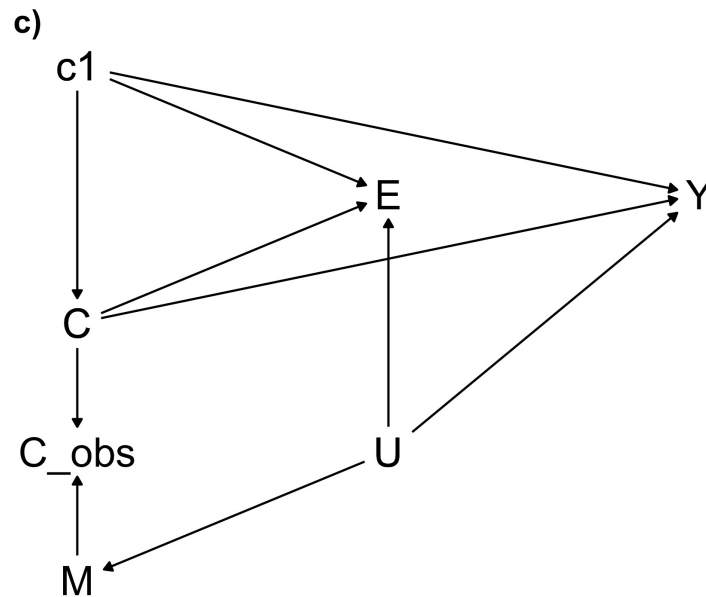
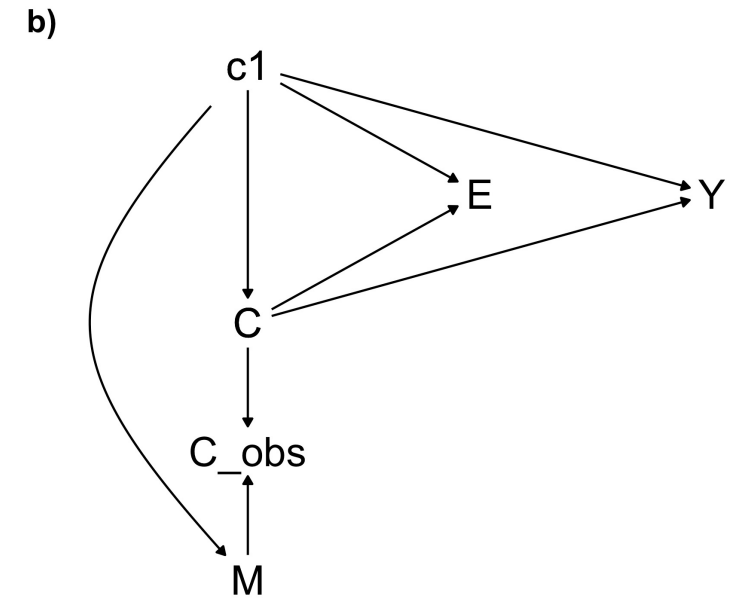
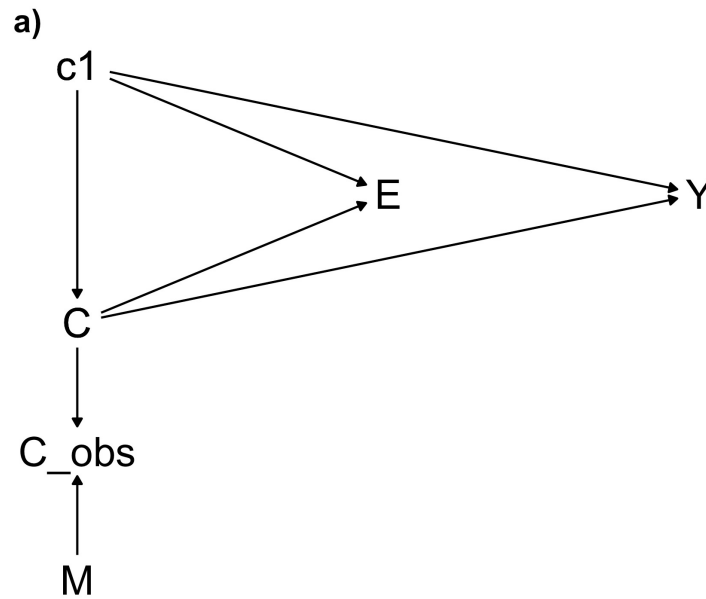
Co covariate vector: 62 covariates that are not associated with the outcome but may be associated with U, the covariates in the C1 vector and/or the EHR-derived confounder of interest and hence may be used as auxiliary variables to increase the efficiency of the imputation model. Examples are diagnostic codes for smoking, obesity and COPD.



Addition of auxiliary variables (C_0) may be particularly beneficial for scenario c)

Variable	Variable label	Variable type
rx_prior_dm_sulfonylureas	History of prior sulfonylureas use (baseline period)	binary
rx_current_dm_GLP1	Concurrent GLP1 use	binary
rx_current_dm_insulin	Concurrent insulin use	binary
rx_current_dm_metformin	Concurrent metformin use	binary
rx_antibiotics	History of antibiotics use	binary
rx_estrogen	History of estrogen use	binary
rx_oral_corticoids	History of oral corticosteroid use	binary
dx_cancer	History of cancer	binary
dx_fungal_infection	History of fungal infection	binary
dx_immune_infection	History of immune infection	binary
dx_urinary_infection	History of urinary infection	binary
colonoscopy	Colonoscopy procedure	binary
fecal_blood	Fecal blood test	binary
flu_shot	Flu shoot	binary
mammography	Mammography	binary
pap_smear	Pap smear test	binary
pneumoc_vaccine	Pneumococcus vaccine	binary
prostate	Prostate exam	binary
num_microalbuminuria	Number of microalbuminuria (CPT4 procedures)	continuous
num_creatinine	Number of creatinine tests	continuous
visits30_internal	Number of internal physician visits in 30 days prior index date	continuous
visits30_endo	Number of endocrinologist visits in 30 days prior index date	continuous
visits180_endo	Number of endocrinologist visits in 180 days prior index date	continuous
dx_alzheimer	History of alzheimer disease	binary
dx_copd	History of COPD	binary
dx_dementia	History of Dementia	binary
dx_depression	History of depression	binary
dx_htn_nephropathy	History of hypertensive nephropathy	binary
dx_hyperkalemia	History of hyperkalemia	binary
dx_hypertension	History of hypertension	binary
dx_hypotension	History of hypotension	binary
dx_obesity	History of obesity (diagnostic codes)	binary
dx_oth_dysrhythmia	History of other dysrhythmia	binary
dx_psychosis	History of psychosis	binary
dx_pulmonary_htn	History of pulmonary hypertension	binary
dx_pvd	History of PVD	binary
dx_sleep_apnea	History of sleep apnea	binary
dx_smoking	History of smoking (diagnostic codes)	binary
dx_stable_angina	History of stable angina	binary

E	Exposure/treatment
Y	Outcome
C	Confounder of interest
C_obs	Observed portion of C
M	Missingness of C (M=0 fully observed and M=1 fully missing)
c1	Covariates associated with outcome and missingness
c0	Auxiliary covariates
U	Unmeasured covariate/confounder



MNAR(value):

- The value of the confounder of interest itself is used as a linear predictor
- As, in consequence, the information about the missing data is missing itself, it is very hard to predict and impute the missing confounder of interest

Table. Illustration of Simulation Parameters.

Analysis element	Parameters altered in simulation
Exemplary partially observed EHR confounders of interest (type)	<ul style="list-style-type: none"> ● HbA1c value in % (continuous) ● Body mass index [BMI], underweight/normal, overweight, obese (ordinal) ● Smoking, current/former vs. never (binary)
Missingness mechanism	<ul style="list-style-type: none"> ● MCAR ● MAR ● MNAR_{unmeasured} ● MNAR_{value}
Degree of missingness	10% - 50% (incremental increases by 10%)
Strength of confounders	As empirically observed in dataset
Ad-hoc & Imputation methods	<ul style="list-style-type: none"> ● Complete case analysis (baseline method) ● Inverse probability of missingness weighting (IPMW) ● Missingness indicator ● missForest (single imputation random forest algorithm) ● <u>Multiple imputation (m = 5 imputed datasets each):</u> <ul style="list-style-type: none"> ▪ MICE defaults for variable types: <ul style="list-style-type: none"> ○ Predictive mean matching (<i>PMM</i>) (HbA1c) ○ Proportional odds model (<i>Polr</i>) (BMI) ○ Logistic regression (<i>Logreg</i>) (smoking) ▪ Classification and regression tree (CART) ▪ Random forest (RF)
Provided covariates for diagnostics and imputation	<ul style="list-style-type: none"> ● C₁ variables (= true confounders/variables used for outcome generation) ● C₁ variables + C₀ variables as auxiliary variables ● + Outcome (i.e. time-to-event and event indicator)
Modifications to causal effect estimation	<ul style="list-style-type: none"> ● +/- treatment effect modification by EHR confounder of interest ● +/- addition of missingness indicator variable in imputation and outcome model
Sampled cohort size (number of datasets)	1000 patients (100 datasets each)

Diagnostics

Group 1 Diagnostics

Group 2 Diagnostics

Group 3 Diagnostics

	Absolute standardized mean difference (ASMD)	P-value Hotelling/Little	AUC (are under the receiver operating curve)	Log HR (missingness indicator)
Purpose	Averaged median SMD of covariate distributions between patients with vs w/o observed confounder (across all observed covariates)	Little chi-squared test statistic assuming that if the missingness is MCAR, then conditional on the missing indicator, the null hypothesis that there are no differences between the means of different missing-value patterns will hold. Hotelling multivariate t-test with same purpose but one variable at a time	If missing indicator can be predicted as a function of observed covariates, MAR may be a likely scenario and would imply that imputation may be feasible	Fitting an outcome model with the missingness indicator <u>crude and conditional on all other prognostic covariates</u> would indicate a meaningful difference in the outcome between patients with vs w/o the observed confounder conditional on other covariates that could explain that difference.
Example value	0.025	p-value <0.001	0.8	log HR 0.1 (95% CI 0.05, 0.2)
Interpretation	<0.1: missingness is not associated with other observed covariates may be completely at random >0.1: missingness differs between patients and observed covariates can explain difference	High test statistics and low p-values would be indicative for differences in covariate distributions and null hypothesis would be rejected (\neq MCAR)	Values around 0.5 indicate random prediction (MCAR) Values meaningfully above 0.5 indicate stronger correlations between covariates (which can be determined!) and missingness (\sim MAR)	MCAR: No association in neither crude nor adjusted model MAR: Association in crude but not adjusted model MNAR: If there was a meaningful difference also after comprehensive adjustment (log HR) , this may be indicative of differential MNAR scenarios

Imputation Metrics

	Root mean square error (RMSE)	Coverage	Width	% Bias
Purpose	Estimation error: compromise between bias and variance, and evaluates the treatment effect estimate on both accuracy and precision based on the imputed data	Proportion of confidence intervals of the imputation method that contain the true estimate	The average width of the confidence interval	Average amount that actual is greater than predicted as a percentage of the absolute value of actual. The percent bias is calculated by taking the average of (actual - predicted) / abs(actual) across all observations.
Example value	0.135	0.96 (96%)	0.139	.125 (12.5%)
Interpretation	The lower the better	<p>< 90 percent (for a nominal 95 percent interval) indicates poor quality</p> <p>A high CR (e.g., 0.99) may indicate that confidence interval is too wide, so the imputation method is inefficient and leads to inferences that are too conservative. Inferences that are “too conservative” are generally regarded a lesser sin than “too optimistic”.</p>	Indicator of statistical efficiency. The length should be as small as possible , but not so small that the CR will fall below the nominal level.	If a model is unbiased, the % bias percent_bias should be close to zero. For acceptable performance we use an upper limit for PB of 5%. (Demirtas, Freels, and Yucel 2008)

Diagnostic Results Across All HbA1c, BMI and Smoking Cohorts

- Overall results, averaged across all scenarios and simulated HbA1c, BMI and smoking plasmode cohorts
- = total 48,000 plasmode datasets | 12,000 per missingness mechanism)

Missingness diagnostics results overall.

Mechanism	ASMD (95% CI)*	p(Hotelling)	p(Little)	AUC (95% CI)*	log HR(crude) (95% CI)	log HR (95% CI)
MCAR	0.05 (0.05-0.05)	0.50	0.50	0.50 (0.50-0.50)	-0.01 (-0.33-0.32)	0.00 (-0.36-0.36)
MAR	0.20 (0.20-0.20)	<.001	<.001	0.58 (0.58-0.59)	0.53 (0.23-0.83)	0.00 (-0.37-0.37)
MNAR(unmeasured)	0.09 (0.09-0.09)	0.02	0.02	0.54 (0.54-0.54)	0.43 (0.13-0.74)	0.31 (-0.03-0.66)
MNAR(value)	0.06 (0.06-0.06)	0.10	0.10	0.53 (0.53-0.53)	0.04 (-0.27-0.36)	0.10 (-0.26-0.45)

ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, CI = Confidence interval

* Confidence intervals are computed based on empirical standard errors.

Group 1 diagnostics

Group 2 diagnostics

Group 3 diagnostics

Diagnostic Results by HbA1c, BMI and Smoking

- Averaged by confounder of interest
- = 4,000 datasets per missingness mechanism and confounder of interest

Missingness diagnostics results by EHR confounder of interest.

Mechanism	ASMD (95% CI)*	p(Hotelling)	p(Little)	AUC (95% CI)*	log HR(crude) (95% CI)	log HR (95% CI)
HbA1c						
MCAR	0.05 (0.05-0.05)	0.50	0.50	0.50 (0.50-0.50)	-0.01 (-0.30-0.29)	-0.01 (-0.33-0.32)
MAR	0.21 (0.20-0.21)	<.001	<.001	0.59 (0.58-0.59)	0.52 (0.25-0.80)	0.00 (-0.34-0.33)
MNAR(unmeasured)	0.10 (0.09-0.10)	0.02	0.02	0.54 (0.54-0.54)	0.42 (0.15-0.70)	0.32 (0.01-0.63)
MNAR(value)	0.07 (0.07-0.07)	0.09	0.09	0.53 (0.53-0.53)	0.05 (-0.24-0.35)	0.12 (-0.21-0.46)
Body Mass Index						
MCAR	0.05 (0.05-0.05)	0.50	0.50	0.50 (0.50-0.50)	-0.01 (-0.35-0.34)	0.00 (-0.39-0.38)
MAR	0.19 (0.19-0.19)	<.001	<.001	0.58 (0.58-0.58)	0.50 (0.18-0.82)	0.00 (-0.40-0.39)
MNAR(unmeasured)	0.09 (0.09-0.09)	0.03	0.03	0.54 (0.54-0.54)	0.45 (0.12-0.77)	0.31 (-0.06-0.68)
MNAR(value)	0.06 (0.06-0.06)	0.17	0.17	0.53 (0.53-0.53)	-0.01 (-0.35-0.34)	-0.02 (-0.41-0.38)
Smoking						
MCAR	0.05 (0.05-0.05)	0.50	0.50	0.50 (0.50-0.50)	0.00 (-0.33-0.32)	0.00 (-0.37-0.36)
MAR	0.20 (0.20-0.20)	<.001	<.001	0.59 (0.58-0.59)	0.57 (0.28-0.87)	0.00 (-0.37-0.37)
MNAR(unmeasured)	0.09 (0.09-0.09)	0.02	0.02	0.54 (0.54-0.54)	0.43 (0.12-0.74)	0.31 (-0.04-0.66)
MNAR(value)	0.07 (0.07-0.07)	0.02	0.02	0.54 (0.54-0.54)	0.08 (-0.21-0.38)	0.18 (-0.16-0.52)

ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, CI = Confidence interval

* Confidence intervals are computed based on empirical standard errors.

Diagnostic Results Across all HbA1c, BMI and Smoking Cohorts

- Averaged by proportion missing / = 2,400 datasets per missingness mechanism and proportion

Missingness diagnostics results by proportion of missingness.

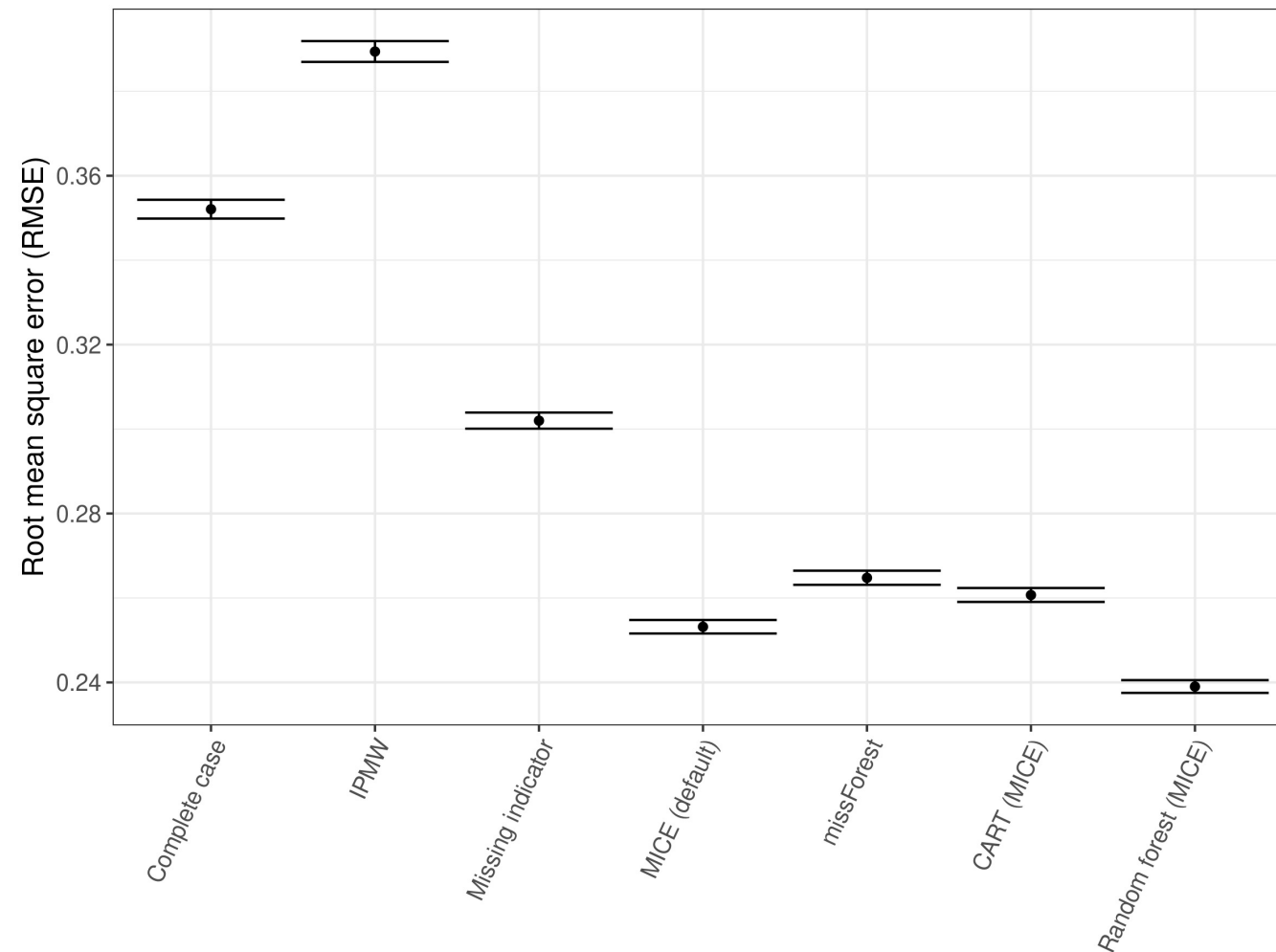
Mechanism	ASMD (95% CI)*	p(Hotelling)	p(Little)	AUC (95% CI)*	log HR(crude) (95% CI)	log HR (95% CI)
Proportion (missing) 10%						
MCAR	0.07 (0.07-0.07)	0.51	0.51	0.50 (0.50-0.50)	-0.01 (-0.46-0.44)	0.00 (-0.51-0.50)
MAR	0.22 (0.22-0.23)	<.001	<.001	0.53 (0.53-0.53)	0.61 (0.23-0.98)	0.00 (-0.46-0.46)
MNAR(unmeasured)	0.11 (0.11-0.11)	0.05	0.06	0.51 (0.51-0.52)	0.49 (0.09-0.88)	0.36 (-0.09-0.80)
MNAR(value)	0.08 (0.08-0.09)	0.16	0.16	0.52 (0.52-0.52)	0.06 (-0.36-0.48)	0.12 (-0.35-0.59)
Proportion (missing) 20%						
MCAR	0.05 (0.05-0.05)	0.49	0.48	0.50 (0.50-0.50)	0.00 (-0.33-0.33)	0.00 (-0.37-0.38)
MAR	0.20 (0.20-0.21)	<.001	<.001	0.56 (0.56-0.56)	0.56 (0.26-0.86)	0.00 (-0.37-0.38)
MNAR(unmeasured)	0.10 (0.10-0.10)	0.02	0.02	0.53 (0.52-0.53)	0.45 (0.15-0.76)	0.33 (-0.02-0.68)
MNAR(value)	0.07 (0.07-0.07)	0.10	0.10	0.53 (0.53-0.53)	0.05 (-0.27-0.37)	0.10 (-0.26-0.47)
Proportion (missing) 30%						
MCAR	0.05 (0.05-0.05)	0.50	0.50	0.50 (0.50-0.50)	-0.01 (-0.30-0.29)	-0.01 (-0.33-0.32)
MAR	0.19 (0.19-0.20)	<.001	<.001	0.59 (0.59-0.59)	0.51 (0.24-0.79)	-0.01 (-0.35-0.33)
MNAR(unmeasured)	0.09 (0.09-0.09)	0.01	0.01	0.54 (0.54-0.54)	0.43 (0.15-0.71)	0.31 (-0.01-0.63)
MNAR(value)	0.06 (0.06-0.06)	0.08	0.08	0.54 (0.53-0.54)	0.04 (-0.25-0.33)	0.09 (-0.23-0.42)
Proportion (missing) 40%						
MCAR	0.04 (0.04-0.04)	0.50	0.50	0.50 (0.50-0.50)	0.00 (-0.27-0.27)	0.00 (-0.30-0.30)
MAR	0.19 (0.19-0.19)	<.001	<.001	0.62 (0.62-0.62)	0.49 (0.22-0.76)	-0.01 (-0.34-0.32)
MNAR(unmeasured)	0.08 (0.08-0.08)	0.01	0.01	0.55 (0.55-0.55)	0.41 (0.14-0.68)	0.29 (-0.01-0.60)
MNAR(value)	0.06 (0.06-0.06)	0.06	0.06	0.54 (0.54-0.54)	0.04 (-0.23-0.31)	0.09 (-0.22-0.40)
Proportion (missing) 50%						
MCAR	0.04 (0.04-0.04)	0.51	0.50	0.51 (0.51-0.51)	-0.01 (-0.27-0.26)	-0.01 (-0.31-0.29)
MAR	0.19 (0.18-0.19)	<.001	<.001	0.63 (0.62-0.63)	0.48 (0.21-0.75)	0.01 (-0.32-0.34)
MNAR(unmeasured)	0.08 (0.08-0.08)	0.02	0.02	0.56 (0.56-0.56)	0.39 (0.12-0.66)	0.28 (-0.03-0.58)
MNAR(value)	0.05 (0.05-0.05)	0.07	0.08	0.55 (0.55-0.55)	0.03 (-0.24-0.30)	0.07 (-0.23-0.38)

ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, CI = Confidence interval

* Confidence intervals are computed based on empirical standard errors.

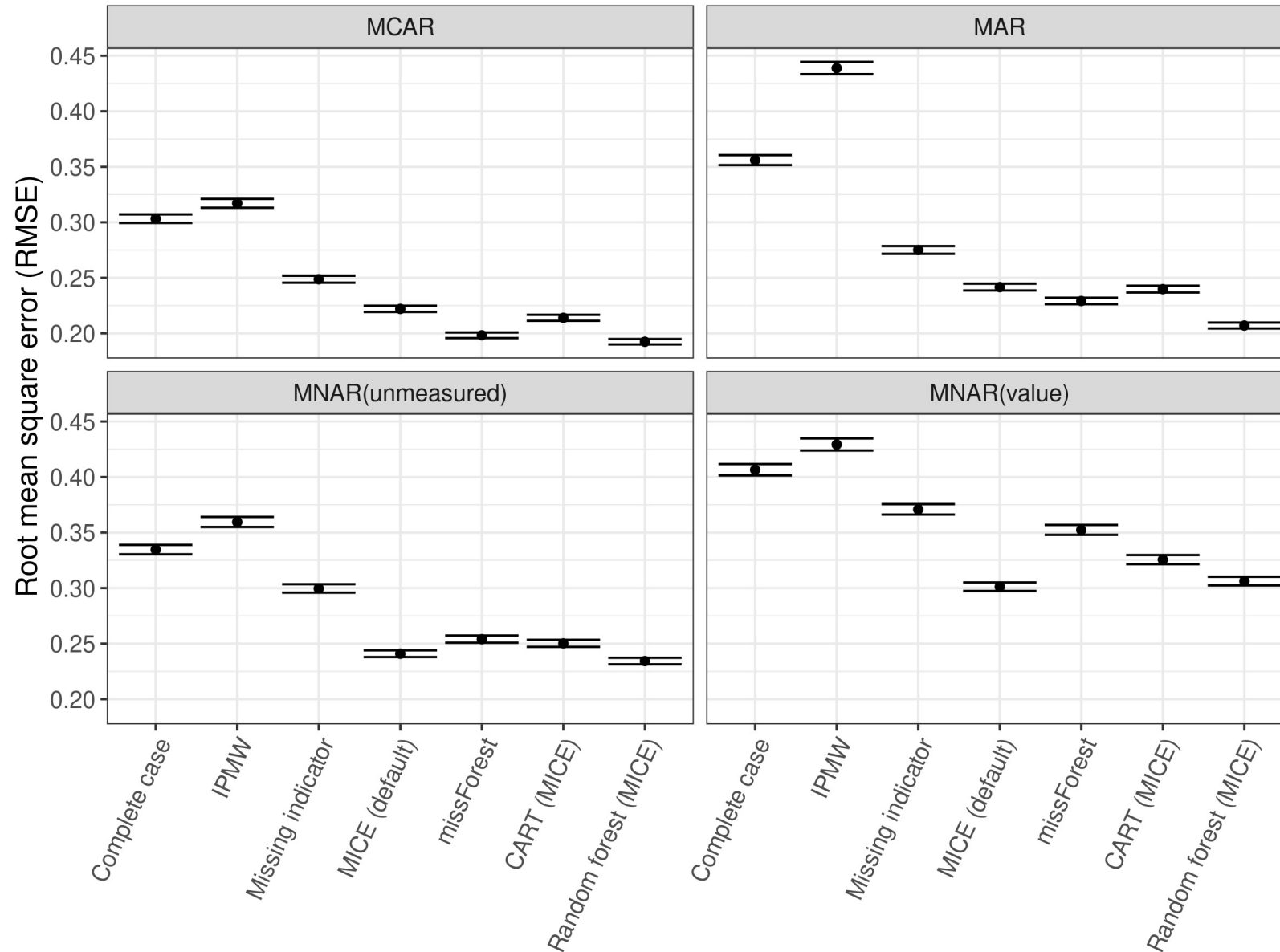
Imputation Results Across all HbA1c, BMI and Smoking Cohorts

- Overall results, averaged across all scenarios and simulated HbA1c, BMI and smoking plasmode cohorts

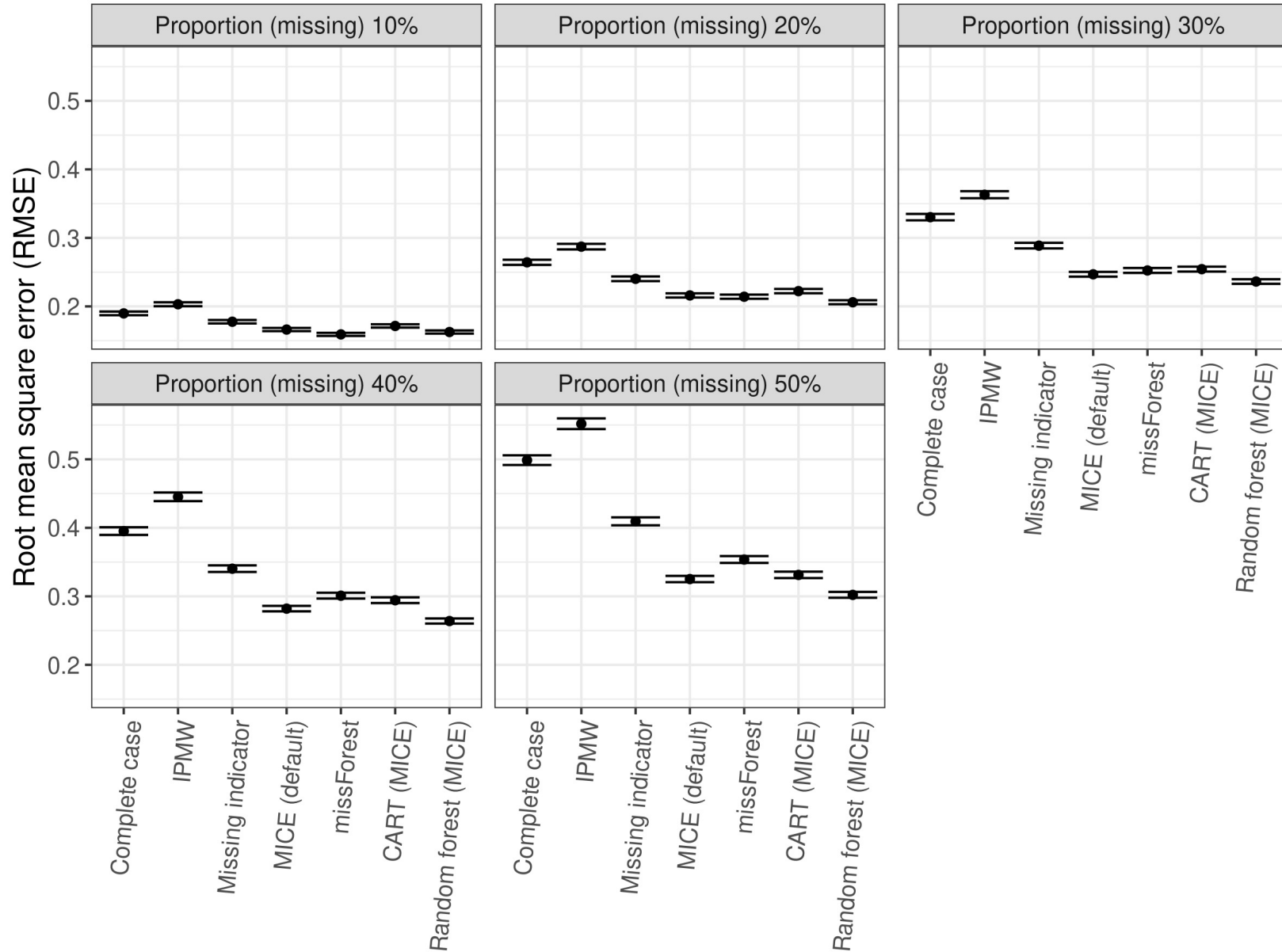


Results in line with recently published manuscript by Getz K, Hubbard RA, Linn KA. Performance of Multiple Imputation Using Modern Machine Learning Methods in Electronic Health Records Data. *Epidemiology*. 2023 Mar 1;34(2):206-215.

Imputation Results by Missing Mechanism



Imputation Results by Proportion Missing

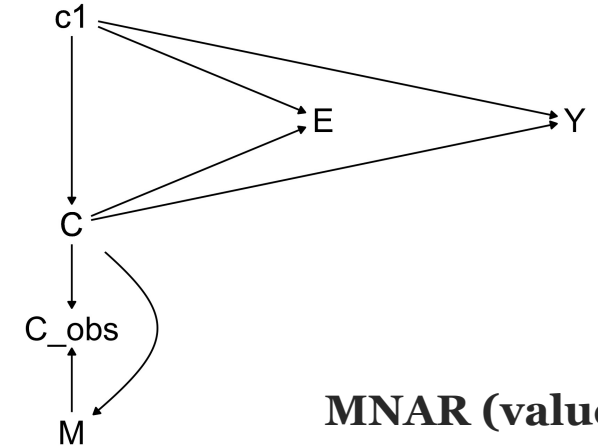


Sensitivity Analyses to Test the Robustness of Analytical Decisions – MNAR (value)

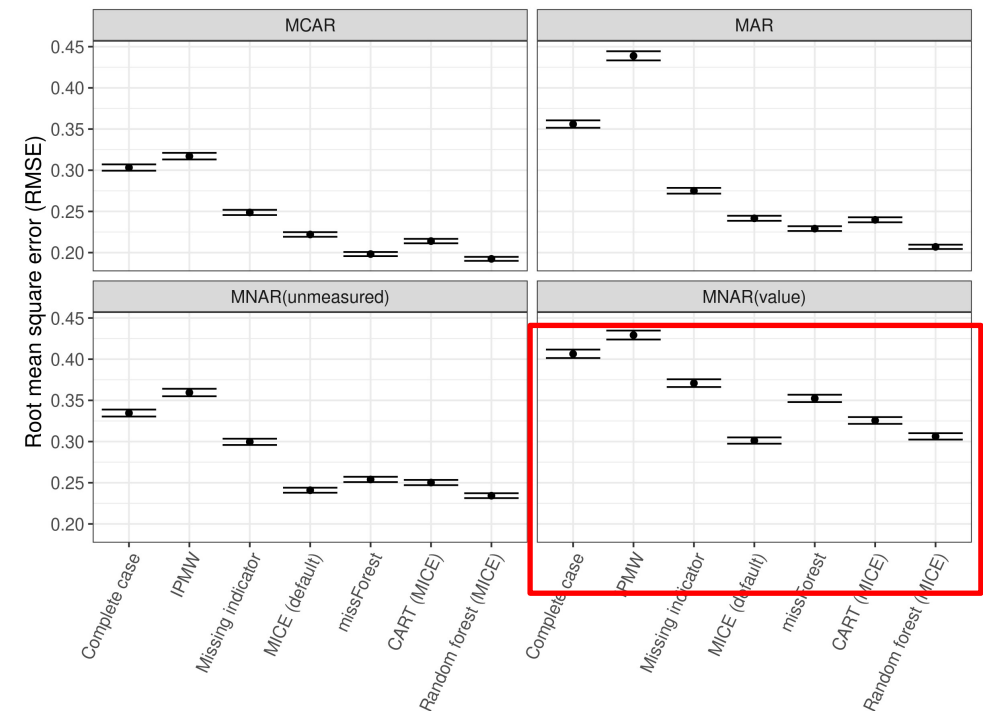
- In reality, we will not be able to know 100% what the true missingness mechanism is
- However, we have tools that enable us to mitigate bias/reach unbiased conclusions when data are MCAR or MAR or (to some extent) MNAR (unmeasured)

Residual uncertainty in case of MNAR (value):

- Hard to differentiate MCAR and MNAR (value)
- Difficult to impossible to impute the marginal distribution based on observed and auxiliary variables
- Much stronger bias



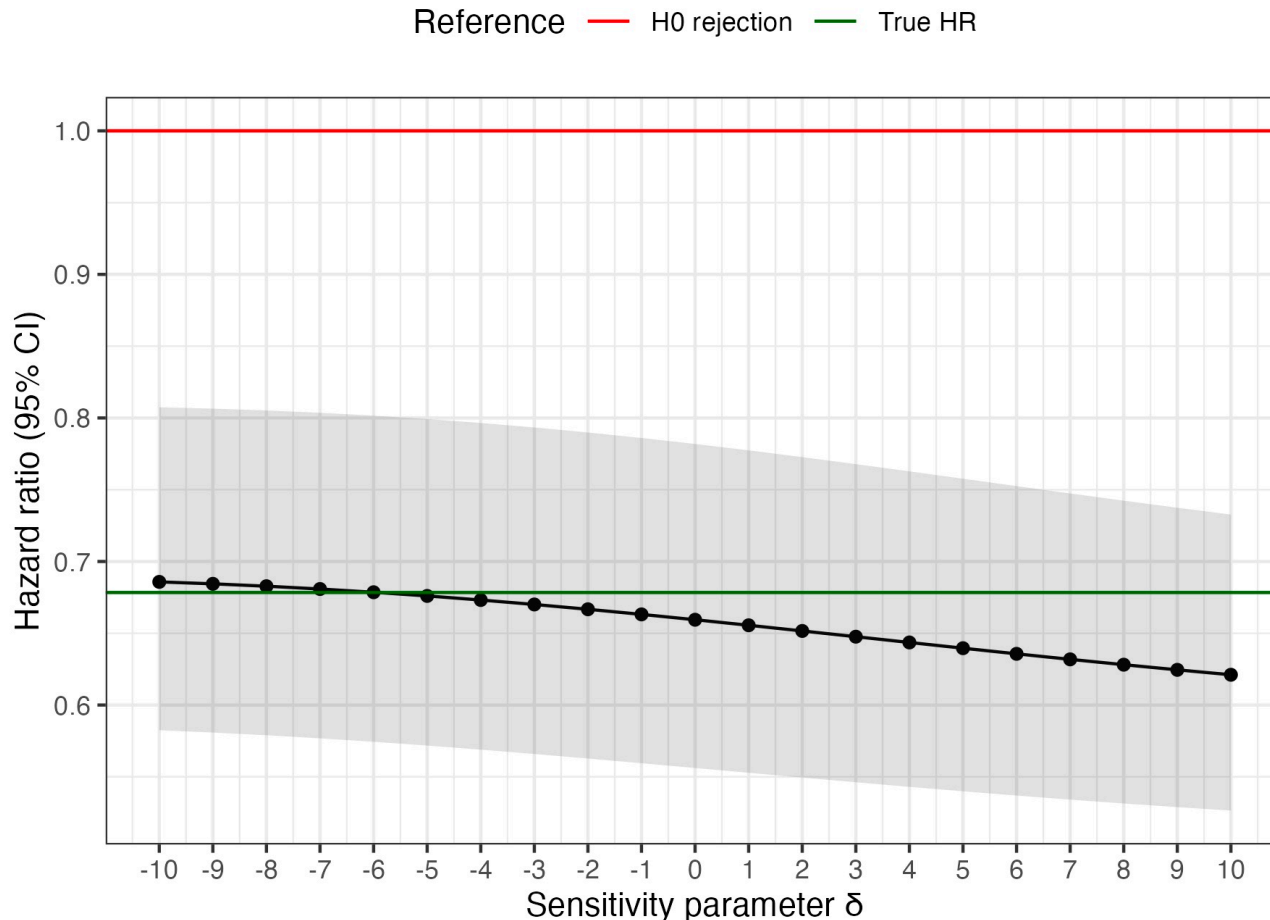
MNAR (value)



Not-at-random Fully Conditional Specification (NARFCS)

Sensitivity Analysis

NARFCS tipping point analysis - how sensitive are results to a departure from MAR?



- Range of MI analyses are run over a range of different conditional sensitivity parameters δ (x-axis)
- Corresponding effect estimates show sensitivity to potential departures from MAR
- Tipping point: δ where confidence interval would cross a pre-specified threshold and discard qualitative conclusion of main analysis
- Hard to illustrate if many variables with sensitivity parameters are modeled

Figure: Example based on simulated data illustrating an MNAR(value) scenario in which younger patients would be systematically more likely to be missing

Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Stat Med.* 2018 Jul 10;37(15):2338-2353

Toolkit - R Package

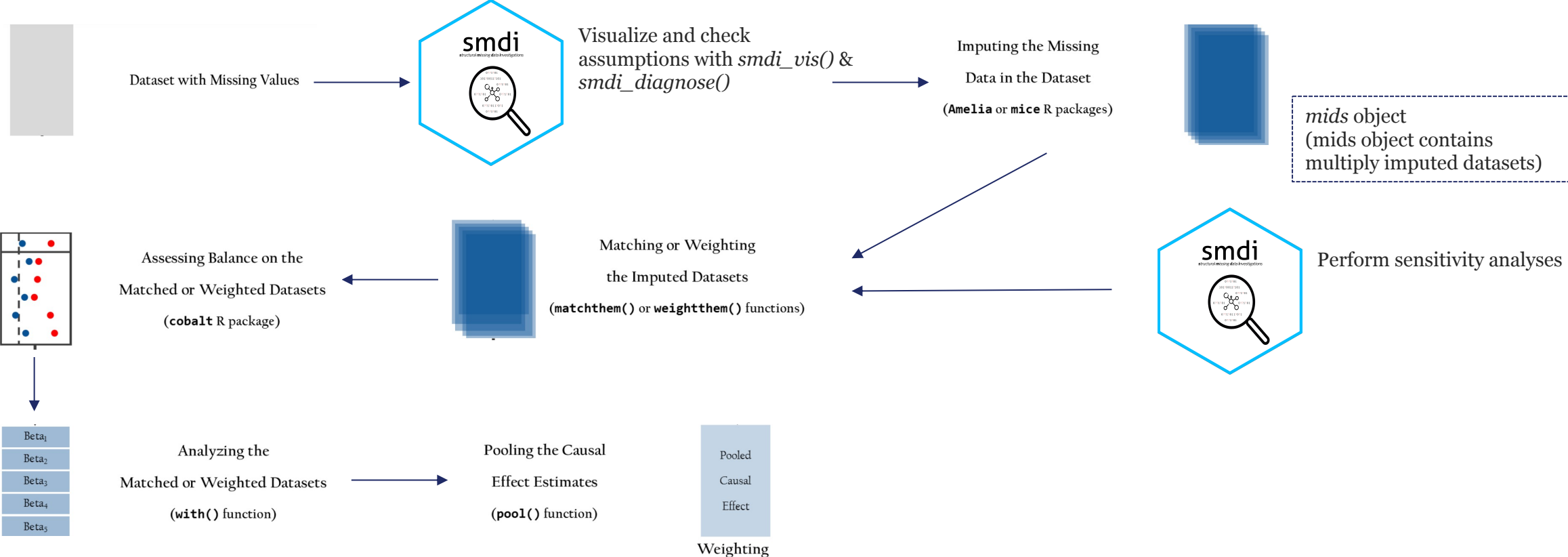
Easy implementation of **routine structural missing data investigations (smdi)**

- Selected functions (S3 method):
 - ***smdi_diagnose()*** – flagship function that will return all three group diagnostics evaluated in simulation study
 - ***smdi_summarize()*** & ***smdi_vis()*** – easy and quick visualization of proportion missingness as (variables can be specified; if not specified, all variables with NA will be displayed)
 - More...
- Duke to implement smdi toolkit as part of the validation and analysis of empirical study question



janickweberpals.gitlab-pages.partners.org/smdi

Practical Implementation, Principled Workflow and Compatibility with *Mice* and *MatchThem* R Packages



Adapted from: Pishgar, F., N. Greifer, C. Leyrat, and E. Stuart. Matchthem: matching and weighting after multiple imputation. *RJ*. 2021; 13: 292–305. doi: 10.32614. RJ-2021-073.

Acknowledgements

Mass General Brigham

- Rishi Desai
- Robert J. Glynn
- Shamika More
- Luke Zobotka

Duke

- Sudha Raman
- Brad Hammill

Kaiser Washington

- Pamela Shaw

Harvard Pilgrim/SOC

- Darren Toh
- John Connolly
- Kimberly J. Dandreo Gegear

FDA

- Fang Tian
- Wei Liu
- Hana Lee
- Jenni Li
- Jose Hernandez

Thank You

Contact: jweberpals@bwh.harvard.edu



Backup Slides

Supplementary Table 2. Overview of C1 covariates.

Variable	Variable label	Variable type
age	Age at index date	continuous
male	Gender (male)	binary
subsidy	Medicare subsidy flag	continuous
combined_score	Combined comorbidity score	continuous
generics	Number of generics used	continuous
rx_anticoagulant	History of anticoagulant use	binary
rx_antiplatelets	History of antiplatelets use	binary
rx_statin	History of statins use	binary
rx_antihypertensive	History of antihypertensive use	binary
rx_current_dm_sulfonylureas	Concurrent sulfonylureas use	binary
num_diab_meds_on_index	Number of unique DM generics on/overlapping index date	continuous
num_hba1c_test	Number of HbA1c tests	continuous
visits180_internal	Number of internal physician visits in 180 days prior index date	continuous
dx_afib	History of atrial fibrillation	binary
dx_anemia	History of Anemia	binary
dx_cardiomyopathy	History of cardiomyopathy	binary
dx_diab_nephropathy	History of diabetic nephropathy	binary
dx_D_circ	Diabetes with peripheral circulatory disorders	binary
dx_D_neur	Diabetic neuropathy	binary
dx_Hyperglycemia	History of hyperglycemia	binary
dx_stroke_isch	History of ischemic stroke	binary
hosp_adm	At least one hospital admission	binary
er_visits	Number of ER visits	continuous
raceWhite	White race (vs other)	binary
index_year_2013_2016	Index year between 2013-2016 (vs after 2016)	binary

C₁ vector

= additional covariates used for outcome data generation and, consequently, for fitting the true outcome model in simulations

Outcome Generation

1. Time to outcome (MACE) $\sim \alpha_Y * Treatment + \theta_Y * HbA1c + \sum_{k=1}^n \beta_{Yk} x_k$
2. Time to censoring $\sim \alpha_{(Y-1)} * Treatment + \theta_{Y-1} * HbA1c + \sum_{k=1}^n \beta_{(Y-1)k} x_k$

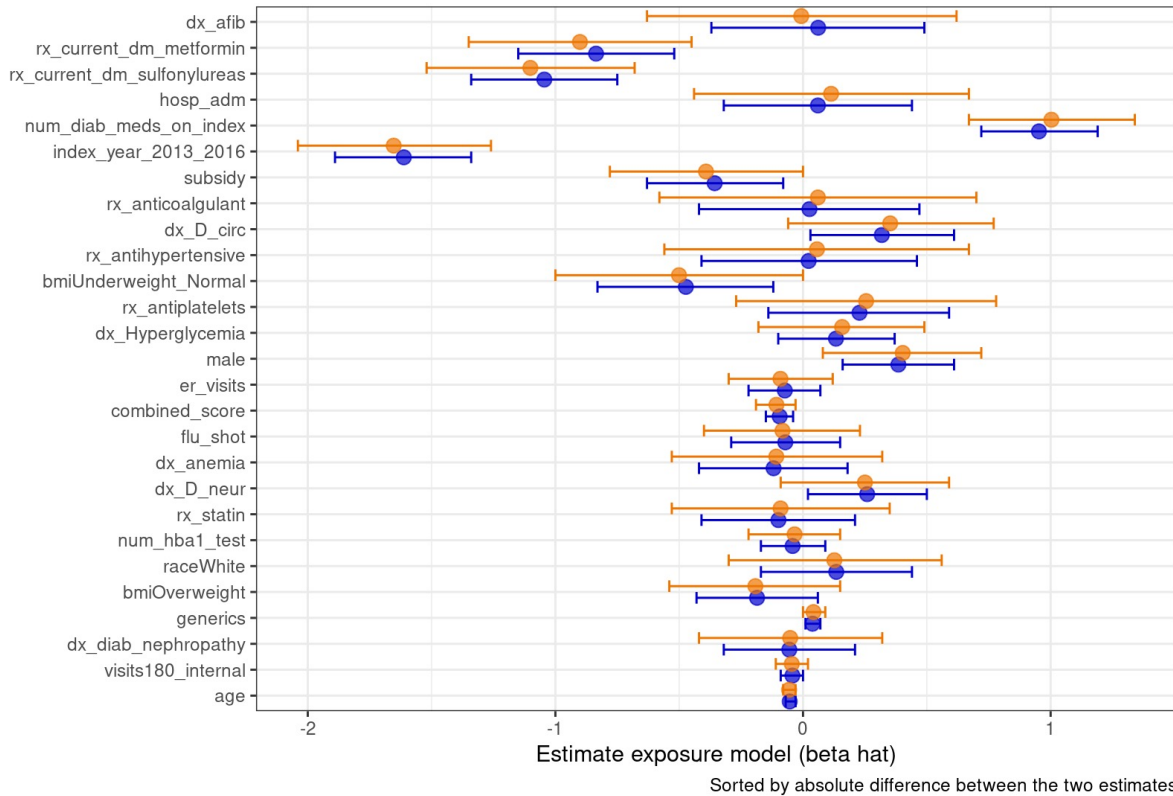
- recommended to specify a set of covariates that are believed to be associated with the outcome
- we have selected **25 covariates (C₁ vector)** to be used for outcome data generation:

Element	Covariates
Exposure	SGLT2 versus DPP4
Confounder of interest → {MCAR, MAR, MNAR}	HbA1c BMI smoking
Predictors for outcome (= cardiovascular composite; in the following also referred to as <u>C₁ covariate vector</u>)	<p>Demographics: age, sex, race, Medicare subsidy, year of index date</p> <p>Comorbidities at baseline: Combined comorbidity score, hyperglycemia, cardiomyopathy, afib, anemia, stroke/ischemia, diabetic neuropathy, diabetes with peripheral circulatory disorders, diabetic nephropathy</p> <p>Use of concomitant drugs at baseline: statin, sulfonylureas (current use), anticoagulants, antiplatelets, antihypertensive</p> <p>Healthcare utilization: number of internal physician visits, hospital admission, generics, # diabetes rx, number of HbA1c tests during baseline period, number of emergency room visits</p>



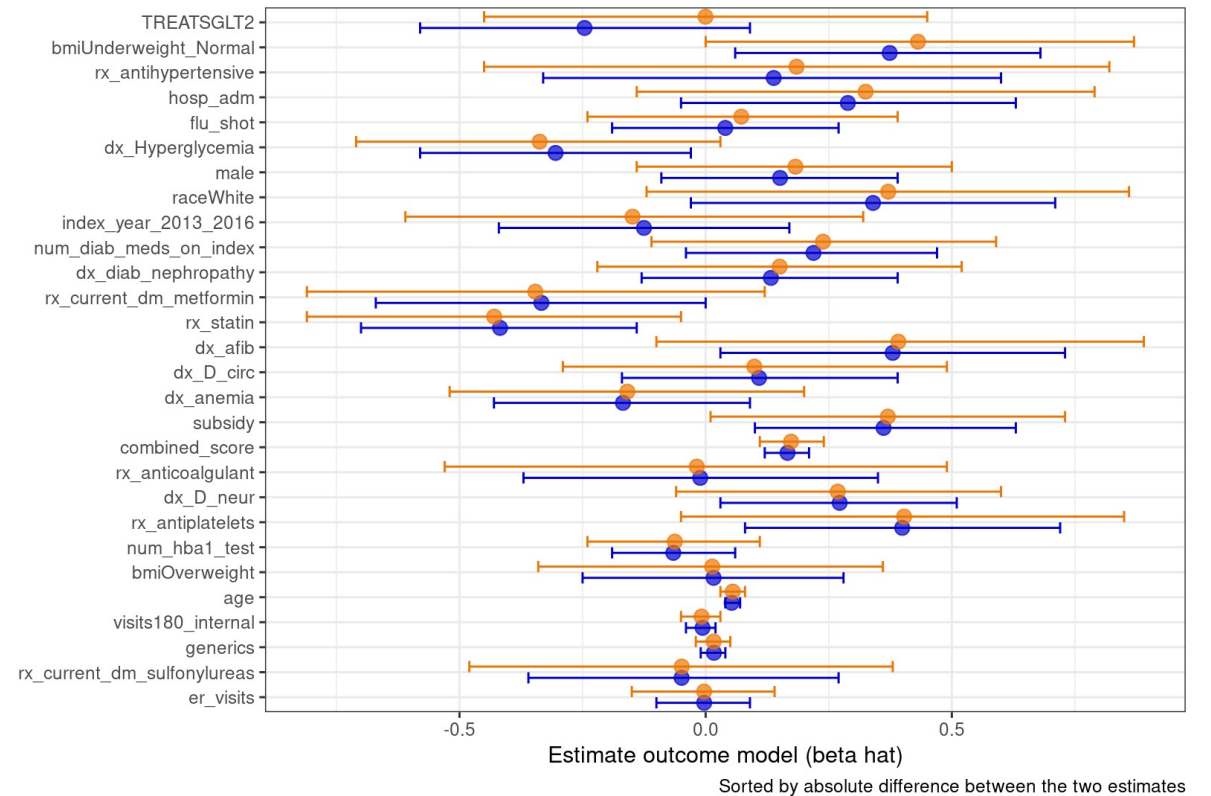
Quality Checks – Plasmode Cohorts (BMI example)

● Original Data ● Plasmode Data (averaged over 100 cohorts)



Exposure model

● Original Data ● Plasmode Data (averaged over 100 cohorts)



Outcome model

Algorithm 1 Plasmode simulation pseudocode

$N_{plasmode} = 100$ ▷ Plasmode datasets we created
 $N_{confounder} = 3$ (HbA1c, BMI, smoking)
 $N_{mechanism_{missing}} = 4$ (MCAR, MAR, $MNAR_{unmeasured}$, $MNAR_{value}$)
 $N_{proportion} = 5$ (10%, 20%, 30%, 40%, 50%)
 $N_{covars_{auxiliary}} = 2$ (with, without empirical auxiliary variables)
 $N_{indicator_{missing}} = 2$ (with, without missing indicator in imputation)
 $N_{interaction} = 2$ (with, without treatment effect heterogeneity)

Require: Matrix that lists all combinations of above vectors

for i of $1:nrow_{Matrix}$ **do**

Select $plasmode\ dataset[i]$ of $confounder[i]$ and $interaction\ term[i]$

Compute $True_{HR}$

Introduce $missing_{mechanism}[i]$ with $proportion[i] \leftarrow plasmode[i]_{missing}$

for $plasmode[i]_{missing}$ **do**

DIAGNOSTICS

Compute standardized differences (SMD)

Perform Hotelling's multivariate t-test

Fit CoxPH(Outcome $\sim confounder[i]_{indicator_{missing}} + \mathbf{x}_{covariates}$)

Fit random forest and predict $confounder[i]_{indicator_{missing}}$

IMPUTATION

Compute $HR_{non-imputed}$ with $indicator_{missing}[i]$

(Multiple) imputation with $indicator_{missing}[i]$ & $covars_{auxiliary}[i]$

Compute $HR_{imputed}$

return Results[i]

end for

end for

Overview of Model Specifications Used in Simulation

Notation

exposure = SGLT2 versus DPP4

coi = EHR confounder of interest, i.e., HbA1c, BMI, smoking

coi:exposure = interaction term for coi and exposure

coi_missing_indicator = binary variable indicating if coi is missing (=0) or observed (=1)

C_1 = Covariates used to generate outcome (consequently also used as covariates in outcome model)

C_o = All remaining (auxiliary) covariates \neq {Exposure, HbA1c, BMI, smoking, C_1 }

{ } = Indicates inclusion in model as a simulation parameter that is altered

Model	Formula
Diagnostics	
Predicting missingness	$\text{coi_missing_indicator} \sim \text{exposure} + \text{TIME} + \text{EVENT} + C_1 + \{C_o\}$
Diagnostics outcome model (differential missingness)	$\text{Surv}(\text{TIME}, \text{EVENT}) \sim \text{exposure} + \text{coi_missing_indicator} + C_1 + \{C_o\}$ (Co not considered yet)
Weighting/Imputation	
Inverse probability of missing weights (IPMW) ^a	$\text{coi_missing_indicator} \sim \text{exposure} + \text{TIME} + \text{EVENT} + C_1 + \{C_o\}$
Imputation models	$\text{coi} \sim \text{exposure} + \text{TIME} + \text{EVENT} + \{\text{coi_missing_indicator}\} + C_1 + \{C_o\}$
Outcome models	
True outcome model	$\text{Surv}(\text{TIME}, \text{EVENT}) \sim \text{exposure} + \text{coi} + \{\text{coi:exposure}\} + C_1$
Complete case/IPMW outcome model	$\text{Surv}(\text{TIME}, \text{EVENT}) \sim \text{exposure} + \text{coi} + \{\text{coi:exposure}\} + C_1$ (<i>complete cases only</i>)
Missing indicator outcome model	$\text{Surv}(\text{TIME}, \text{EVENT}) \sim \text{exposure} + \text{coi} + \{\text{coi:exposure}\} + \text{coi_missing_indicator} + C_1$
Outcome model across imputed datasets	$\text{Surv}(\text{TIME}, \text{EVENT}) \sim \text{exposure} + \text{coi} + \{\text{coi:exposure}\} + \{\text{coi_missing_indicator}\} + C_1$

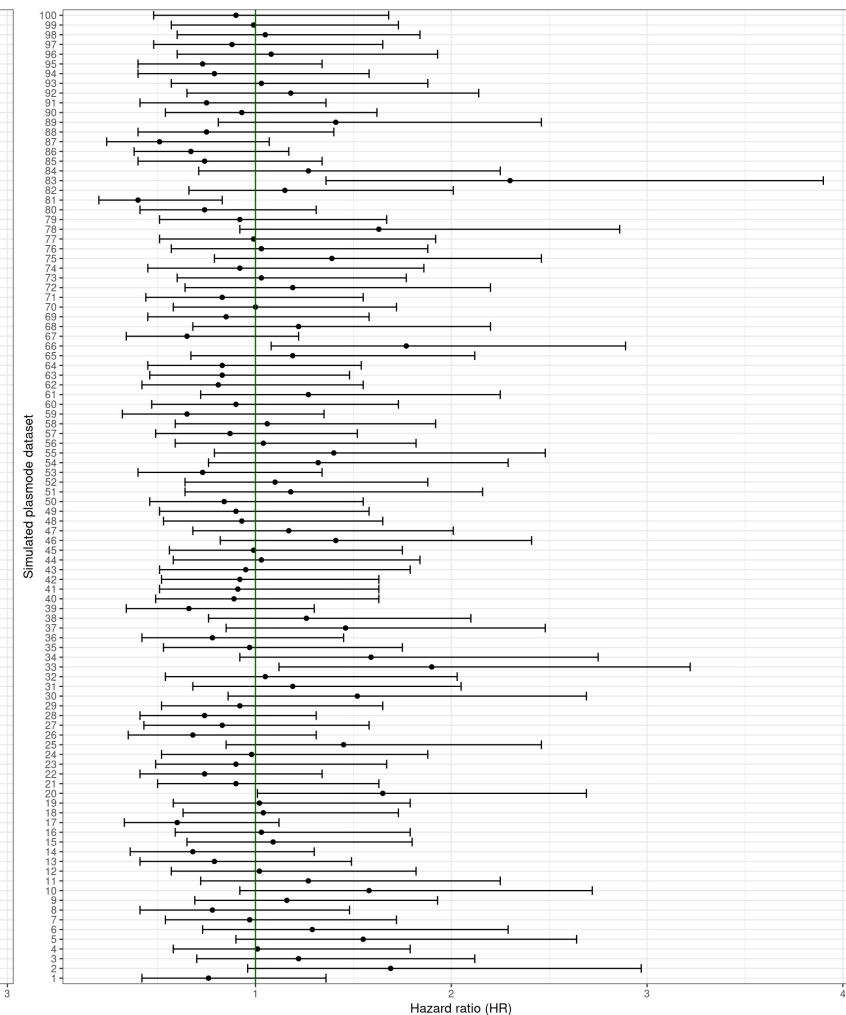
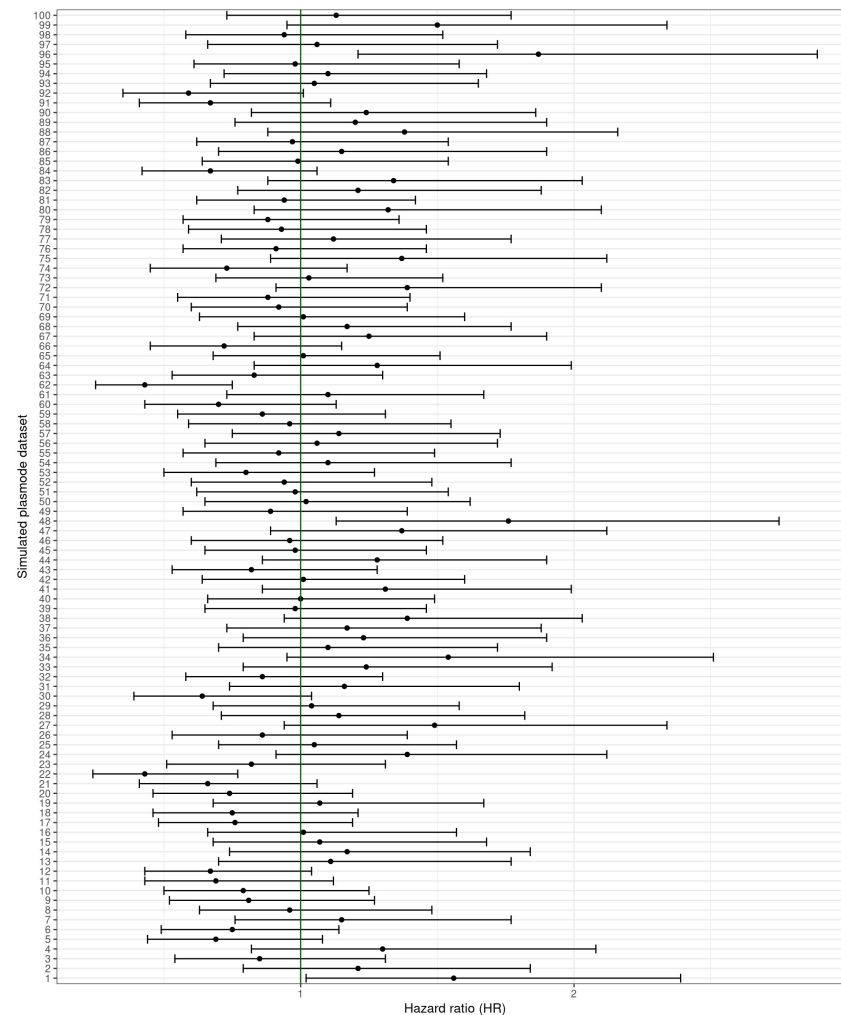
^aWeights are automatically trimmed to the 1% and 99% percentile; robust variance estimator is used in outcome model to estimate standard error

Quality/Sanity Checks – True Outcome Model (BMI example)

Effect modification	log HR (95% CI)	HR (95% CI)	Standard error
FALSE	0.00 (-0.45-0.45)	1.00 (0.64-1.57)	0.23
TRUE	0.00 (-0.59-0.59)	1.00 (0.55-1.80)	0.30

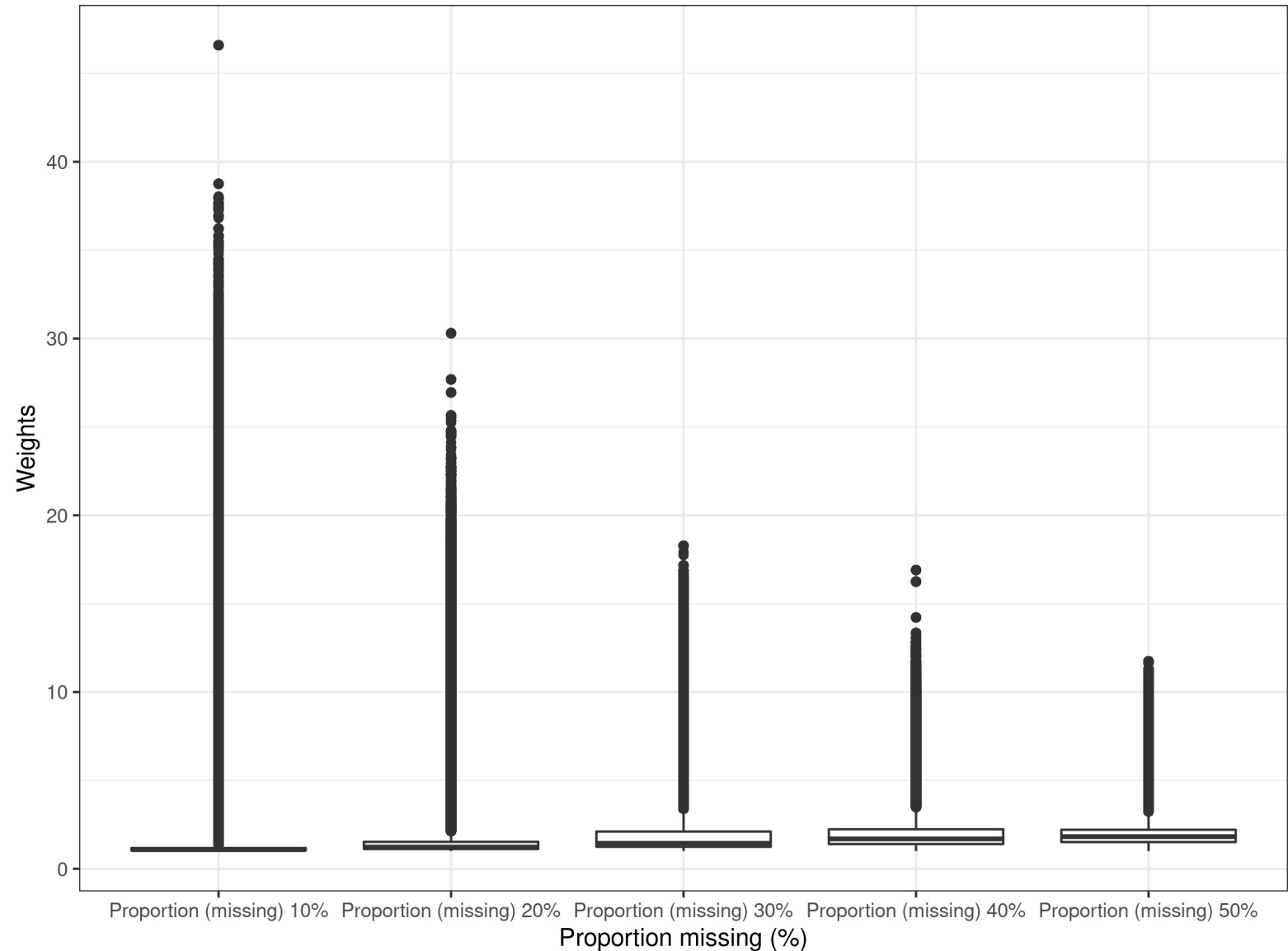
Simulated missingness proportion by mechanism and expected true proportion (quality check).

Mechanism	True proportion	Simulated proportion
Proportion missing = 10%		
MCAR	0.1	0.10
MAR	0.1	0.10
<i>MNAR_{unmeasured}</i>	0.1	0.10
<i>MNAR_{value}</i>	0.1	0.10
Proportion missing = 20%		
MCAR	0.2	0.20
MAR	0.2	0.20
<i>MNAR_{unmeasured}</i>	0.2	0.20
<i>MNAR_{value}</i>	0.2	0.20
Proportion missing = 30%		
MCAR	0.3	0.30
MAR	0.3	0.30
<i>MNAR_{unmeasured}</i>	0.3	0.30
<i>MNAR_{value}</i>	0.3	0.30
Proportion missing = 40%		
MCAR	0.4	0.40
MAR	0.4	0.39
<i>MNAR_{unmeasured}</i>	0.4	0.39
<i>MNAR_{value}</i>	0.4	0.40
Proportion missing = 50%		
MCAR	0.5	0.50
MAR	0.5	0.50
<i>MNAR_{unmeasured}</i>	0.5	0.49
<i>MNAR_{value}</i>	0.5	0.49



Post-hoc simulation analytics (by proportion)

IPMW performed very poorly due to extreme weights (especially when missingness was less frequent) for some patients



Thank You

Please visit www.sentinelinitiative.org for more information.