The missing link: Cross-species EHR data linkage offers new opportunities for improving One Health

Kathleen Mullen

University of North Carolina at Chapel Hill

Translational and Integrative Sciences Lab

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

TISLab

2024 OHDSI Global Symposium

23 October 2024

One Health

is an integrative multidisciplinary effort focused on achieving optimal health for people, animals, and their shared environments.

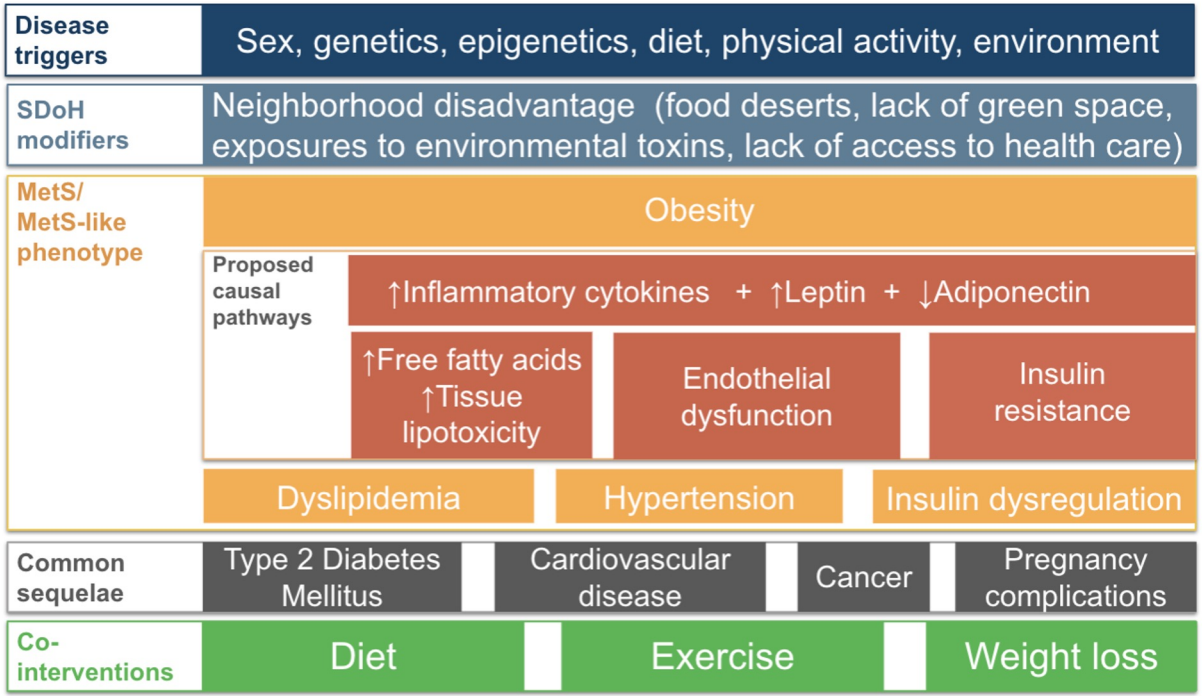
There are significant opportunities for **learning across species living in a shared environment**.

A mechanism to study the household co-occurrence of disease is needed.

Metabolic syndrome is similar in people and animals!

Image credit: Sarah M. Reuss

Miniature donkeys with equine metabolic syndrome phenotype.

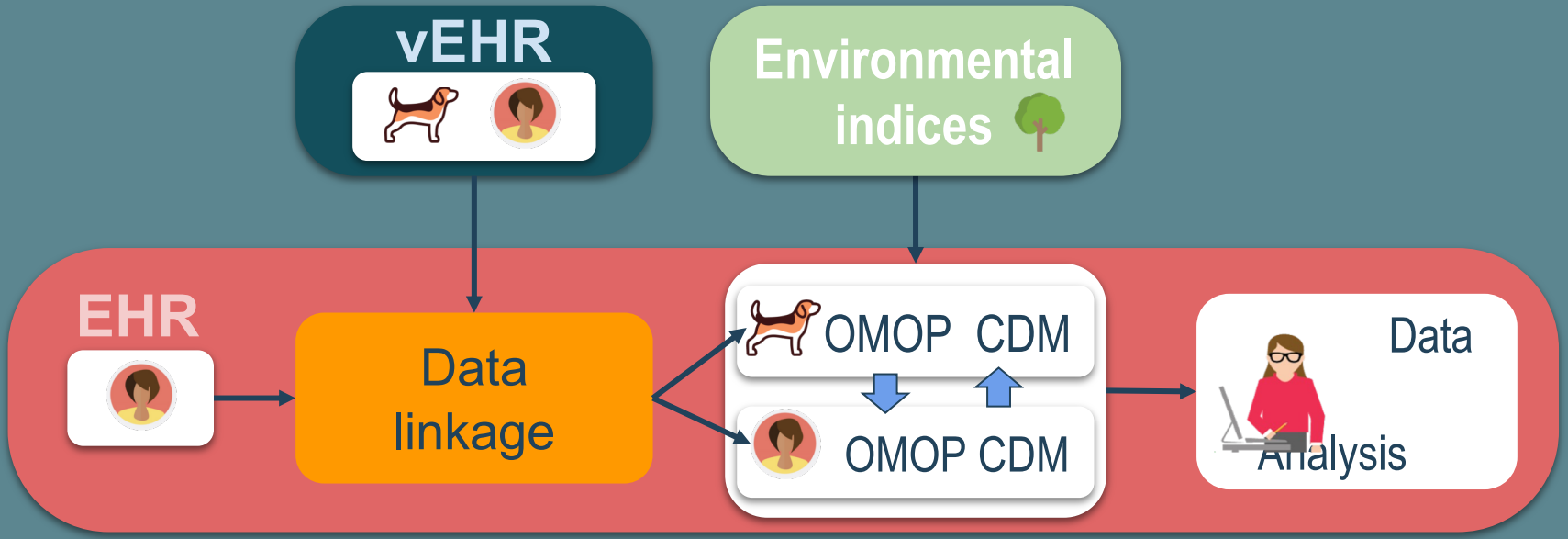


Commonalities in MetS in people and MetS-like in companion animals.

How can we explore the causes of metabolic syndrome in a single household?

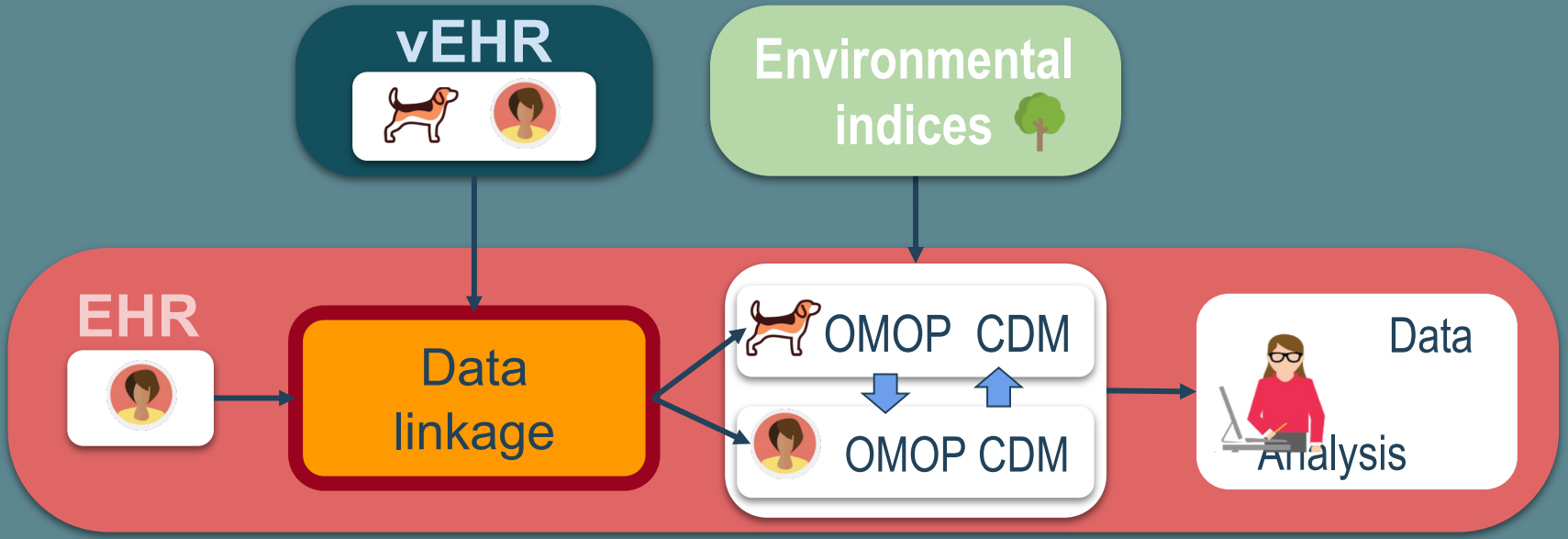
Project goal

Create a secure, pet-patient registry linking people, their animals & the environment.

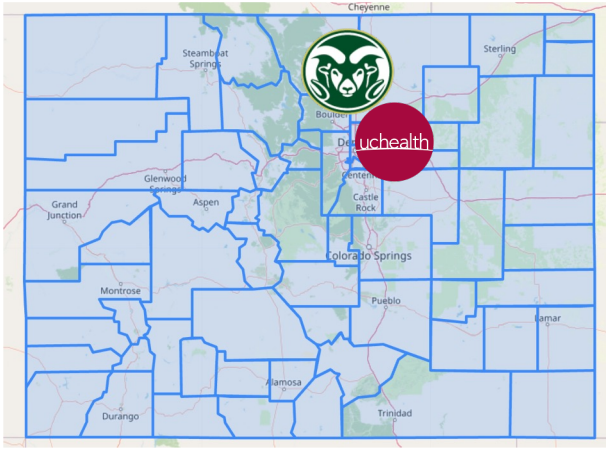


Project goal

Putting the household back together with EHR data linkage.



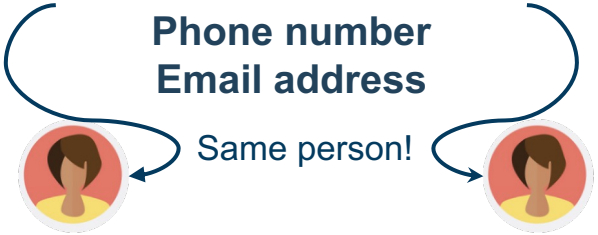
How many animals and owners can be linked via their EHRs?



Colorado map with locations of CSU-VTH and CU Medical Campus (UCHealth).

Linkage Honest Broker

First name, Last name
Street address
Phone number
Email address



How many animals and owners can be linked via their EHRs?

Vet. EHR (vEHR) time span: 2019-2024

 # animal owners: 41,081

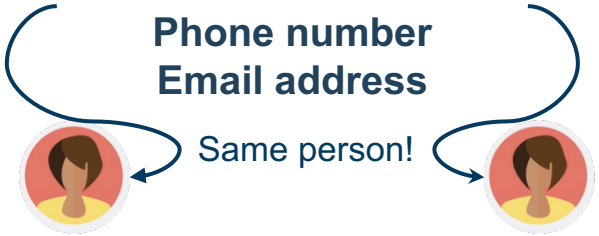
 # animals patients: 76,282

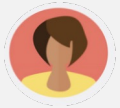
Cats:	13.0%
Dogs:	55.5%
Horses:	15.3%
Other:	16.2%

Female animals patients: 47.9%

Linkage Honest Broker

First name, Last name
Street address
Phone number
Email address





EHR time span: 2015-2024

 # human patients: 3,282,860

Female human patients: 53.4%

How many animals and owners can be linked via their EHRs?

Vet. EHR (vEHR) time span: 2019-2024

 # animal owners: 41,081

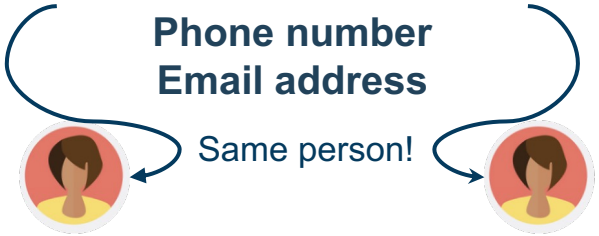
 # animals patients: 76,282

Cats:	13.0%
Dogs:	55.5%
Horses:	15.3%
Other:	16.2%

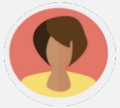
Female animals patients: 47.9%

Linkage Honest Broker

First name, Last name
Street address
Phone number
Email address



12,115 vEHR-EHR pairs!



uchealth

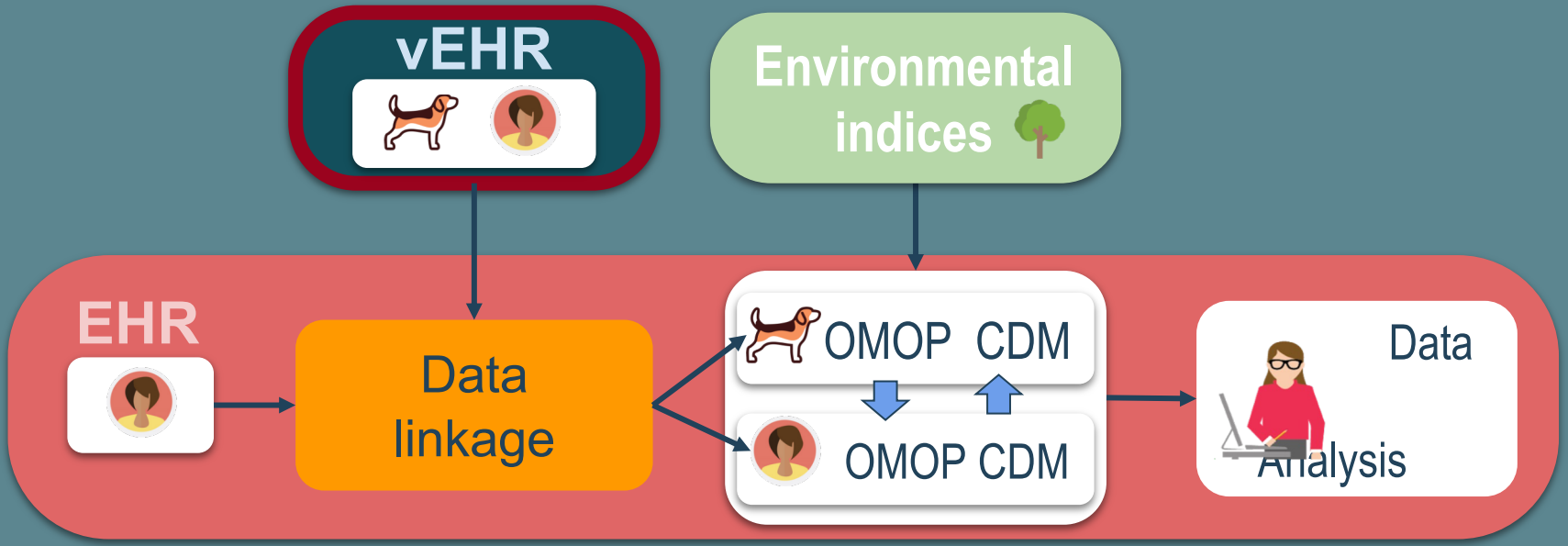
EHR time span: 2015-2024

 # human patients: 3,282,860

Female human patients: 53.4%

Project goal

Interrogating the vEHR for metabolic-like syndrome features.



Do key indicators of MetS-like exist in the vEHR?

Key indicators

- Elevated BCS
- Overweight
- Diabetes
- Obesity
- Over-conditioned*
- Equine metabolic syndrome
- Cresty neck*

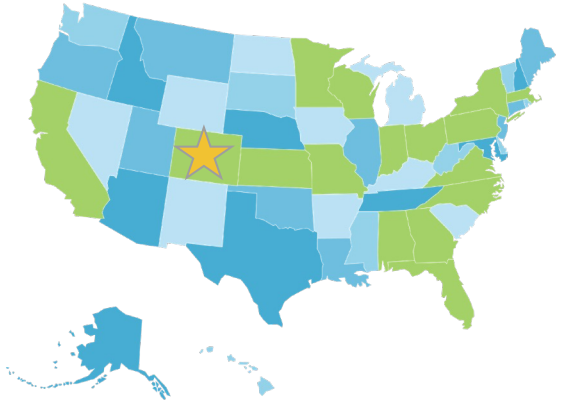
Prevalence of MetS-like key indicators in the CSU-VTH vEHR for companion animals.

Species	Animal patients (N)	Prevalence (%)
Cats*	3,037	51.0%
Dogs*	13,672	43.9%
Horses*	1,027	11.8%

*Significant differences in the prevalence of MetS-like key indicators by species ($p<0.001$).

That's a lot of fat cats!

The CSU-CU pet-patient data registry is a blueprint for **One Health** study



States with paired veterinary colleges and academic medical centers

One Health issues for the pet-patient data registry:

- Antimicrobial resistance
- Chronic diseases
- Environmental health
- Mental health
- Trauma
- Vector-borne diseases
- Zoonotic diseases

. . . And more!

We need you and your use cases! Visit poster #114 for more details!

Acknowledgements

Nadia Saklou

Adam Kiehl

Joe Strecker

Tracy Webb

Sue VandeWoude

Ian Brooks

Toan Ong

Sabrina Toro

Melissa Haendel

OHDSI Community

Support: NIH/NIAMS K12AR084226, NIH/NCATS Colorado CTSA UM1 TR004399 & UNC Department of Genetics

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

TISLab

Comparing probabilistic and rule-based phenotype algorithms for hypotension and angioedema to the experience observed in randomized clinical trials.

Joel Swerdel, PhD MS MPH[1,2], Martijn Schuemie, PhD[1,2], Judith A. Racoosin, MD, MPH[2,3], and Patrick Ryan, PhD[1,2]

[1]Janssen R&D LLC, Titusville, NJ USA; [2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA [3]United States Food and Drug Administration, Silver Spring, MD, USA

Presented at OHDSI Symposium; October 23, 2024; New Brunswick, NJ USA

Disclosures

Joel Swerdel, Martijn Schuemie, and Patrick Ryan are employees of Janssen Research and Development and shareholders of Johnson & Johnson. Judith Racoosin has no disclosures and notes that the views and opinions expressed in this presentation are those of the authors and should not be construed to represent the views or policies of the FDA.

Background

- Rule-based phenotype algorithms (PAs) are the standard for identifying outcomes in observational data.
- However, the performance characteristics of the PAs, such as sensitivity and positive predictive value, are estimated to be low for many phenotypes.
- Probabilistic PAs, e.g., PAs based on logistic regression models, offer an alternative to the rule-based method. Prior efforts have demonstrated the potential for probabilistic phenotyping as an alternative to rule-based PAs.[1-3]

Objective

The objective of this study was to develop a methodology for creating probabilistic phenotypes and to show examples of its use in hypotension, i.e., low blood pressure, and angioedema, a subcutaneous tissue swelling triggered by an allergic reaction.

Methods – Building rule-based and probabilistic phenotypes

Developing Rule-based and Probabilistic Phenotypes

Example - Rule-based phenotype

Cohort Entry Events:

People enter the cohort when observing any of the following:

condition occurrences of 'Angioedema'.

Code	Name	Vocabulary
T78.3	Angioneurotic oedema	ICD10
995.1	Angioneurotic edema, not elsewhere classified	ICD9CM

Cohort Exit:

The cohort end date will be offset from index event's end date plus 7 days.

Developing Rule-based and Probabilistic Phenotypes

Example - Rule-based phenotype

Cohort Entry Events:

People enter the cohort when observing any of the following:

condition occurrences of 'Angioedema'.

Code	Name	Vocabulary
T78.3	Angioneurotic oedema	ICD10
995.1	Angioneurotic edema, not elsewhere classified	ICD9CM

Cohort Exit:

The cohort end date will be offset from index event's end date plus 7 days.

Example - Probabilistic phenotype

Beta Coefficient	Covariate Name
4.82	condition_era group during day 0 through 10 days relative to index: **Angioedema**
3.88	condition_era group during day 0 through 10 days relative to index: **Allergic disposition**
2.35	visit_occurrence concept count during day 0 through 10 concept_count relative to index: **Emergency Room Visit**
1.88	drug_era group during day 0 through 10 days relative to index: **prednisone**
1.87	condition_era group during day 0 through 10 days relative to index: **Anaphylaxis**
1.78	drug_era group during day 0 through 10 days relative to index: **ACE INHIBITORS, PLAIN**
1.58	drug_era group during day 0 through 10 days relative to index: **H2-receptor antagonists**
1.57	observation during day 0 through 10 days relative to index: **Adverse reaction to substance**
1.52	condition_era group during day 0 through 10 days relative to index: **Angioedema and/or urticaria**
1.38	condition_era group during day 0 through 10 days relative to index: **Edema**
1.34	drug_era group during day 0 through 10 days relative to index: **Sympathomimetics in glaucoma therapy**
1.28	drug_era group during day 0 through 10 days relative to index: **CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN**
1.26	condition_era group during day 11 through 20 days relative to index: **Angioedema and/or urticaria**
1.20	condition_era group during day 0 through 10 days relative to index: **Acute allergic reaction**
...	

Developing Rule-based and Probabilistic Phenotypes

Example - Rule-based phenotype

Cohort Entry Events:

People enter the cohort when observing any of the following:

condition occurrences of 'Angioedema'.

Example - Probabilistic phenotype

Beta Coefficient	Covariate Name
4.82	condition_era group during day 0 through 10 days relative to index: **Angioedema**
3.88	condition_era group during day 0 through 10 days relative to index: **Allergic disposition**
2.35	visit_occurrence concept count during day 0 through 10 concept_count relative to index: **Emergency Room Visit**
1.88	drug_era group during day 0 through 10 days relative to index: **prednisone**
1.87	condition_era group during day 0 through 10 days relative to index: **Anaphylaxis**
1.78	drug_era group during day 0 through 10 days relative to index: **ACE INHIBITORS, PLAIN**

Beta Coefficient	Covariate Name
4.82	condition during day 0 through 10 days: **Angioedema**
3.88	condition during day 0 through 10 days: **Allergic disposition**
2.35	Visit occurrence during day 0 through 10 days: **Emergency Room Visit**

Developing Rule-based and Probabilistic Phenotypes (cont.)

Rule-based phenotype

1. Create a rule-based phenotype algorithm
2. Find subjects during the time-at-risk in the cohort of interest satisfying algorithm

Developing Rule-based and Probabilistic Phenotypes (cont.)

Rule-based phenotype

1. Create a rule-based phenotype algorithm
2. Find subjects during the time-at-risk in the cohort of interest satisfying algorithm

Probabilistic phenotype

1. Use noisy labeled positive and negative controls to develop a supervised learning probabilistic model using LASSO regularized regression
2. Apply model at each appropriate time point during the time-at-risk for each subject in the cohort of interest
3. Select the highest probability among the different time points within the time-at-risk for each subject
4. Use a designated probability cut-point, e.g., 70%, to determine those with the outcome

Evaluating the model

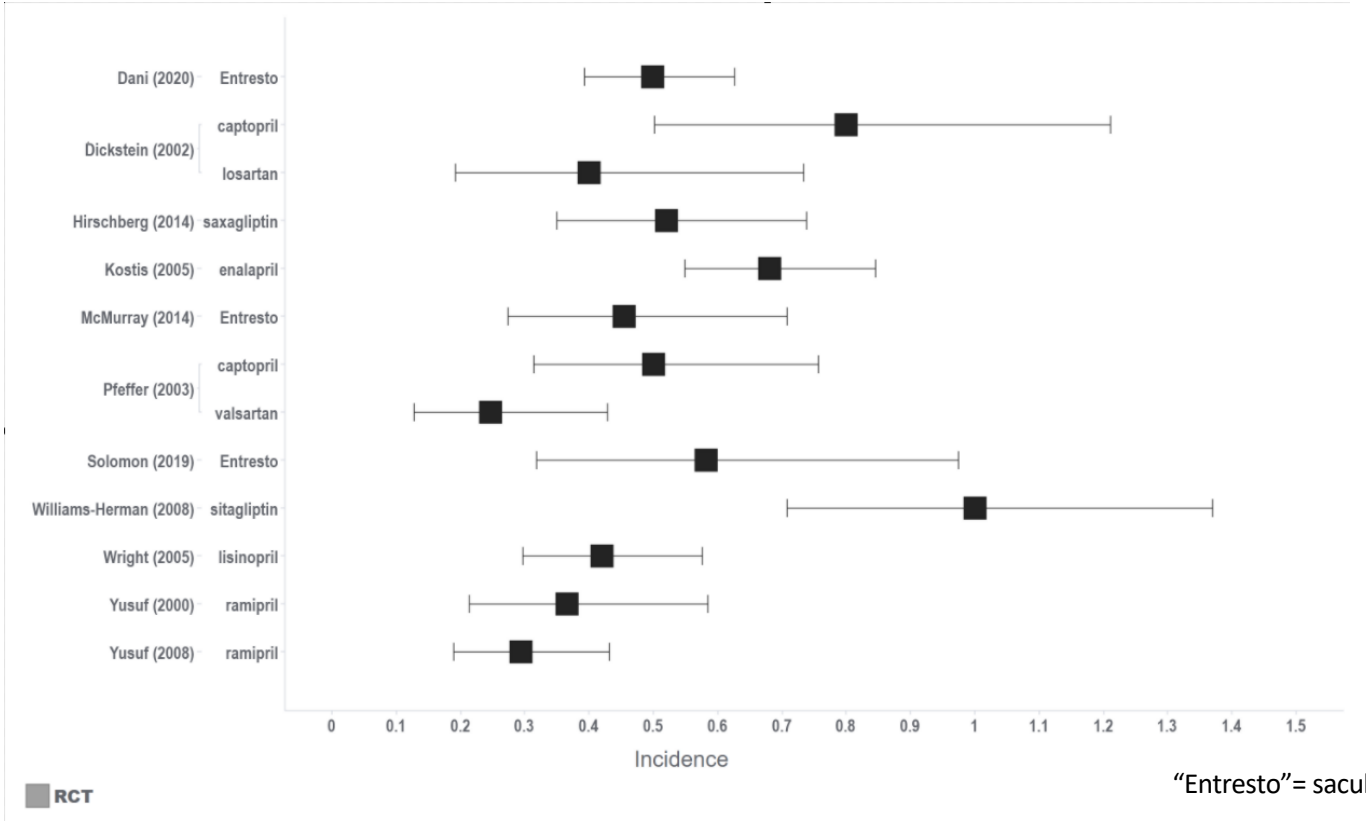
- Analysis conducted in 5 administrative claims datasets
- Rule-based algorithm used an occurrence of a diagnosis code for hypotension or angioedema
- Developed probabilistic phenotypes and examined the results using probability cut-points of 0.50, 0.60, 0.67, 0.70, 0.75, 0.80, and 0.90.
- Estimated incidence of angioedema, while on-treatment, for 7 anti-hypertensive and 2 anti-diabetic (DPP-4 inhibitors) drugs
- Using both the rule-based and probabilistic phenotypes, we performed the analysis on 9 new user drug cohorts from 2010 to 2023

Evaluating the model - Metrics

- Extracted incidence estimates and 95% confidence intervals from randomized clinical trials (RCTs) as a basis for comparison.
- Computed the proportion of incidence estimates for the rule-based and probabilistic algorithms that fell within the 95% confidence intervals (CI) of the incidence estimates from the clinical trials.
- Assessed the performance characteristics, e.g., positive predictive value (PPV) and sensitivity, of the rule-based and probabilistic phenotypes using the OHDSI tool PheValuator.

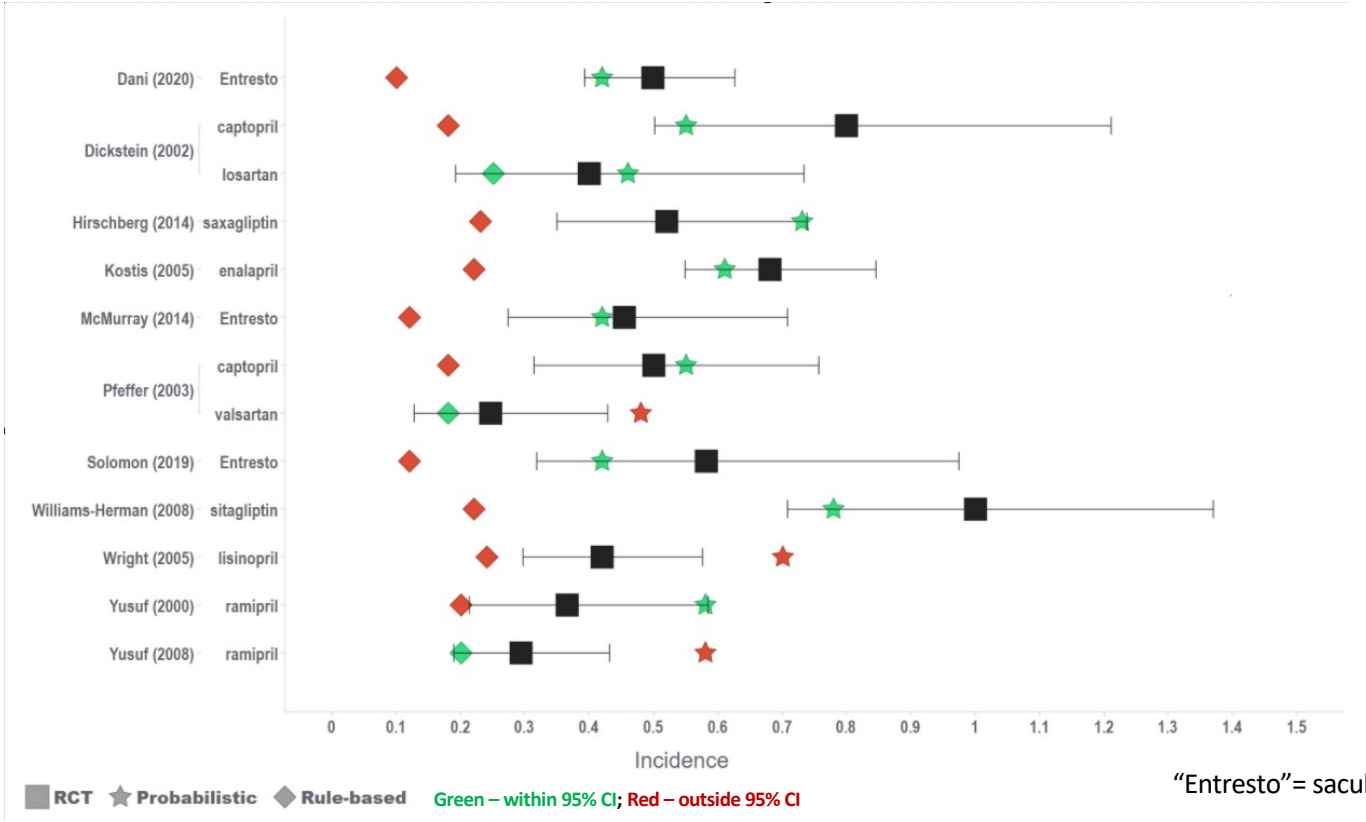
Results

Results: Angioedema - Probabilistic (Cut-point 0.67) v. Rule-based



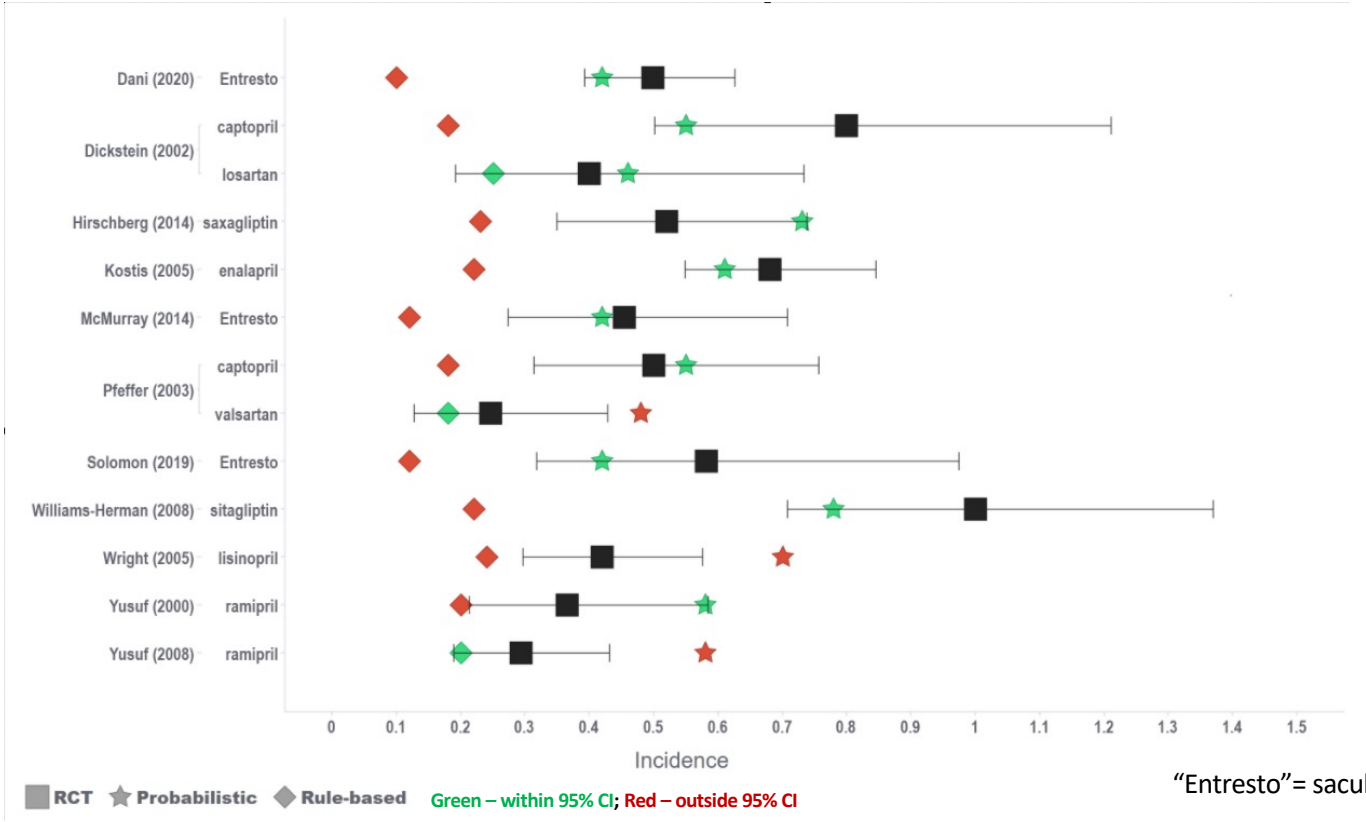
"Entresto"= sacubitril/valsartan

Results: Angioedema - Probabilistic (Cut-point 0.67) v. Rule-based



"Entresto"= sacubitril/valsartan

Results: Angioedema - Probabilistic (Cut-point 0.67) v. Rule-based



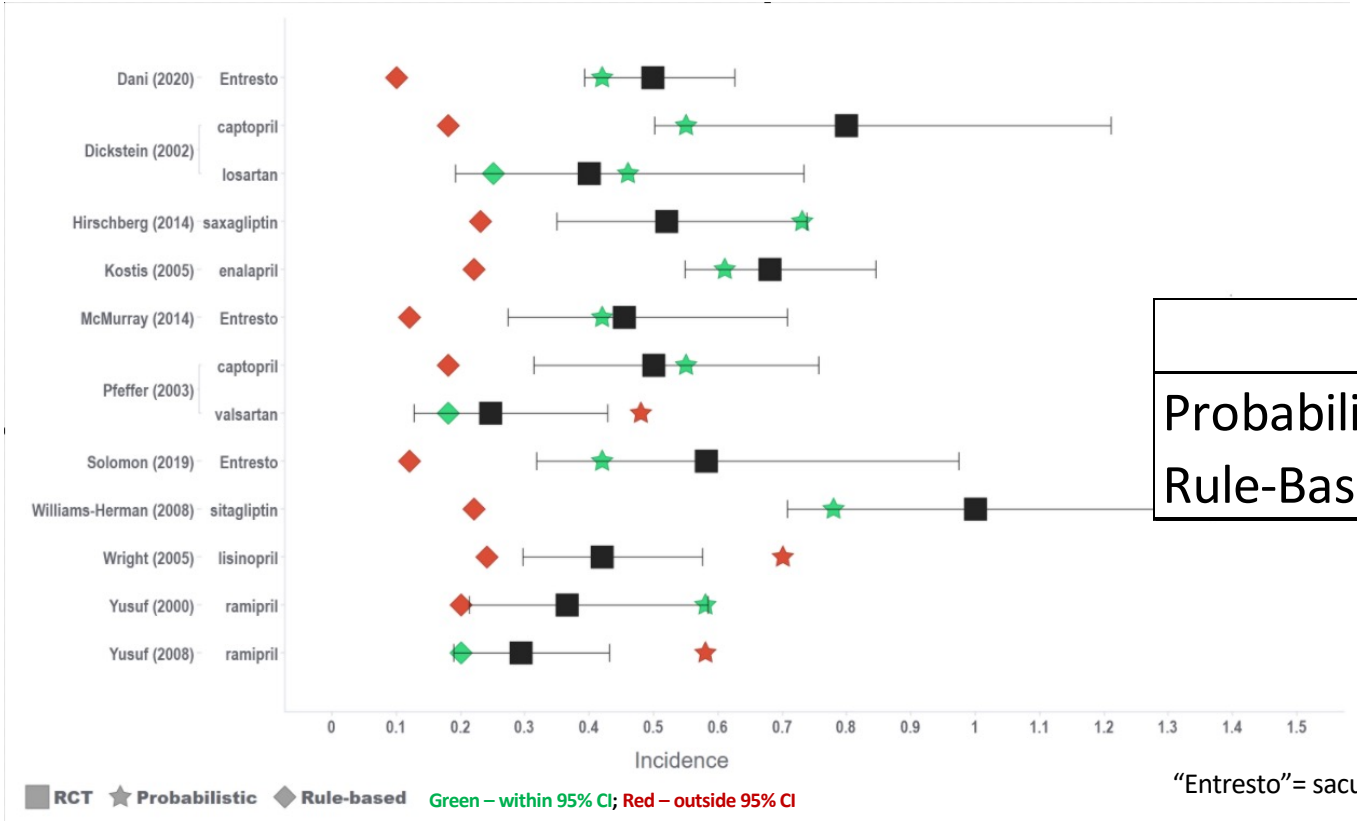
Within 95% CI:

Probabilistic: 77%

Rule-based: 23%

"Entresto"= sacubitril/valsartan

Results: Angioedema - Probabilistic (Cut-point 0.67) v. Rule-based



Within 95% CI:

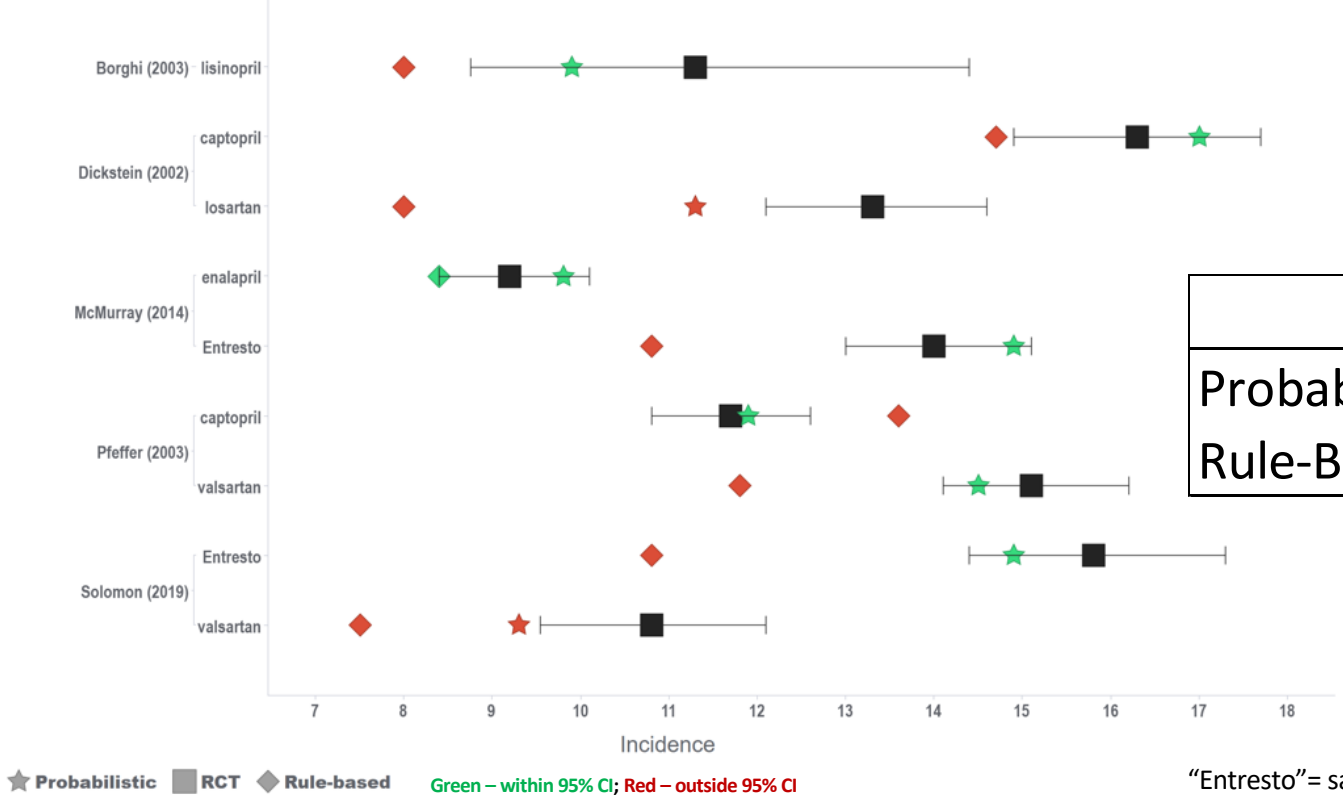
Probabilistic: 77%

Rule-based: 23%

	Sensitivity	PPV
Probabilistic	0.71	0.87
Rule-Based	0.23	0.78

"Entresto"= sacubitril/valsartan

Results: Hypotension - Probabilistic (Cut-point 0.67) v. Rule-based



Within 95% CI:

Probabilistic: 78%

Rule-based: 11%

	Sensitivity	PPV
Probabilistic	0.82	0.89
Rule-Based	0.56	0.72

"Entresto"= sacubitril/valsartan

Results: Probabilistic v. Rule-based

For both angioedema and hypotension:

- The 0.67 cut-point provided the closest match for the RCT results.
- Lower cut-points – produced higher average incidence estimates compared to RCT results.
- Higher cut-points – produced lower average incidence estimates compared to RCT results.

Conclusions

- Probabilistic phenotype algorithms (PA) for angioedema and hypotension estimated incidence closer to the results from RCTs than rule-based PAs.
- The performance of probabilistic PAs was superior to rule-based PAs on PPV and sensitivity.
- Future research is needed to evaluate the performance of probabilistic PAs in postmarket safety settings and to determine how they could potentially be used to estimate the incidence of drug adverse effects.

References

References:

1. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016;23(6):1166-73.
2. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017:48-57.
3. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018;1:53-68.
4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75.
5. Swerdel JN, Schuemie M, Murray G, Ryan PB. PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. J Biomed Inform. 2022:104177.
6. Dickstein K, Kjekshus J. Comparison of the effects of losartan and captopril on mortality in patients after acute myocardial infarction: the OPTIMAAL trial design. Optimal Therapy in Myocardial Infarction with the Angiotensin II Antagonist Losartan. Am J Cardiol. 1999;83(4):477-81.
7. Pfeffer MA, McMurray JJ, Velazquez EJ, Rouleau JL, Køber L, Maggioni AP, et al. Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. N Engl J Med. 2003;349(20):1893-906.
8. Kostis JB, Kim HJ, Rusnak J, Casale T, Kaplan A, Corren J, et al. Incidence and characteristics of angioedema associated with enalapril. Arch Intern Med. 2005;165(14):1637-42.
9. Dani SS, Ganatra S, Vaduganathan M. Angioedema with sacubitril/valsartan: Trial-level meta-analysis of over 14,000 patients and real-world evidence to date. Int J Cardiol. 2021;323:188-91.
10. McMurray JJ, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, et al. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med. 2014;371(11):993-1004.
11. Solomon SD, McMurray JJV, Anand IS, Ge J, Lam CSP, Maggioni AP, et al. Angiotensin-Neprilysin Inhibition in Heart Failure with Preserved Ejection Fraction. N Engl J Med. 2019;381(17):1609-20.
12. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. N Engl J Med. 2000;342(3):145-53.
13. Yusuf S, Teo KK, Pogue J, Dyal L, Copland I, Schumacher H, et al. Telmisartan, ramipril, or both in patients at high risk for vascular events. N Engl J Med. 2008;358(15):1547-59.
14. Hirshberg B, Parker A, Edelberg H, Donovan M, Iqbal N. Safety of saxagliptin: events of special interest in 9156 patients with type 2 diabetes mellitus. Diabetes Metab Res Rev. 2014;30(7):556-69.
15. Williams-Herman D, Round E, Swern AS, Musser B, Davies MJ, Stein PP, et al. Safety and tolerability of sitagliptin in patients with type 2 diabetes: a pooled analysis. BMC Endocr Disord. 2008;8:14.
16. Borghi C, Ambrosioni E. Double-blind comparison between zofenopril and lisinopril in patients with acute myocardial infarction: results of the Survival of Myocardial Infarction Long-term Evaluation-2 (SMILE-2) study. Am Heart J. 2003;145(1):80-7.
17. Wright JT, Dunn JK, Cutler JA, Davis BR, Cushman WC, Ford CE, et al. Outcomes in Hypertensive Black and Nonblack Patients Treated With Chlorthalidone, Amlodipine, and Lisinopril. JAMA. 2005;293(13):1595-608.

Thank you!

Questions?

Come to Poster

115

2024 Global OHDSI lightning talk

Exploring the interplay between metabolic syndrome and brain volume in depression

: Basis for Phenotype-Based Classification

Department of Biomedical Science and Biomedical Informatics

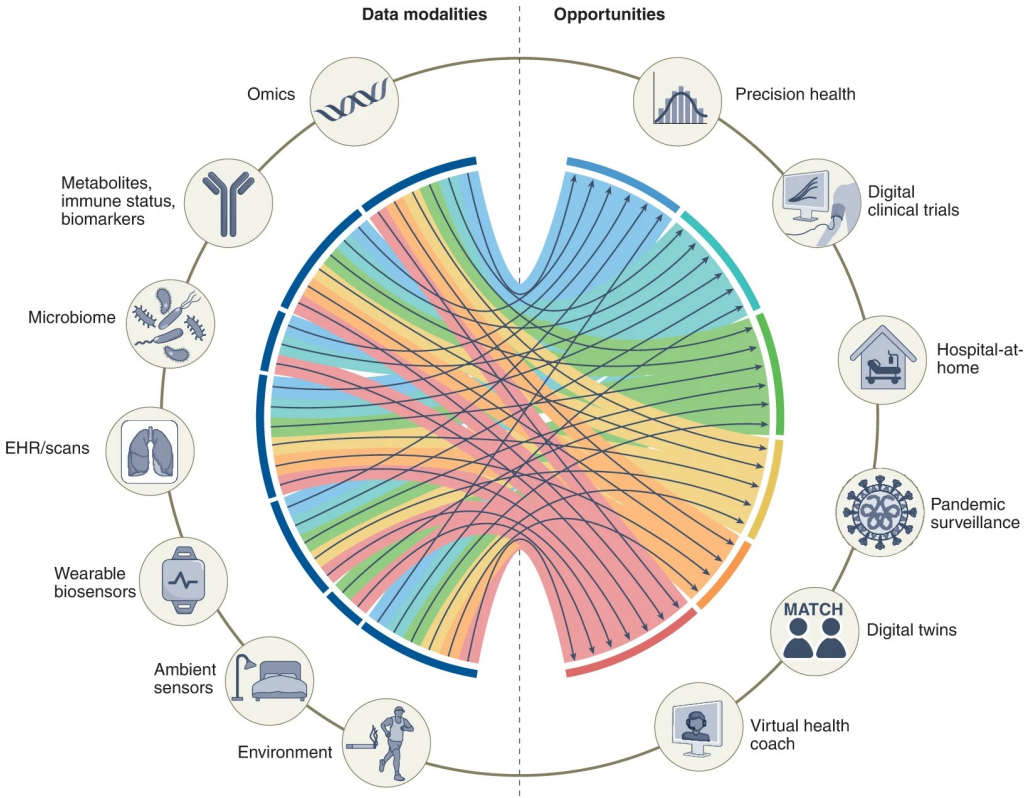
Sujin Gan

PhD student, Ajou University School of Medicine

Advisor: Professor Rae Woong Park

Latest Research Trends

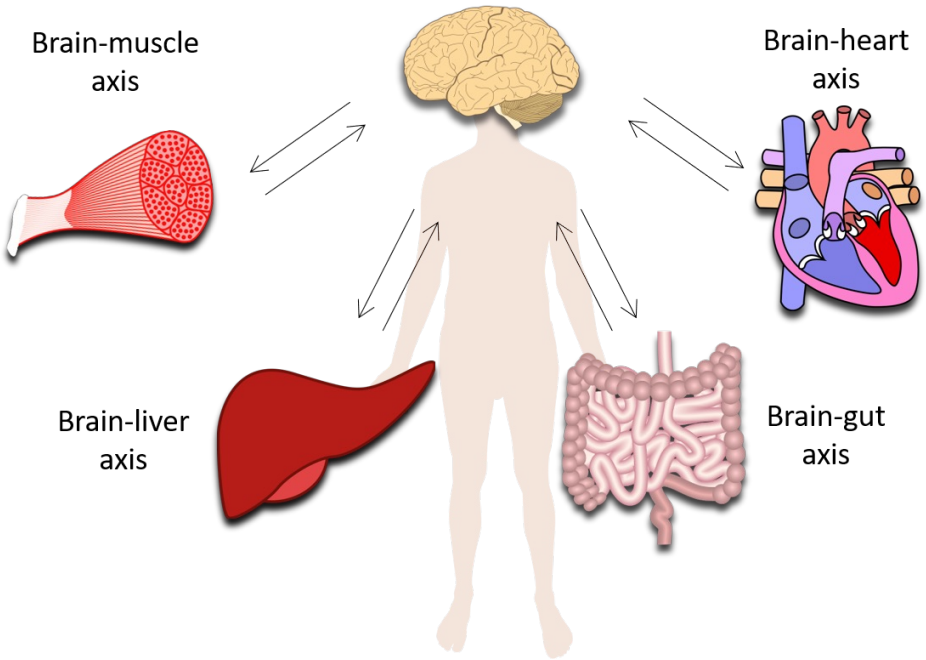
Growing utilization of multimodal approaches in medical research



Latest Research Trends

Multi-organ Interaction

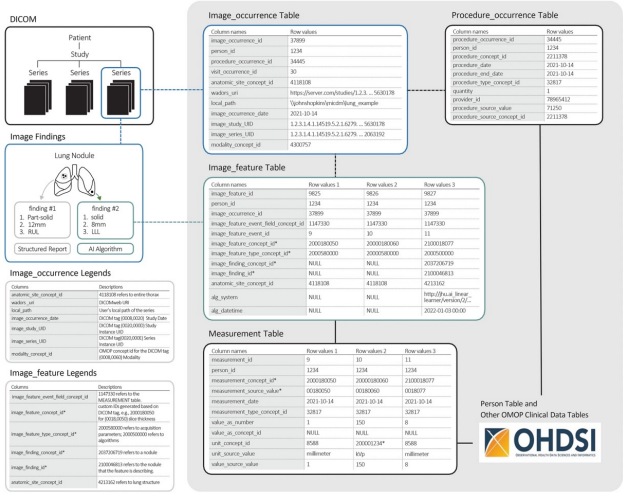
How do these interconnected systems between the various organs contribute to overall health and disease mechanisms?



Understanding the interaction axes between the brain and various

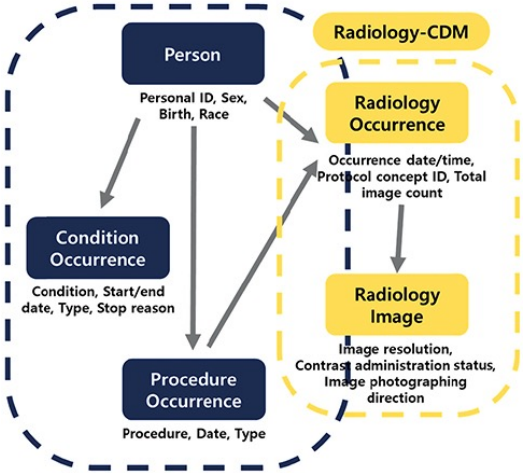
Multimodal Integration in OMOP CDM

OMOP CDM Medical Imaging Extension



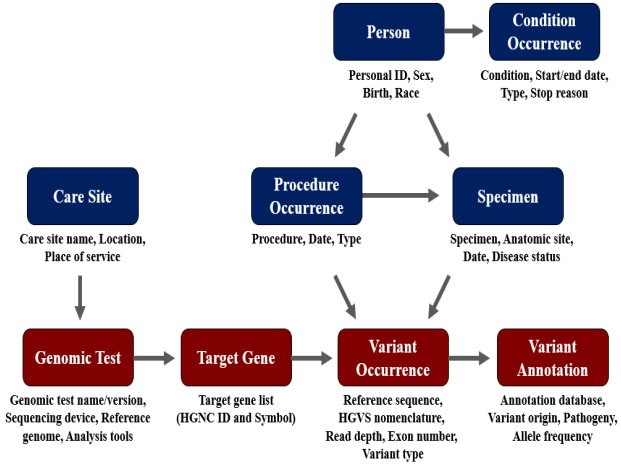
Park, W.Y., Jeon, K., et al. J Digit Imaging. Inform. med. 2024

R-CDM



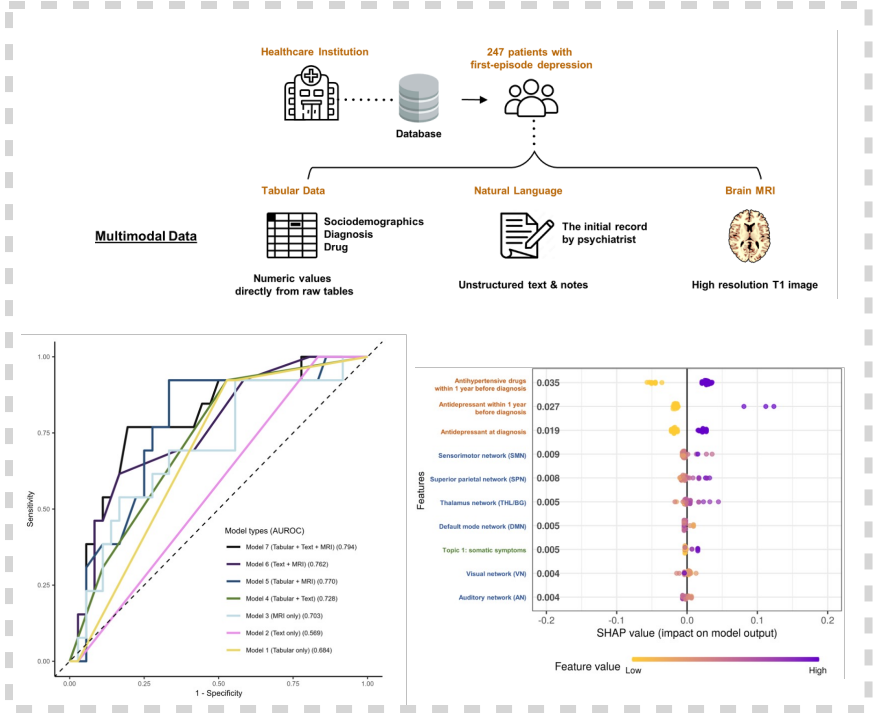
Park C, You SC, et al. Yonsei Med J. 2022

G-CDM

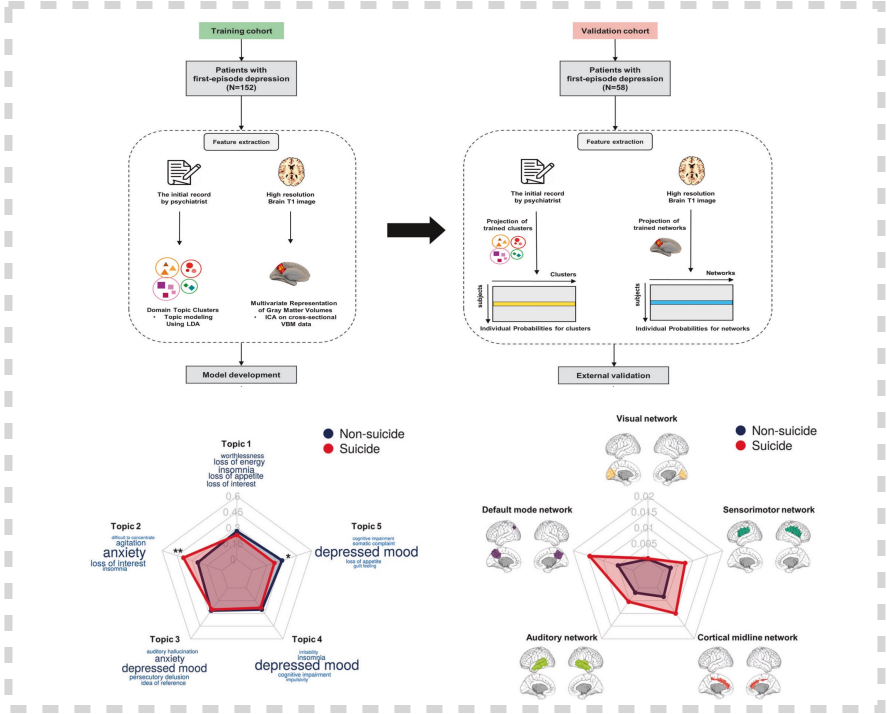


Shin SJ, You SC, Park RW. et al. J Med Internet Res. 2019

Multimodal Integration in OMOP CDM

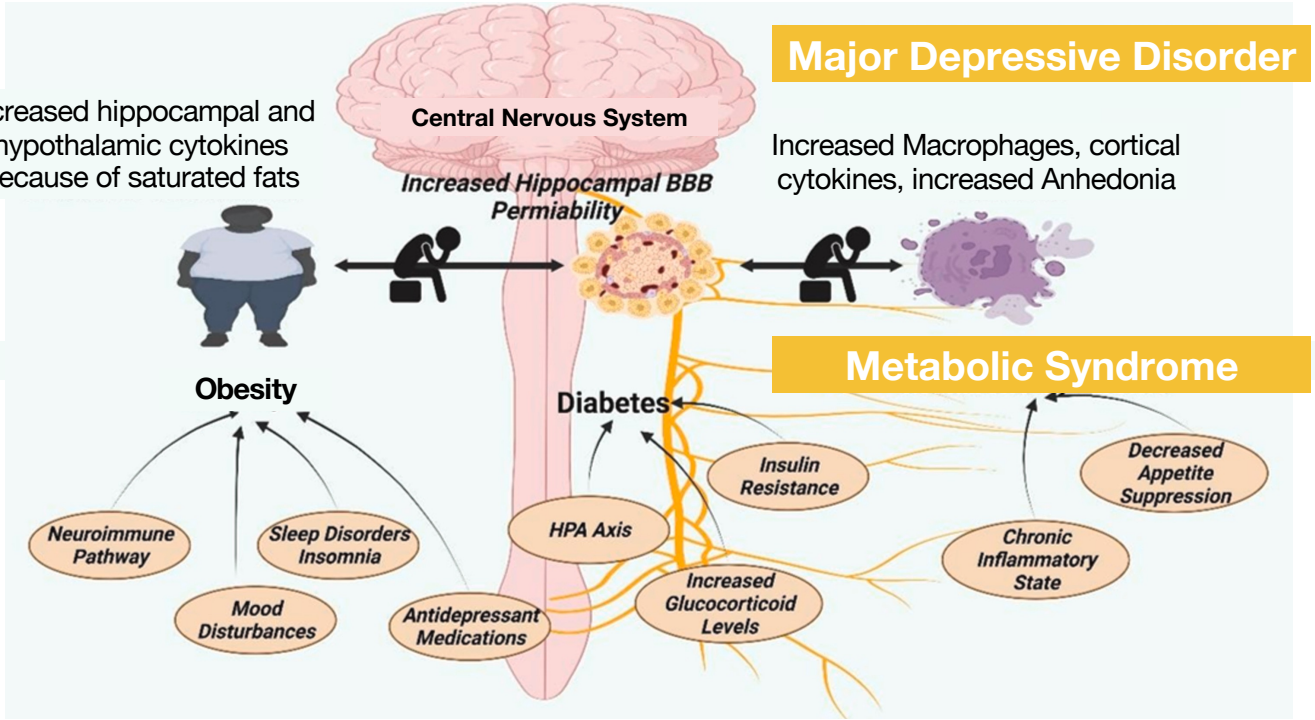


Lee, D.Y, Kim N.R., et al, Psychiatry Research, 2024



Lee, D.Y., Byeon, G, et al. Transl Psychiatry, 2024

Bi-directional relationship between Major Depressive Disorder and Metabolic Syndrome



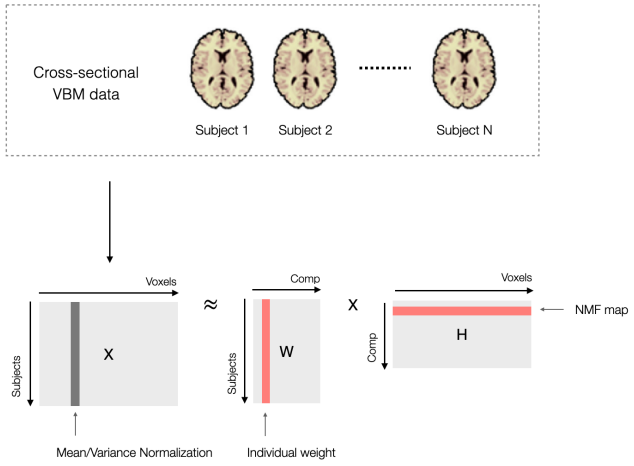
Research Hypothesis

- **The bidirectional interaction** between depression and metabolic syndrome is **mediated** by specific **brain volume components** and **peripheral laboratory markers**.

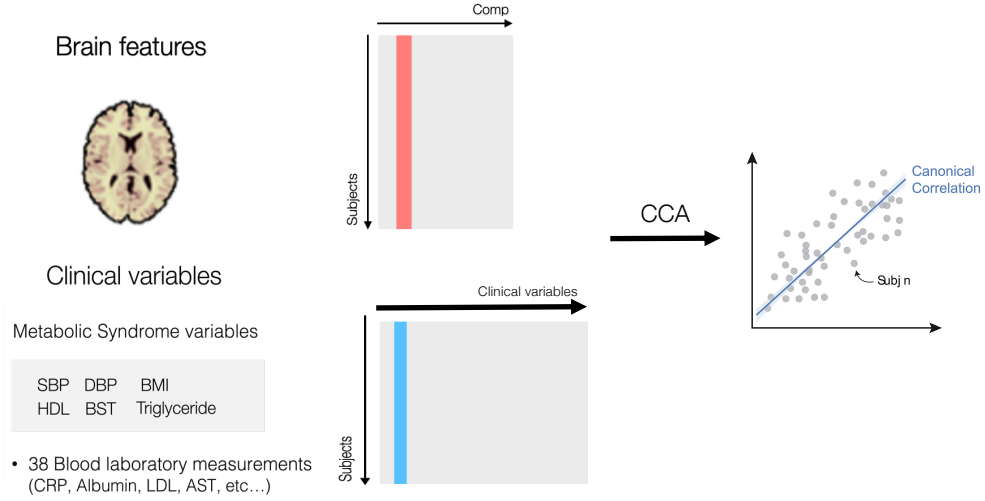
- These components can serve as biomarkers **to classify the presence of metabolic syndrome** in patients with major depressive disorder (MDD).

Overview

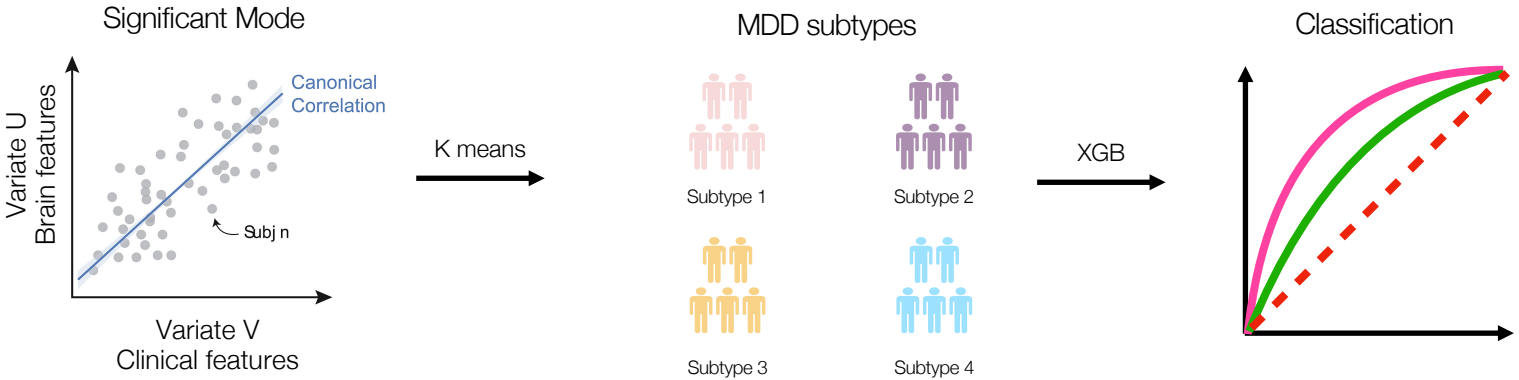
Dimension Reduction in VBM



Canonical Correlation Analysis



Classification



Study population

OMOP CDM

- Electronic Health Records database from Ajou University School of Medicine (AUSOM)
- January 1994 to July 2023 (OMOP-CDM 534)

Cohort

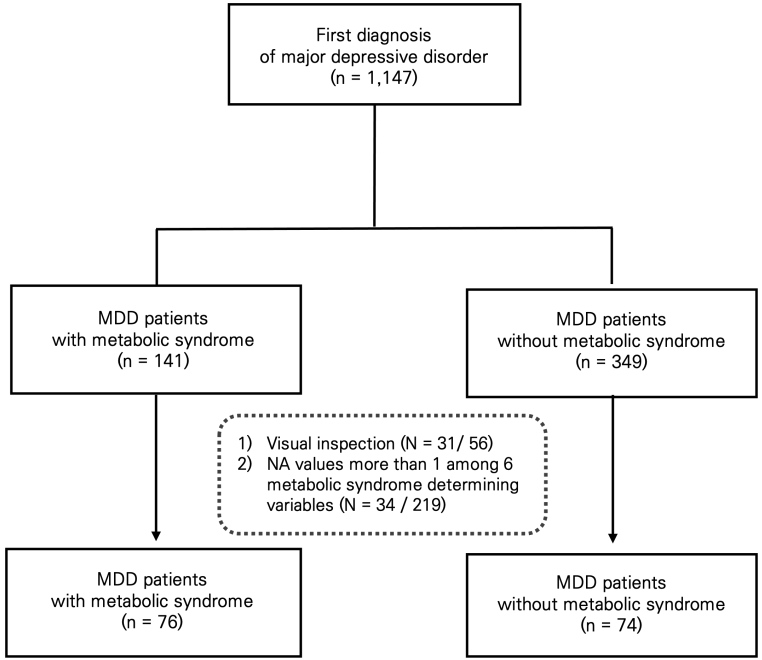
- Major Depressive Disorder diagnosis, for the first time
- Brain MRI procedures (- 365 d ~ + 30 d from the index date)
- No history of psychiatric comorbidities (bipolar disorder, schizophrenia, or dementia), substance disorders, brain injury, hydrocephalus

Definition for Metabolic syndrome

- More than 2 out of the following MetS criteria

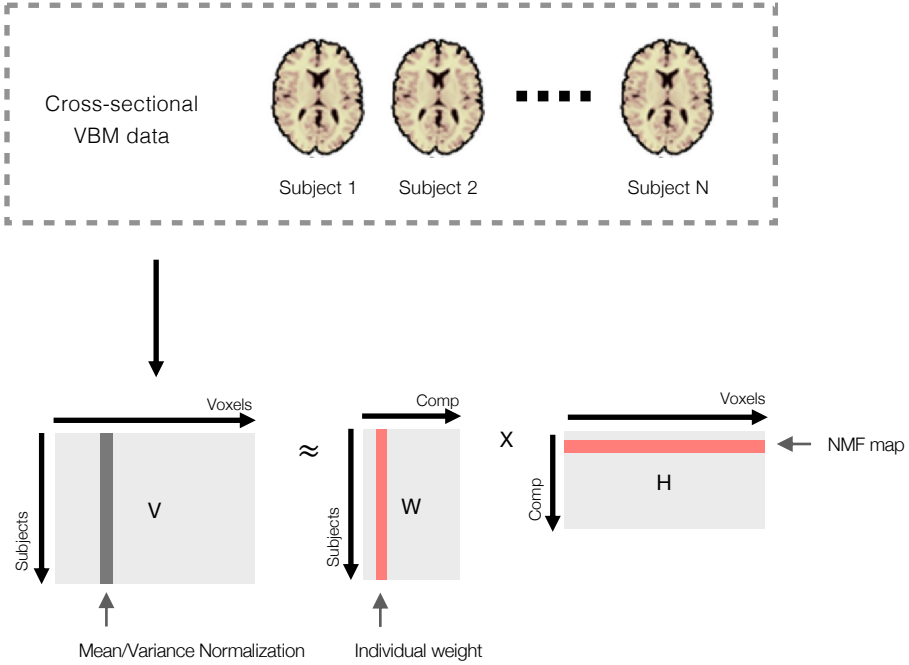
1) Antihypertensive drugs uses or **systolic blood pressure (SBP)** > 130
2) Antidiabetic drugs uses or **blood sugar level (BST)** > 100
3) Low **HDL-cholesterol** levels (<40 mg/dL for men and <50 mg/dL for women)
4) Hypertriglyceridemia (**Triglyceride** ≥ 150 mg/dL)
5) **BMI** ≥ 30

Study population flow chart



Dimension Reduction

Non-negative Matrix Factorization



$$V \approx W \cdot H$$

The matrix **V** is represented by the two smaller matrices **W** and **H**, which, when multiplied, approximately reconstruct **V**

$$\mathbf{H}^{n+1}_{[i,j]} \leftarrow \mathbf{H}^{n}_{[i,j]} \frac{((\mathbf{W}^n)^T \mathbf{V})_{[i,j]}}{((\mathbf{W}^n)^T \mathbf{W}^n \mathbf{H}^n)_{[i,j]}}$$

and

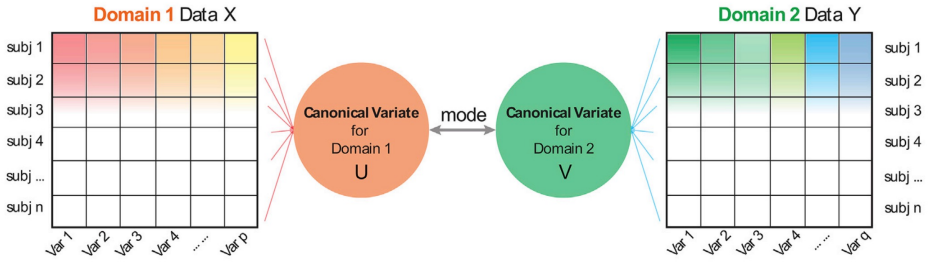
$$\mathbf{W}^{n+1}_{[i,j]} \leftarrow \mathbf{W}^{n}_{[i,j]} \frac{(\mathbf{V}(\mathbf{H}^{n+1})^T)_{[i,j]}}{(\mathbf{W}^n \mathbf{H}^{n+1} (\mathbf{H}^{n+1})^T)_{[i,j]}}$$

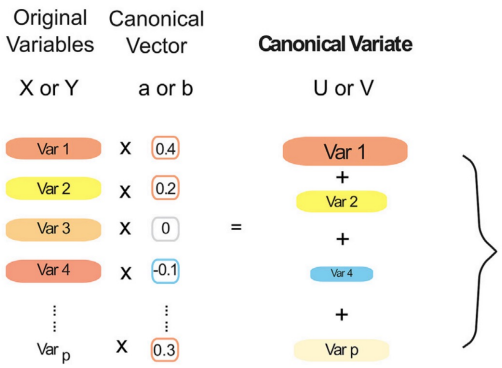
Initialize: W and H non negative.

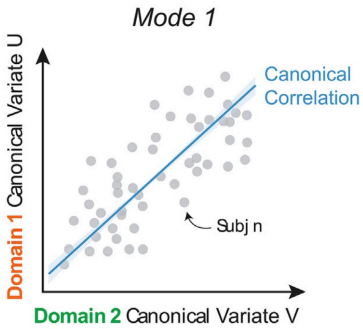
Until W and H are stable.

Dimension Reduction

Canonical Correlation Analysis







$$\Sigma_{XX} = Cov(X,X) = X^T X \text{ and } \Sigma_{YY} = Cov(Y,Y) = Y^T Y$$

$$\Sigma_{XY} = Cov(X,Y) = X^T Y$$

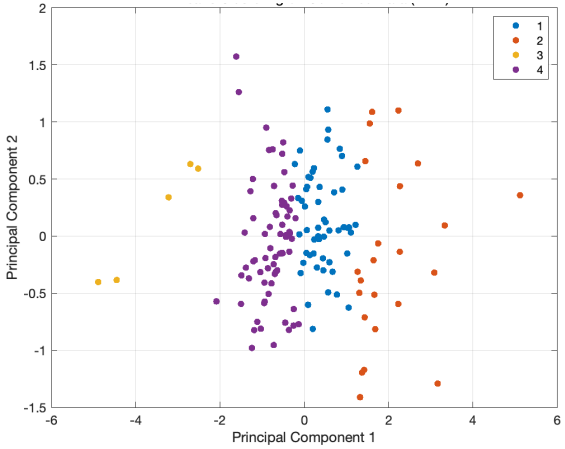
$$\rightarrow \quad U = c^T \Sigma_{xx}^{-1/2} X = a^T X$$

$$V = d^T \Sigma_{yy}^{-1/2} Y = b^T Y.$$

Clustering and Classification

K means clustering analysis

- Using **only the first component in rCCA**
- Determination of the optimal number of k (1:10)
  - based on Cophenetic Correlation, Silhouette Coefficient, Residual Sum of Squares (RSS)



Optimal Number of K = 4

Mean silhouette: 0.523

→ **4 MDD Subtypes**

XGBoost

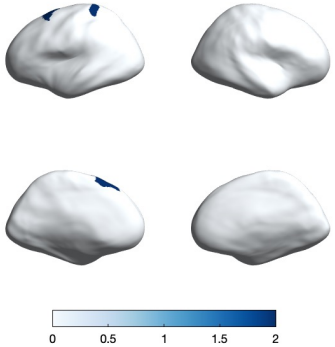
- Dataset split
  - Train set : Test set (75 : 25)
  - 5 folds cross validation
- Parameters
  - Numbers of estimators: 50, 100, 200
  - Learning rate: 0.01, 0.1, 0.2
  - Max depth: 3, 4, 5
  - Colsample by tree: 0.8, 0.9, 1.0
  - Subsample: 0.8, 0.9, 1.0

Model description

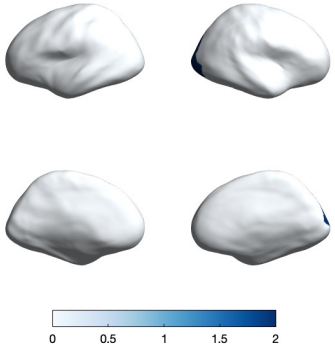
Model 1	Demographics + MetS
Model 2	Demographics + MetS + VBM NMF
Model 3	Demographics + MetS + VBM NMF + Subtypes

NMF-derived structural networks

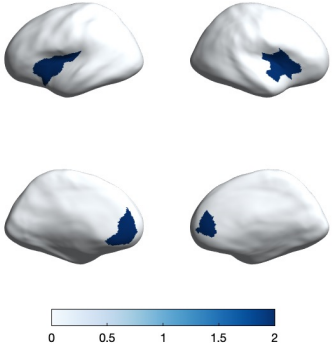
NMF Component 1



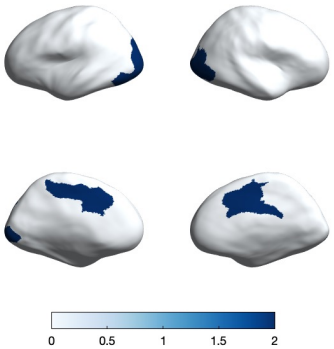
NMF Component 2



NMF Component 3



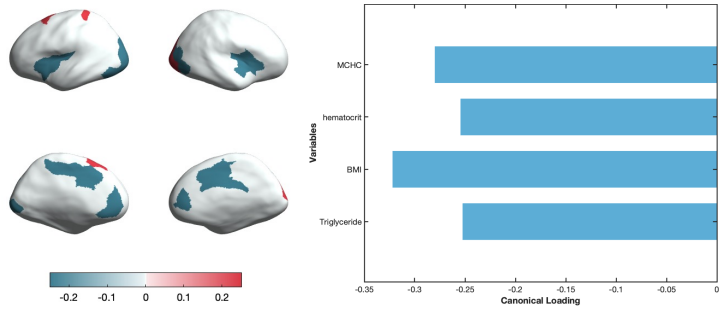
NMF Component 4



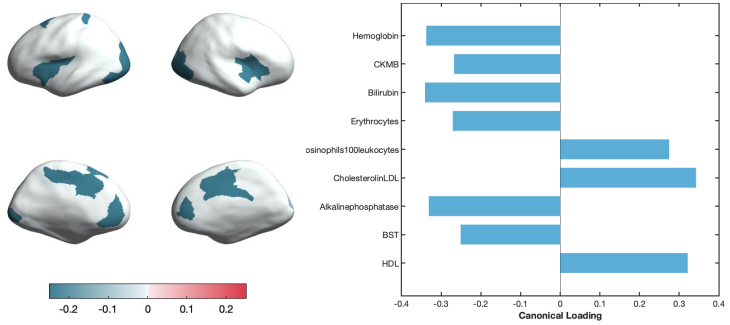
AJOU UNIVERSITY

CCA Multivariate patterns of brain imaging and clinical variables

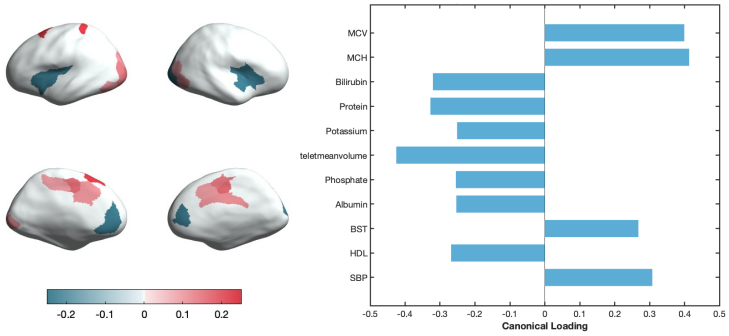
CCA Component 1



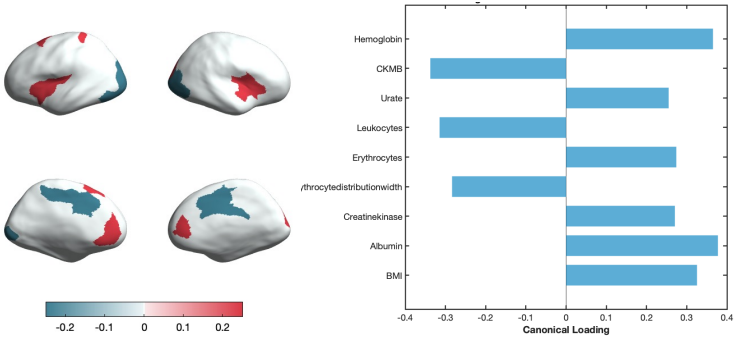
CCA Component 2



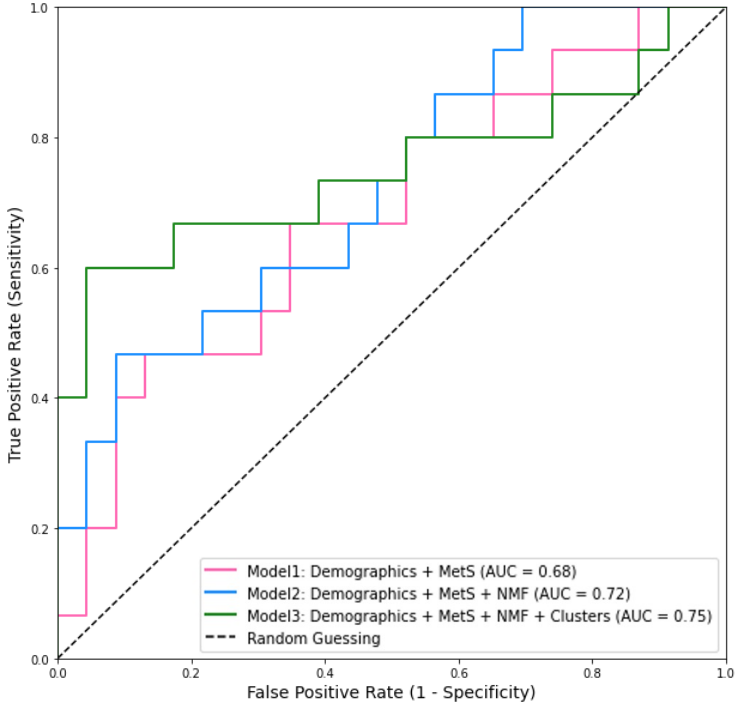
CCA Component 3



CCA Component 4



Classification Model Performance



	Model 1	Model 2	Model 3
AUROC	0.680	0.720	0.750
Accuracy	0.605	0.580	0.632
Precision	0.500	0.610	0.640
Recall	0.667	0.610	0.733
F1-score	0.571	0.580	0.611

Research Summary

- Through **the initial exploration of brain-body interactions**, we applied OMOP CDM to integrate brain imaging and clinical data for **classifying metabolic syndrome in MDD patients**
- Combined brain imaging data (reduced via NMF) with clinical variables using CCA
- Demonstrates **the potential of CDM** in pioneering **multimodal studies** and future scalability in **multi-organ interactions research**

Thank you

Authors

Sujin Gan, R.N

Narae Kim, M.S

Advisors

Rae Woong Park, M.D., PhD

Bumhee Park, PhD

CohortConstructor – an R package to support cohort building pipelines

Ed Burn

Cohorts

- Cohorts are a key building block in research studies – people fulfilling some criteria for some amount of time
- Established OHDSI tools (such as ATLAS/ Capr and CIRCE) allow us to define cohorts that can be instantiated in a database and stored in a library for future re-use
- **However**,
  - computational challenges remain when making many cohorts, and
  - bespoke cohort logic may not be supported by current tools

Many cohorts, all at once

Building cohorts by domain

- In OHDSI studies cohorts are typically defined independently and instantiated sequentially
- *CohortConstructor* builds cohorts by domain instead

```r
# libraries
library(CDMConnector)
library(CodelistGenerator)
library(CohortConstructor)

# connect to eunomia data
con <- DBI::dbConnect(duckdb::duckdb(), eunomia_dir())
cdm <- cdm_from_con(con,
                    cdm_schema = "main",
                    write_schema = "main")

# get drug codes
meds_cs <- getDrugIngredientCodes(cdm = cdm,
                                  name = c("acetaminophen",
                                           "aspirin",
                                           "celecoxib",
```

2:12 (Top Level)

Console | Terminal | Background Jobs

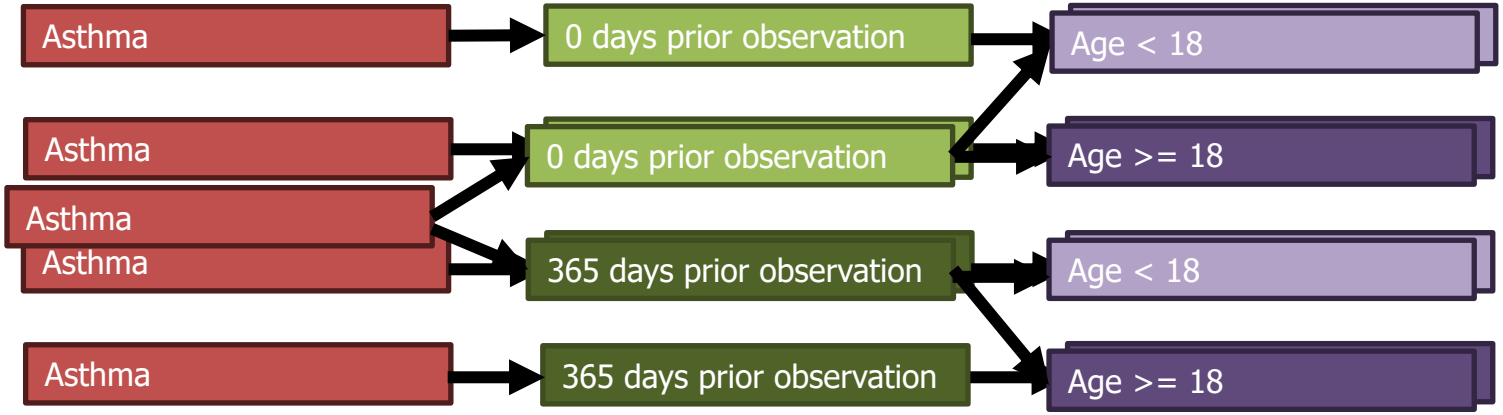
R 4.4.0 · ~/

```
                <int>          <int>          <int>
1                   1            137            137
2                   2           4379           1927
3                   3           1800           1800
4                   4          13908           2679
5                   5            830            830
6                   6             35             35
> cdm_disconnect(cdm)
>
```

From one cohort to many

Deriving cohorts from other cohorts

- Often, we build many cohorts that only vary slightly
- *CohortConstructor* encourages creating base cohorts from which study cohorts can be derived



```r
cdm$drug_cohorts <- cdm$drug_cohorts |>
  requirePriorObservation(c(0, 365))

settings(cdm$drug_cohorts)

cohortCount(cdm$drug_cohorts)

cdm$drug_cohorts <- cdm$drug_cohorts |>
```

33:1 (Top Level)

Console Terminal Background Jobs

R 4.4.0 · ~/

>

Flexible cohort pipelines

Cohort utilities

- Often, custom study-specific requirements need to be applied to a cohort
- CohortConstructor provides various utilities to support building bespoke cohorts
  - Add or subtract days from cohort entry and exit
  - Create cohorts based on age (e.g. entry on 18th birthday)
  - Reset entry/ exit on first/ last of set of date variables
  - Require cohort subjects are present in (or absence from) another cohort or table in some time window
  - Take a random sample of each cohort
  - Persist cohort entry across multiple observation periods (*next release*)

```r
cdm$hip_fx <- conceptCohort(cdm,
                            conceptSet = list("hip_fracture" = 4230399L),
                            name = "hip_fx",
                            exit = "event_start_date") |>
  padCohortEnd(180) |>
  requireIsFirstEntry() |>
  sampleCohorts(n = 100)

settings(cdm$hip_fx)

cohortCount(cdm$hip_fx)
```

1:1 (Top Level)

Console Terminal Background Jobs

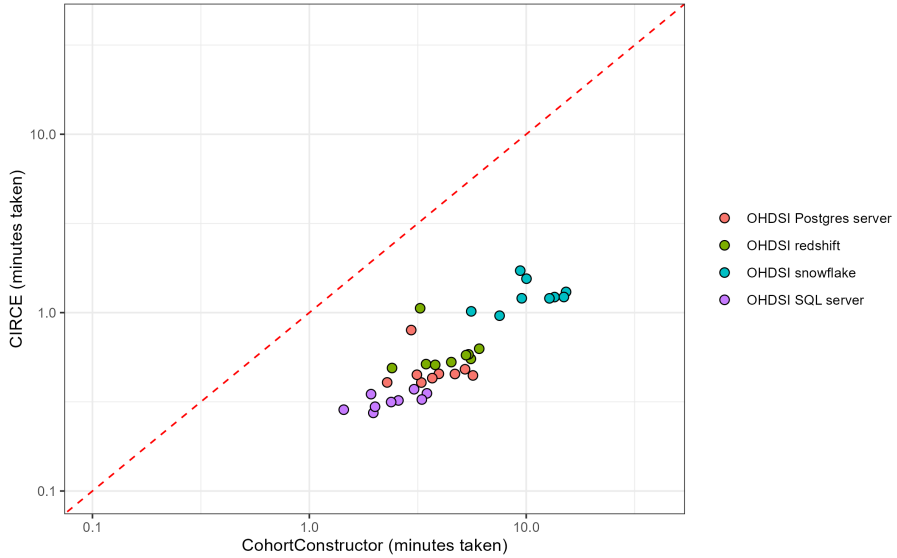
R 4.4.0 · ~/

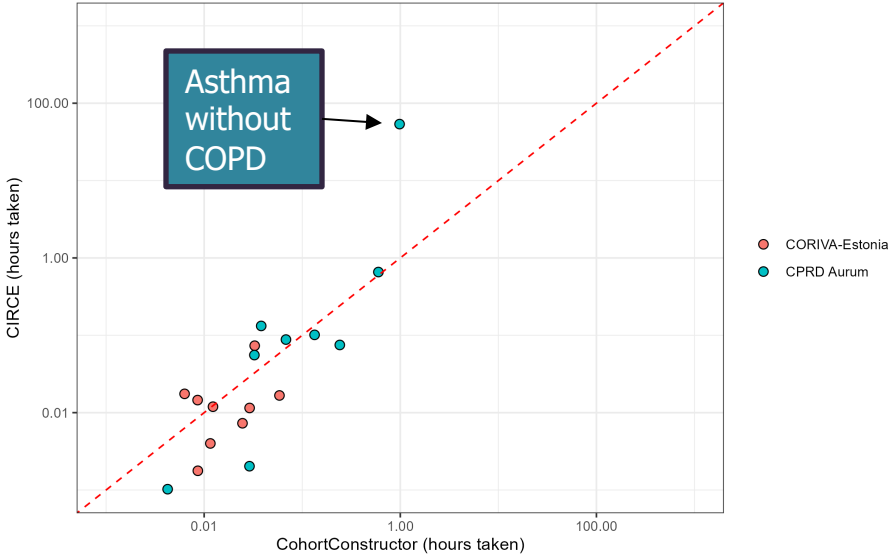
>

Benchmarking

Benchmark results

- Selected 9 cohorts from the OHDSI phenotype library





More information

Package website	ohdsi.github.io/CohortConstructor
GitHub	github.com/OHDSI/CohortConstructor
Benchmarking code	github.com/oxford-pharmacoepi/BenchmarkCohortConstructor

Contributors 9

Unlocking Efficiency in Real-world Collaborative Studies

A Multi-site International Study with **COLA-GLMM**

(**C**ollaborative **O**ne-shot **L**ossless **A**lgorithm for Generalized Linear Mixed Model)

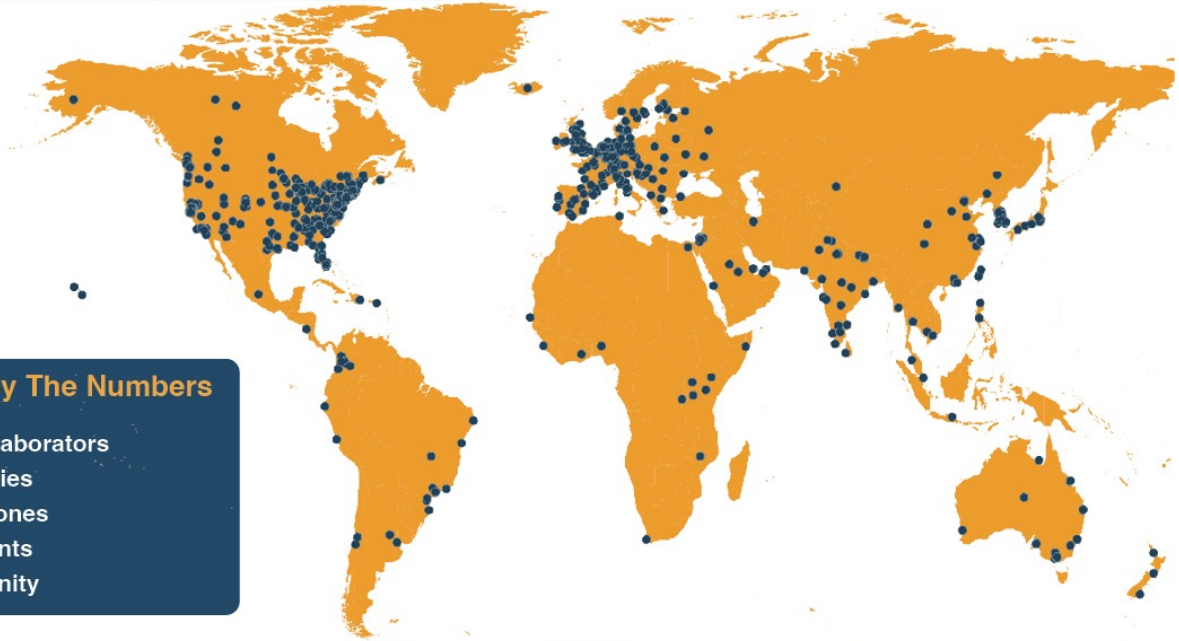
Jiayi (Jessie) Tong, Assistant Professor
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Oct 23, 2024, OHDSI 2024 Global Symposium

Map of Collaborators

The OHDSI community brings together volunteers from around the world to establish open community data standards, develop open-source software, conduct methodological research, and apply scientific best practices to both answer public health questions and generate reliable clinical evidence.

OHDSI Collaborators

Our community is ALWAYS seeking new collaborators. Do you want to focus on data standards or methodological research? Are you passionate about open-source development or clinical applications? Do you have data that you want to be part of global network studies? Do you want to be part of a global community that truly values the benefits of open science? Add a dot to the map below and JOIN THE JOURNEY!



OHDSI By The Numbers

- 3,266 collaborators
- 80 countries
- 21 time zones
- 6 continents
- 1 community

A **Primary Challenge** in Multi-site International Study

Individual Patient-level Data (IPD) cannot be shared across sites

- Regulatory Approval Processes
- Country-Specific Laws (e.g., HIPAA in USA, PIPEDA in Canada)
- Institutional Policies Data Sharing Restrictions

Privacy-Preserving Federated Learning Algorithms

- Enables fitting statistical models in a federated manner
- Requires summary statistics, instead of IPD
- Ensures data privacy and security

Multi-site Collaborative Study with Observational Data

-- An OHDSI Study Using Distributed **Linear** Mixed Model (DLMM)


nature COMMUNICATIONS

OHDSI

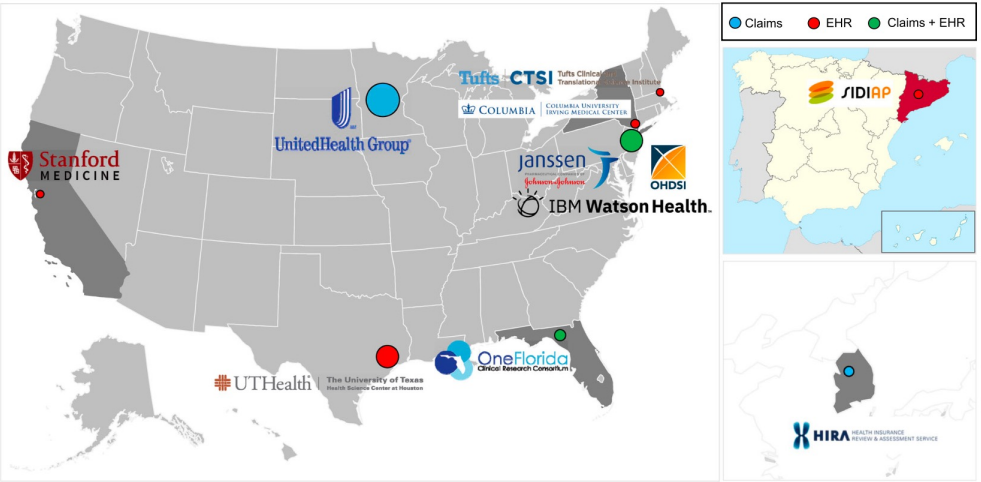
ARTICLE

https://doi.org/10.1038/s41467-022-29160-4 OPEN

DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models

Chongliang Luo[1,2], Md. Nazmul Islam[3], Natalie E. Sheils[3], John Buresh[3], Jenna Reps[4], Martijn J. Schuemie[4], Patrick B. Ryan[4], Mackenzie Edmondson[1], Rui Duan[1,5], Jiayi Tong[1], Arielle Marks-Anglin[1], Jiang Bian[6], Zhaoyi Chen[6], Talita Duarte-Salles[7], Sergio Fernández-Bertolín[7], Thomas Falconer[8], Chungsoo Kim[9], Rae Woong Park[9,10], Stephen R. Pfohl[11], Nigam H. Shah[11], Andrew E. Williams[12], Hua Xu[13], Yujia Zhou[13], Ebbing Lautenbach[1,14,15], Jalpa A. Doshi[16,17], Rachel M. Werner[16,17,18], David A. Asch[16,17] & Yong Chen[1]✉

Luo et al, 2022, Nature Communications



11 databases from 3 countries

Outcome of interest: Length of Stay (continuous outcome)

Investigation on the associations between demographic and clinical characteristics and length of hospital stay in COVID-19 patients

Luo C, Islam MN, Sheils NE, Buresh J, Reps J, Schuemie MJ, Ryan PB, Edmondson M, Duan R, Tong J, Marks-Anglin A. DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. Nature communications. 2022 Mar 30;13(1):1678.

Two Ideal Properties of Federated Learning Algorithms

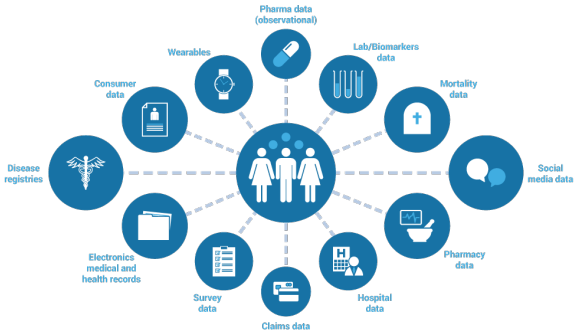
To date, only a few algorithms have successfully achieved both lossless and one-shot properties simultaneously:

- Linear Regression (i.e., Chen et al., 2006, IEEE)
- Linear mixed models (i.e., Luo et al., 2022, Nature Communications)

Luo C, Islam MN, Sheils NE, Buresh J, Reps J, Schuemie MJ, Ryan PB, Edmondson M, Duan R, Tong J, Marks-Anglin A. DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. Nature communications. 2022 Mar 30;13(1):1678.

Yixin Chen et al. Regression Cubes with Lossless Compression and Aggregation. IEEE Trans Knowl Data Eng 18, 1585–1599 (2006).

Challenges in Real-world Data



Non-continuous outcomes

- Binary outcome
- Categorical outcome
- Count outcome
- …

Between-site heterogeneity

We need **Federated Learning Algorithms** for **Generalized Linear Mixed Model (GLMM)**

GLMM Model: $y_{ij} = b_i + \boldsymbol{\beta} x_{ij} + \epsilon_{ij},$ **where** $\epsilon_{ij} \sim N(0, \sigma^2), b_i \sim N(0, \tau^2)$ **is site-specific random effect for k-th site,** β **is fixed effect.**

Existing Works on Federated Learning Algorithms for GLMM


Bioinformatics, 36, 2020, i128–i135
doi: 10.1093/bioinformatics/btaa478
ISMB 2020

Privacy-preserving construction of generalized linear mixed model for biomedical computation

Rui Zhu[1], Chao Jiang[2], Xiaofeng Wang[1], Shuang Wang[1], Hao Zheng[3] and Haixu Tang[1,*]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47405, USA, [2]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA and [3]Department of Bioinformatics, Hangzhou Nuowei Information Technology, Hangzhou 310053, China

*To whom correspondence should be addressed.

Zhu et al, 2020, *Bioinformatics*

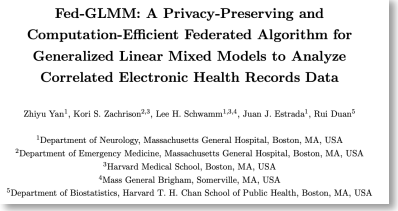

JAMIA
Featured as Editor's Choice

dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling

Chongliang Luo, Md Nazmul Islam, Natalie E Sheils, John Buresh, Martijn J Schuemie, Jalpa A Doshi, Rachel M Werner, David A Asch, Yong Chen

Journal of the American Medical Informatics Association, Volume 29, Issue 8, August 2022, Pages 1366–1371, https://doi.org/10.1093/jamia/ocac067

Luo et al, 2022, *JAMIA*



Fed-GLMM: A Privacy-Preserving and Computation-Efficient Federated Algorithm for Generalized Linear Mixed Models to Analyze Correlated Electronic Health Records Data

Zhiyu Yan[1], Kori S. Zachrison[2,3], Lee H. Schwamm[1,3,4], Juan J. Estrada[1], Rui Duan[5]

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[2]Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[4]Mass General Brigham, Somerville, MA, USA
[5]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

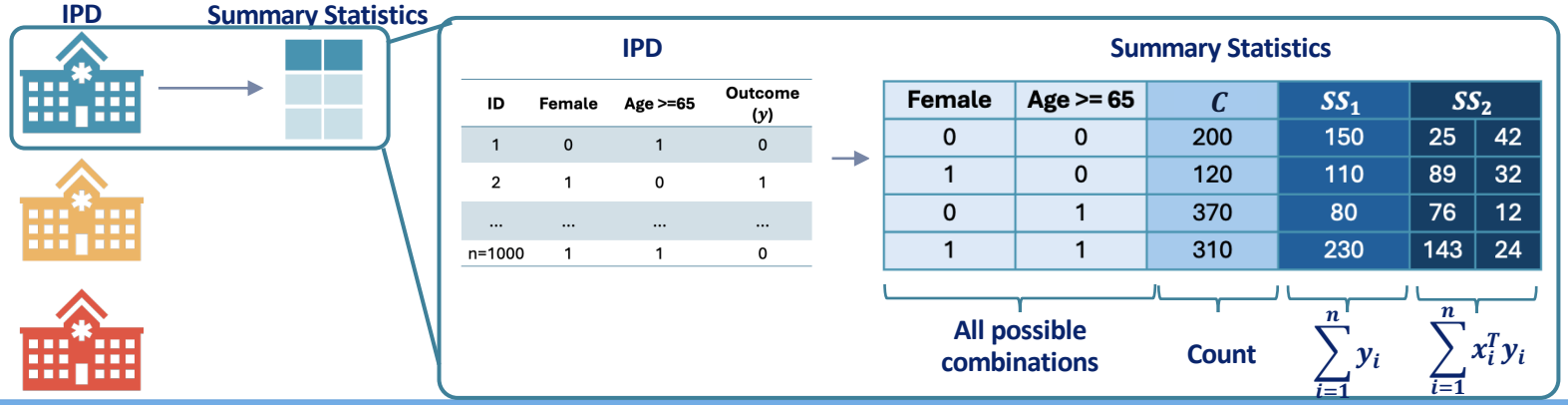
Yan et al, 2022, *arxiv*

	Zhu et al, 2020, *Bioinformatics*	Luo et al, 2022, *JAMIA*	Yan et al, 2022, *arxiv*	COLA-GLMM
Lossless	✗	✓	✗	✓
One-shot	✗	✗	!	✓
Communication Round	Iterative (500~1000 rounds)	< 5 rounds	1 or 2 rounds* (Depends on initialization)	1 round

Proposed Method – COLA-GLMM

Collaborative One-shot Lossless Algorithm for Generalized Linear Mixed Model

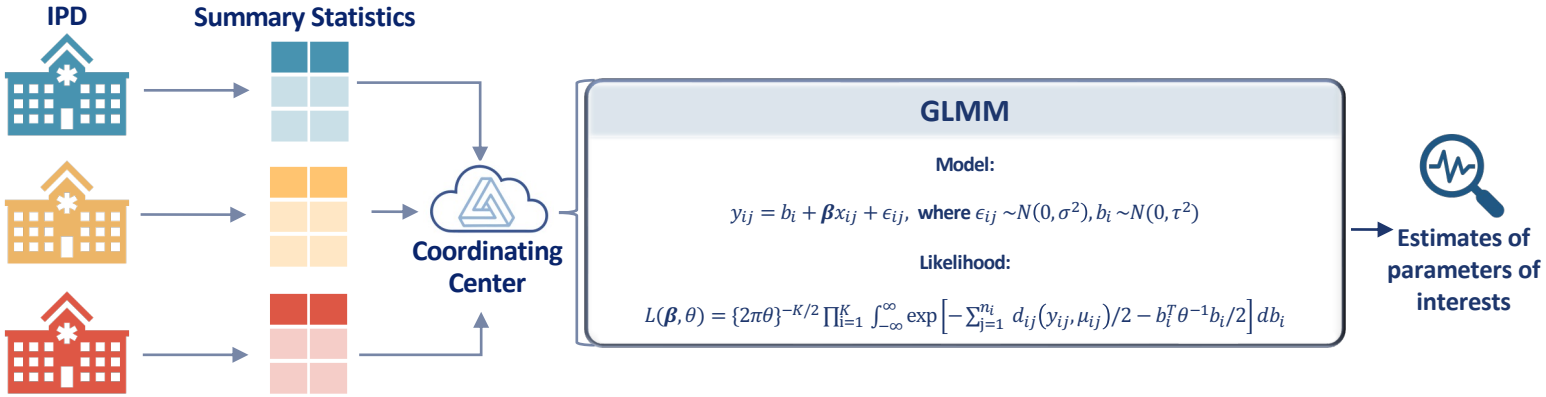
- Suppose that a **common** set of covariates are available at all collaborating sites.
- The covariates have been **standardized into categorical variables**.
- Pipeline:



Proposed Method – COLA-GLMM

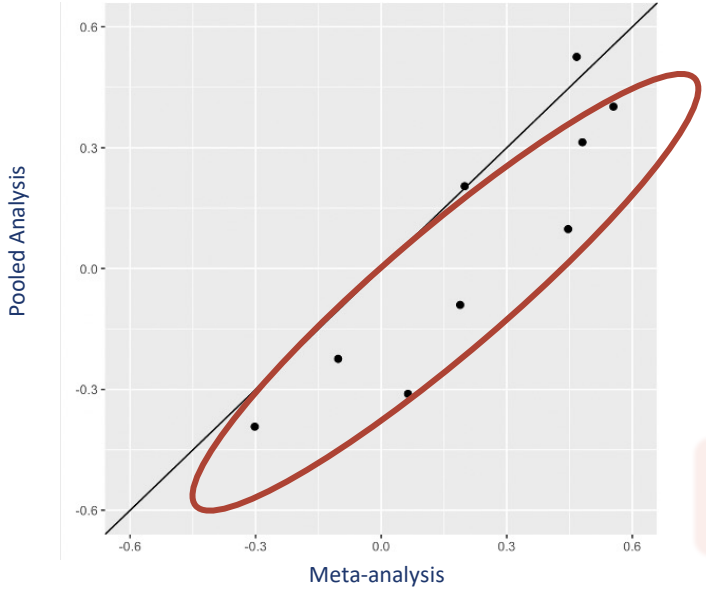
Collaborative One-shot Lossless Algorithm for Generalized Linear Mixed Model

- Suppose that a **common** set of covariates are available at all collaborating sites.
- The covariates have been **standardized into categorical variables**.
- Pipeline:



Simulation Study – Meta-analysis vs Pooled analysis

Estimated fixed effects (in log odds ratio)



Simulation setting:

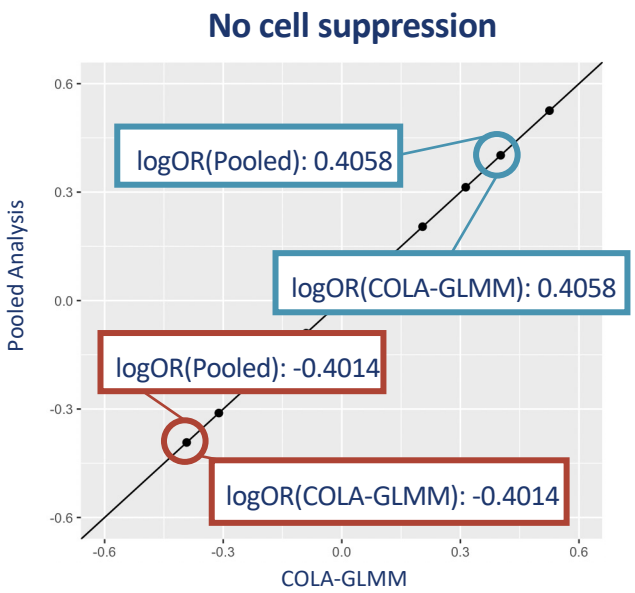
- 8 sites in total
- 9 risk factors
- Binary outcome
- Heterogeneous site-level random effects

Methods to compare:

- Pooled Analysis
- Meta-analysis

Meta-analysis has low accuracy

Simulation Study – Compare Pooled Analysis and COLA-GLMM



U.S. Dept. of Health & Human Services
Guidance Portal

Return to Search

CMS Cell Suppression Policy

Guidance for CMS Cell Suppression Policy Web Page

Final

Issued by: Centers for Medicare & Medicaid Services (CMS)

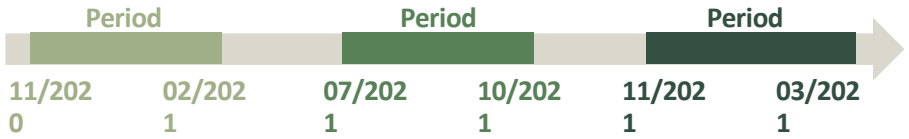
Raw value	Report
0	0
1-10	<11
>=11	as it

Real-world Case Study

- **Scientific Question:**

Identify COVID-19 mortality **risk factors** over **three time periods** among hospitalized patients

- **Study Period:**

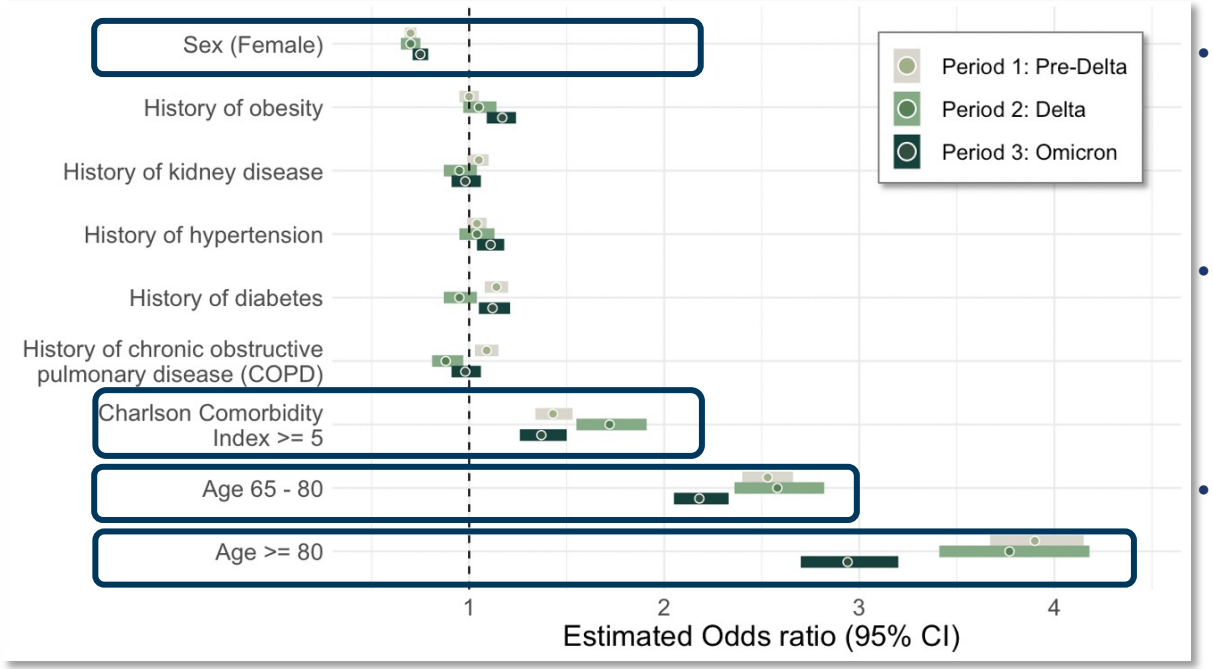


- **Databases (3 countries):**
  - Optum® de-identified Electronic Health Record Dataset (Optum EHR);
  - Optum's Clinformatics® Data Mart (CDM or Clinformatics®);
  - IQVIA Hospital CDM;
  - University of Florida Health;
  - Department of Veterans Affairs;
  - Integrated Primary Care Information (IPCI), The Netherlands;
  - Columbia University Irving Medical Center (CUIMC);
  - Parc Salut Mar Barcelona (PSMAR), Spain.

Inclusion criteria:

- Patients aged 18 years and older
- Had an inpatient visit with either a diagnosis of COVID-19 or a positive test for COVID-19 between 21 days prior to the inpatient visit and the end of the inpatient visit

Real-world Case Study Results



- **Sex (female):**
  - Reference group: Male
  - Female patients consistently exhibit a lower risk of mortality compared to males across all periods
- **Charlson Comorbidity Index (CCI):**
  - Reference group: CCI < 5
  - Higher CCI scores are statistically associated with an increased risk of mortality.
- **Age:**
  - Reference group: Age < 65
  - Higher age indicates significantly increased risk of mortality

Summary – COLA-GLMM

Collaborative **O**ne-shot **L**ossless **A**lgorithm for **G**eneralized **L**inear **M**ixed **M**odel

- **Lossless One-Shot**
- **Summary Statistics Only**
- **Heterogeneity-Aware**
- **Scalable, Applicable, and Implementation-Ready in OHDSI Network**

PDA R Package: 13300+ downloads since 2020

PDA Github Page: https://github.com/Penncil/pda

PDA website: https://pdamethods.org/

PDA-OTA: https://pda-ota.pdamethods.org/

Penn security office certified

Acknowledgements

Poster #117

- Yong Chen, University of Pennsylvania
- David A. Asch, University of Pennsylvania
- Jenna Reps, Janssen Research and Development
- Chongliang Luo, Washington University in St. Louis
- Yiwen Lu, University of Pennsylvania
- Milou T. Brand, Real World Solutions, IQVIA
- Scott L. DuVall, VA Informatics and Computing Infrastructure
- Thomas Falconer, Columbia University
- Juan Manuel Ramirez-Anguita, Hospital del Mar Research Institute (HMRIB)
- Miguel A. Mayer, Hospital del Mar Research Institute (HMRIB)
- Michael E Matheny, VA Informatics and Computing Infrastructure
- Alex Mayer Fuentes, Parc Taulí Hospital Universitari
- Xing He, University of Florida
- Bhavnisha K Patel, VA Informatics and Computing Infrastructure
- Katherine R Simon, VA Informatics and Computing Infrastructure
- Marc A. Suchard, University of California, Los Angeles
- Guojun Tang, University of Calgary
- Benjamin Viernes, VA Informatics and Computing Infrastructure
- Fei Wang, Weill Cornell Medicine
- Ross D. Williams, Erasmus University Medical Center
- Mui van Zandt, Real World Solutions, IQVIA
- Jiang Bian, University of Florida
- Jiayu Zhou, Michigan State University

Correspondence to:

- **Jessie Tong**, jtong20@jhu.edu
- **Yong Chen**, ychen123@upenn.edu

Department of Biostatistics, Epidemiology and Informatics

NCO-Calibrated DID Analysis: Addressing Unmeasured Confounding in Difference-in-Differences Analyses Using Negative Control Outcomes Experiments

Dazheng Zhang, Ph.D. candidate in Biostatistics at the University of Pennsylvania

2024 OHDSI Symposium

Advisor: Yong Chen, Ph.D., Professor of Biostatistics

Director of Center for Health AI & Synthesis of Evidence (CHASE), University of Pennsylvania

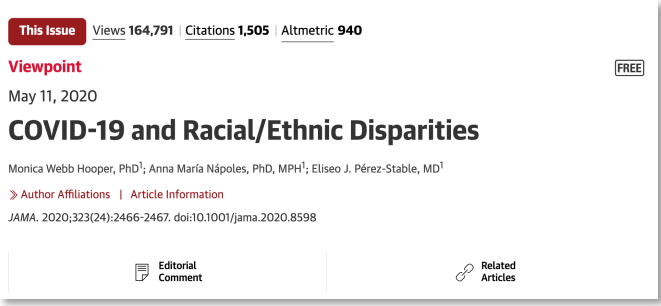
Joint work with Bingyu Zhang, Dr. Huiyuan Wang, Dr. Charles J. Wolock, Yiwen (Iris) Lu, Dr. Yong Chen

Real-World Case Study

- **Racial/Ethnic disparities** long lasting in healthcare.

Real-World Case Study

- **Racial/Ethnic disparities** long lasting in healthcare.
- Does the pandemic worsen racial/ethnic disparities?



This Issue Views 164,791 | Citations 1,505 | Altmetric 940

Viewpoint FREE

May 11, 2020

COVID-19 and Racial/Ethnic Disparities

Monica Webb Hooper, PhD[1]; Anna María Nápoles, PhD, MPH[1]; Eliseo J. Pérez-Stable, MD[1]

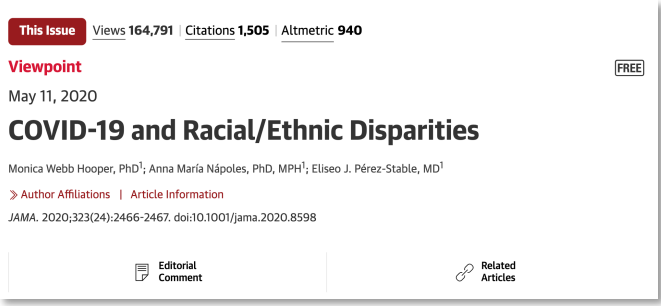
Author Affiliations | Article Information

JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598

Editorial Comment

Related Articles

Real-World Case Study


This Issue Views 164,791 | Citations 1,505 | Altmetric 940

Viewpoint FREE

May 11, 2020

COVID-19 and Racial/Ethnic Disparities

Monica Webb Hooper, PhD[1]; Anna María Nápoles, PhD, MPH[1]; Eliseo J. Pérez-Stable, MD[1]

Author Affiliations | Article Information

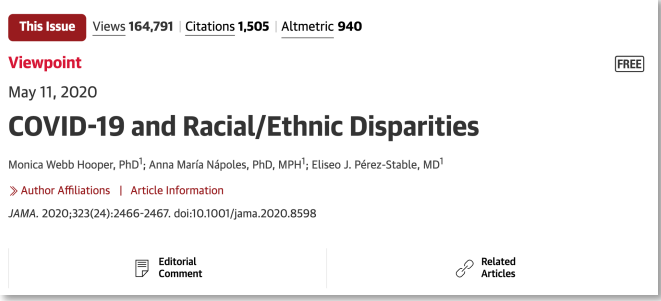
JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598

Editorial Comment

Related Articles

- **Racial/Ethnic disparities** long lasting in healthcare.
- Does the pandemic worsen racial/ethnic disparities?
  - **Difference-in-difference (DiD) approach** finds racial/ethnic disparities attributable to the pandemic while controlling for pre-existing disparities.

Real-World Case Study


This Issue Views 164,791 | Citations 1,505 | Altmetric 940

Viewpoint FREE

May 11, 2020

COVID-19 and Racial/Ethnic Disparities

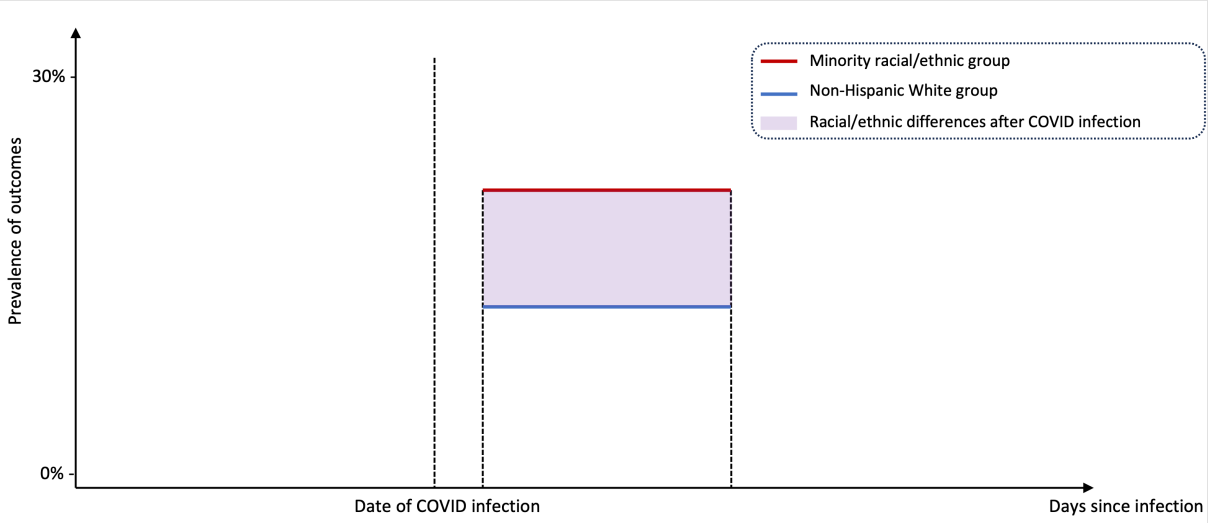
Monica Webb Hooper, PhD[1]; Anna María Nápoles, PhD, MPH[1]; Eliseo J. Pérez-Stable, MD[1]

Author Affiliations | Article Information

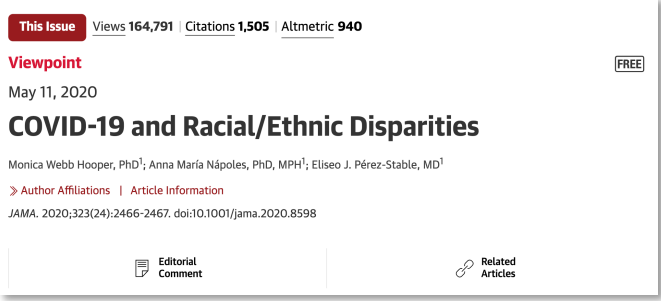
JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598

Editorial Comment | Related Articles

- **Racial/Ethnic disparities** long lasting in healthcare.
- Does the pandemic worsen racial/ethnic disparities?
  - **Difference-in-difference (DiD) approach** finds racial/ethnic disparities attributable to the pandemic while controlling for pre-existing disparities.



Real-World Case Study


This Issue Views 164,791 | Citations 1,505 | Altmetric 940

Viewpoint FREE

May 11, 2020

COVID-19 and Racial/Ethnic Disparities

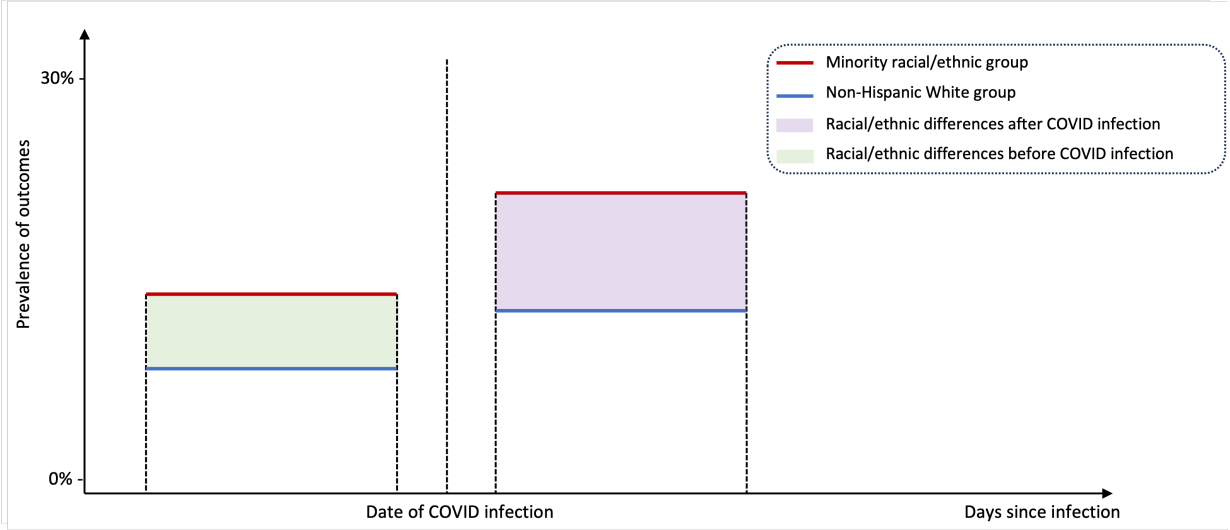
Monica Webb Hooper, PhD[1]; Anna María Nápoles, PhD, MPH[1]; Eliseo J. Pérez-Stable, MD[1]

Author Affiliations | Article Information

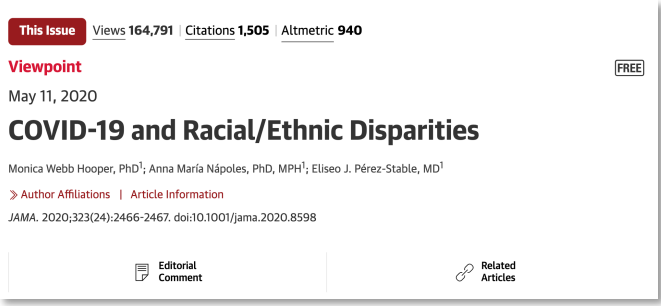
JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598

Editorial Comment | Related Articles

- **Racial/Ethnic disparities** long lasting in healthcare.
- Does the pandemic worsen racial/ethnic disparities?
  - **Difference-in-difference (DiD) approach** finds racial/ethnic disparities attributable to the pandemic while controlling for pre-existing disparities.



Real-World Case Study


This Issue Views 164,791 | Citations 1,505 | Altmetric 940

Viewpoint FREE

May 11, 2020

COVID-19 and Racial/Ethnic Disparities

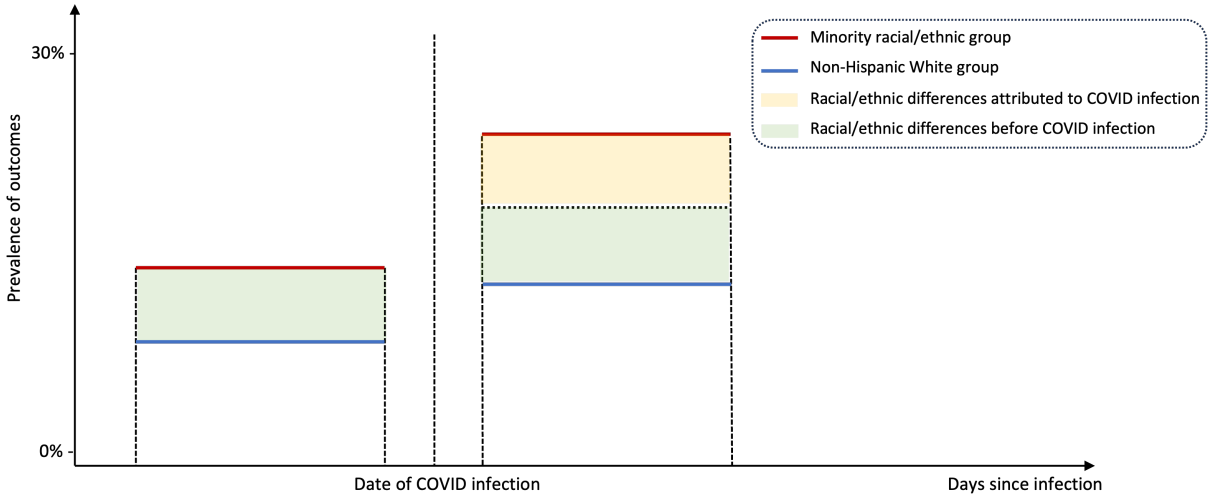
Monica Webb Hooper, PhD[1]; Anna María Nápoles, PhD, MPH[1]; Eliseo J. Pérez-Stable, MD[1]

Author Affiliations | Article Information

JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598

Editorial Comment | Related Articles

- **Racial/Ethnic disparities** long lasting in healthcare.
- Does the pandemic worsen racial/ethnic disparities?
  - **Difference-in-difference (DiD) approach** finds racial/ethnic disparities attributable to the pandemic while controlling for pre-existing disparities.

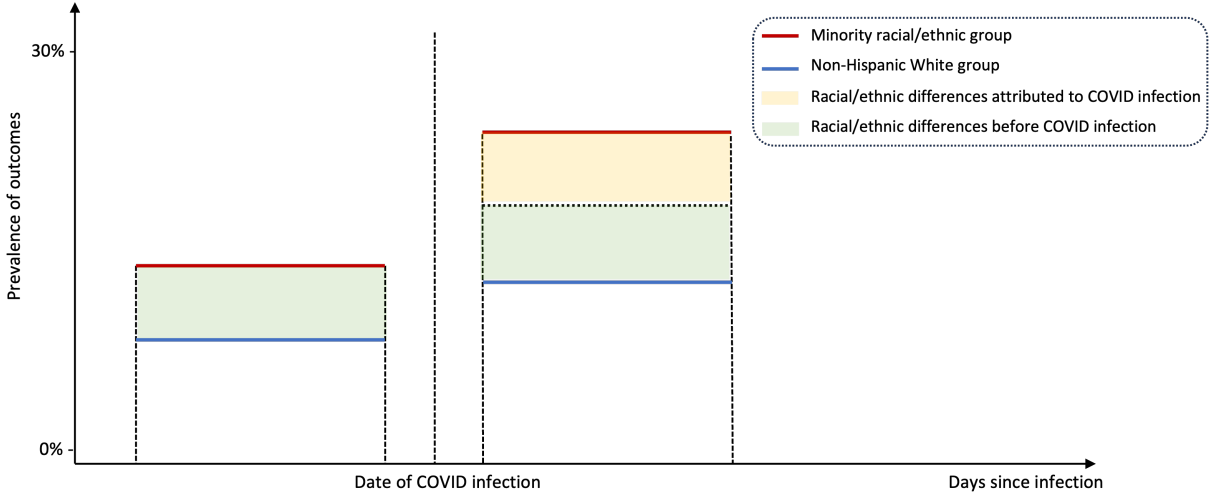


Parallel Trends Assumption in DiD Model

- **Parallel trends assumption**: in the absence of intervention, the before-intervention difference and the after-intervention should be the same.

Parallel Trends Assumption in DiD Model

- **Parallel trends assumption**: in the absence of intervention, the before-intervention difference and the after-intervention should be the same.

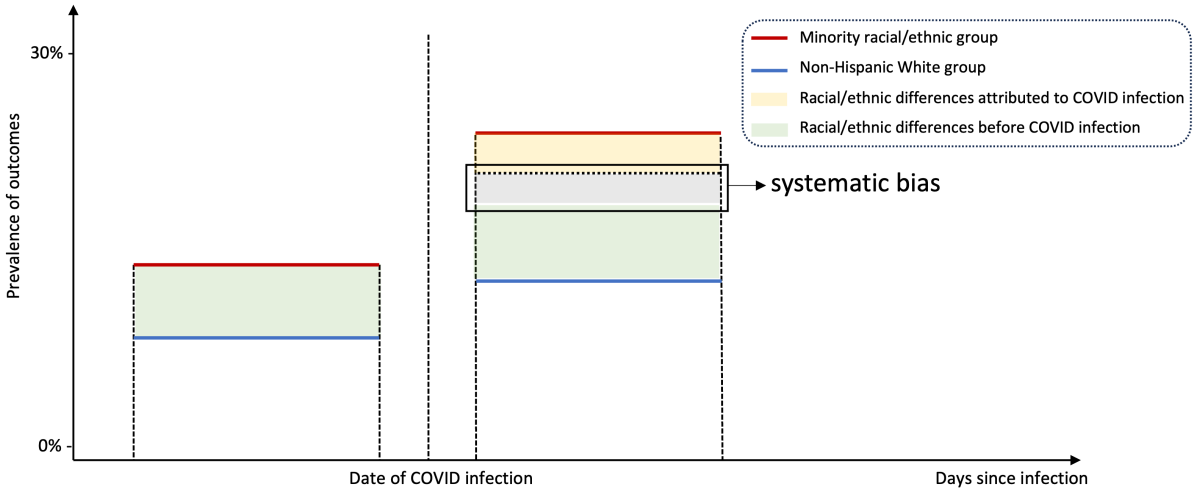


Parallel Trends Assumption in DiD Model

- **Parallel trends assumption**: in the absence of intervention, the before-intervention difference and the after-intervention should be the same.
- **Violation of parallel trends assumption**: systematic bias from unmeasured confounding variables makes the non-parallel trends for two groups.

Parallel Trends Assumption in DiD Model

- **Parallel trends assumption**: in the absence of intervention, the before-intervention difference and the after-intervention should be the same.
- **Violation of parallel trends assumption**: systematic bias from unmeasured confounding variables makes the non-parallel trends for two groups.



Negative Control Experiments Empirical Calibration

Negative control outcome (NCO), known **in priori** to be unrelated to exposure.

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California

Negative Control Experiments Empirical Calibration

Negative control outcome (NCO), known **in priori** to be unrelated to exposure.

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California

Legend study (Suchard et al. 2021 Lancet) used "ingrown nail" as an adverse event that is known to be unrelated to the antihypertension

JOURNAL ARTICLE

Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, Marc A Suchard

Journal of the American Medical Informatics Association, Volume 27, Issue 8, August 2020

Negative Control Experiments Empirical Calibration

Negative control outcome (NCO), known **in priori** to be unrelated to exposure.

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California

Legend study (Suchard et al. 2021 Lancet) used "ingrown nail" as an adverse event that is known to be unrelated to the antihypertension

JOURNAL ARTICLE

Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, Marc A Suchard

Journal of the American Medical Informatics Association, Volume 27, Issue 8, August 2020

An example list of 40 NCOs for a vaccine study.

Related Features VISUAL ABSTRACT

Original Research | 9 January 2024

Real-World Effectiveness of BNT162b2 Against Infection and Severe Diseases in Children and Adolescents

Authors: Qiong Wu, PhD, Jiayi Tong, MS, Bingyu Zhang, MS, Dazheng Zhang, MS, Jiajie Chen, PhD, Yuqing Lei, MS, Yiwen Lu, BS, ... SHOW ALL ..., and Yong Chen, PhD | AUTHOR, ARTICLE, & DISCLOSURE INFORMATION

Publication: Annals of Internal Medicine • Volume 177, Number 2 • https://doi.org/10.7326/M23-1754

Categories	Examples
Infectious and parasitic diseases	Impetigo, Tinea capitis, Tinea corporis, Insect bite
Diseases of the skin tissue	Contact dermatitis, Diaper rash, Acne
Diseases of the musculoskeletal system and connective tissue	Dislocations, Closed fracture of distal end of radius, Sprain of ankle, Scoliosis, Foot pain, Injury of free lower limb, Injury of upper extremity, Injury of right leg, Injury of left leg, Injury of right foot
Diseases of the nervous system	Seizure, Epilepsy, Concussion, Closed injury of head

Negative Control Experiments Empirical Calibration

Negative control outcome (NCO), known **in priori** to be unrelated to exposure.

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California

An example list of 40 NCOs for a vaccine study.

Related Features VISUAL ABSTRACT

Original Research | 9 January 2024

Real-World Effectiveness of BNT162b2 Against Infection and Severe Diseases in Children and Adolescents

Authors: Qiong Wu, PhD, Jiayi Tong, MS, Bingyu Zhang, MS, Dazheng Zhang, MS, Jiajie Chen, PhD, Yuqing Lei, MS, Yiwen Lu, BS, ... SHOW ALL ..., and Yong Chen, PhD AUTHOR, ARTICLE, & DISCLOSURE INFORMATION

Publication: Annals of Internal Medicine • Volume 177, Number 2 • https://doi.org/10.7326/M23-1754

Can we make empirical calibration for DiD model?

Categories	Examples
parasitic diseases	Impetigo, Tinea capitis, Tinea corporis, Insect bite
Diseases of the skin tissue	Contact dermatitis, Diaper rash, Acne
Diseases of the musculoskeletal system and connective tissue	Dislocations, Closed fracture of distal end of radius, Sprain of ankle, Scoliosis, Foot pain, Injury of free lower limb, Injury of upper extremity, Injury of right leg, Injury of left leg, Injury of right foot
Diseases of the nervous system	Seizure, Epilepsy, Concussion, Closed injury of head

"ingrown nail" as an adverse event that is known to be unrelated to the antihypertension

JOURNAL ARTICLE

Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, Marc A Suchard

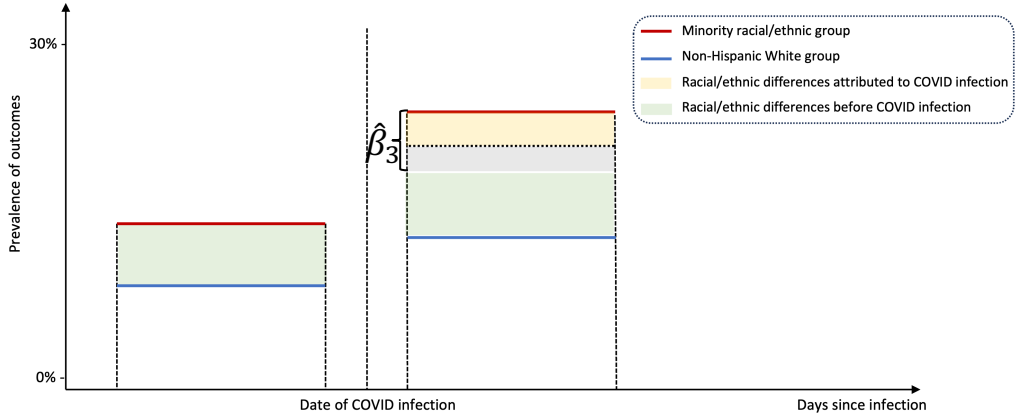
Journal of the American Medical Informatics Association, Volume 27, Issue 8, August 2020

Proposed Method: NCO-DiD

Proposed Method: NCO-DiD

- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$



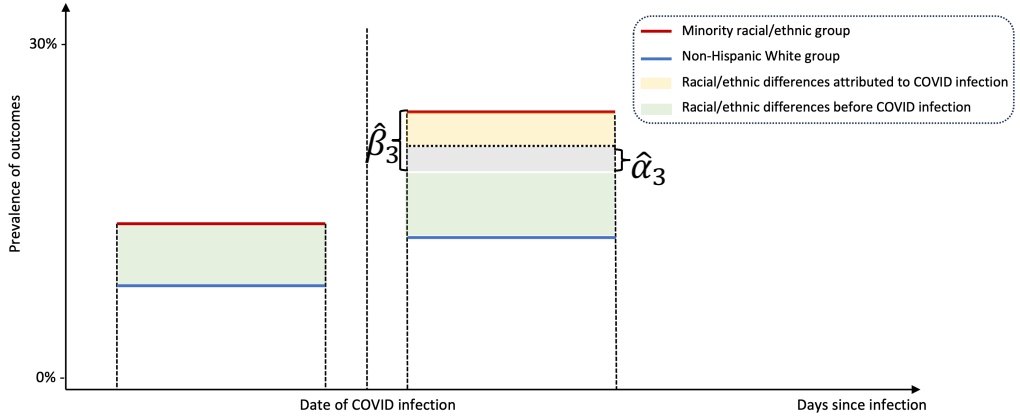
Proposed Method: NCO-DiD

- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$



Proposed Method: NCO-DiD

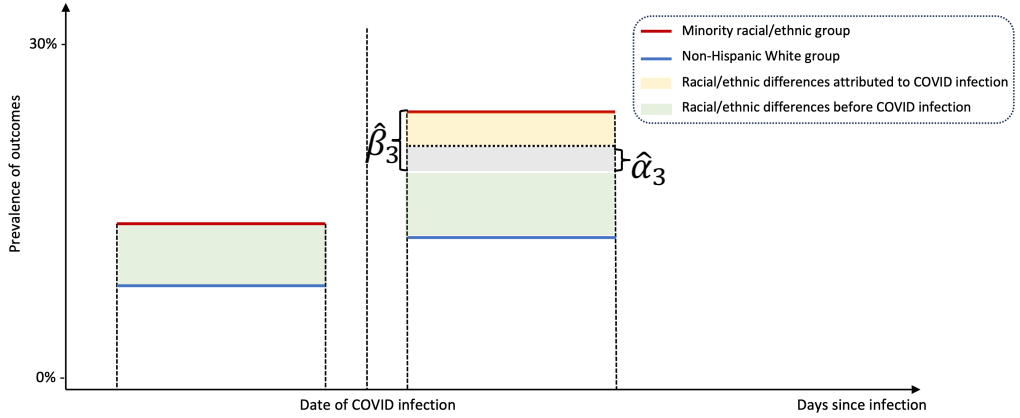
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Proposed Method: NCO-DiD

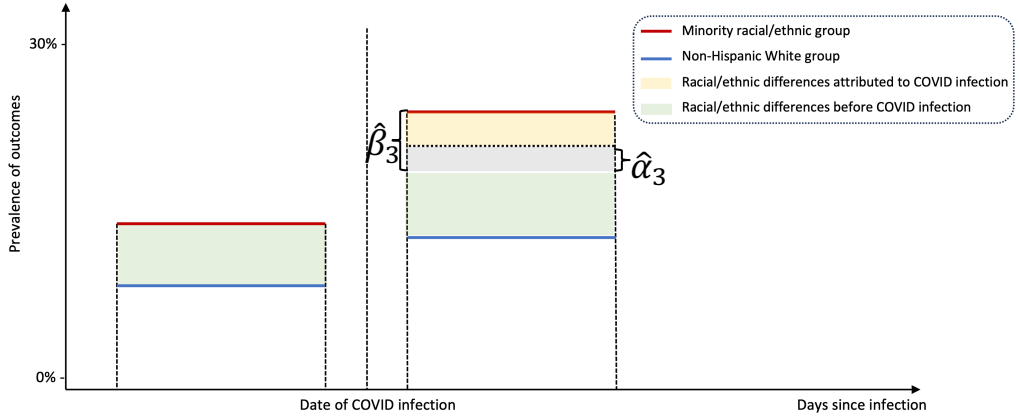
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Estimation for outcome of interest

$\hat{\beta}_3 = \text{systematic bias} + \beta_3$

Proposed Method: NCO-DiD

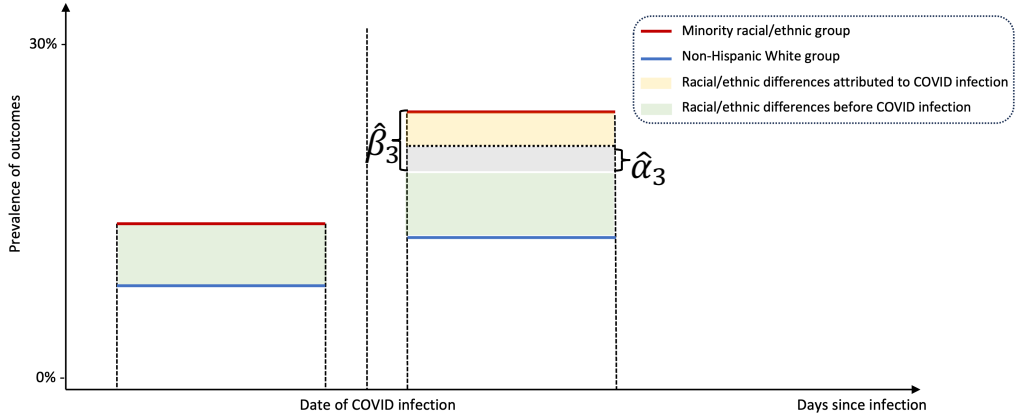
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Estimation for outcome of interest

$$\hat{\beta}_3 = \text{systematic bias} + \beta_3$$

Estimation for NCO:

$$\hat{\alpha}_3 = \text{systematic bias} + \alpha_3$$

Proposed Method: NCO-DiD

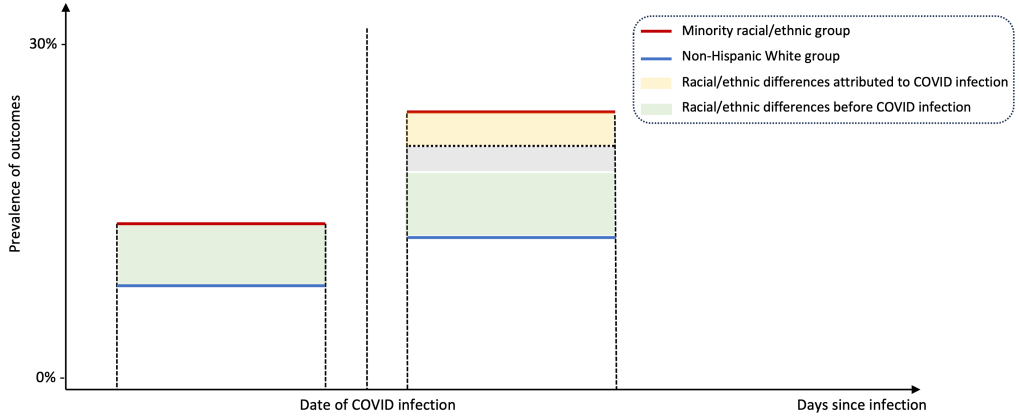
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Estimation for outcome of interest

$$\hat{\beta}_3 = \text{systematic bias} + \beta_3$$

Estimation for NCO:

$$\hat{\alpha}_3 = \text{systematic bias} + \mathbf{0}$$

Proposed Method: NCO-DiD

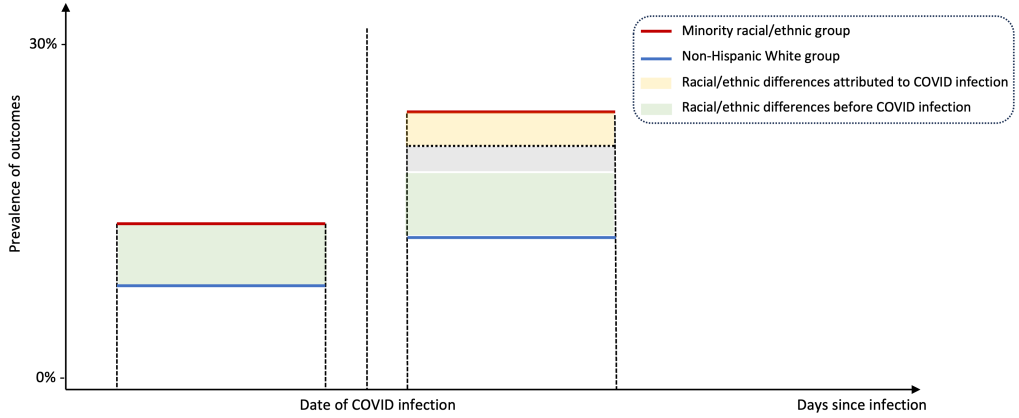
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from an NCO ($\hat{\alpha}_3$; true α_3 = 0).

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Estimation for outcome of interest

$$\hat{\beta}_3 = \text{systematic bias} + \beta_3$$

$$\boxed{\hat{\alpha}_3 \approx \text{systematic bias}}$$

Estimation for NCO:

$$\hat{\alpha}_3 = \text{systematic bias} + \mathbf{0}$$

Proposed Method: NCO-DiD

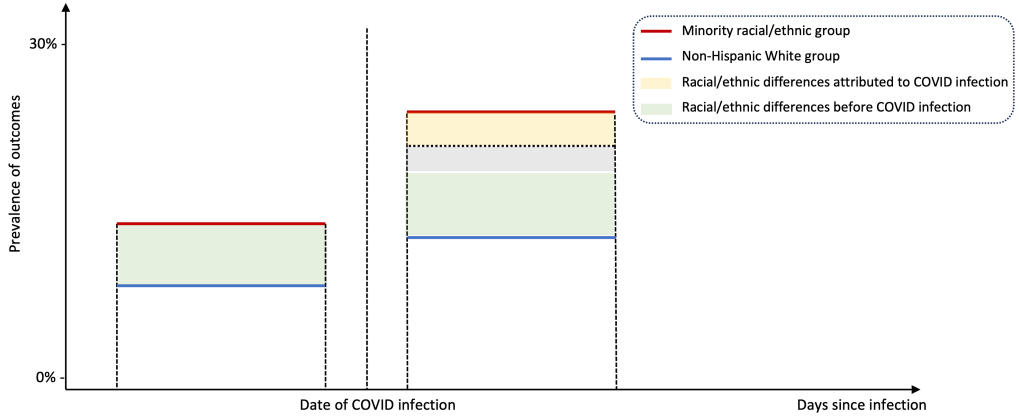
- Step 1: Fit a DiD model on the cohort (matched on measured confounders) to find estimate ($\hat{\beta}_3$)

$$\text{Outcome} \sim \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Intervention} + \beta_3 \text{Time} \times \text{Intervention}$$

- Step 2: Estimate systematic bias from M NCOs.

$$\text{NCO} \sim \alpha_0 + \alpha_1 \text{Time} + \alpha_2 \text{Intervention} + \alpha_3 \text{Time} \times \text{Intervention}$$

- Step 3: Empirically calibrate $\hat{\beta}_3 - \hat{\alpha}_3$



Estimation for outcome of interest

$$\hat{\beta}_3 = \text{systematic bias} + \beta_3$$

$$\boxed{\hat{\alpha}_3 \approx \text{systematic bias}}$$

Estimation for NCO:

$$\hat{\alpha}_3 = \text{systematic bias} + \mathbf{0}$$

Baseline Method vs Proposed Method

Baseline method (DiD)

Matching

- ❑ Learn propensity score.
- ❑ Construct matched cohort.

➡

Inference for intervention effect

- ❑ Fit DiD model.

Proposed method (NCO-DiD)

Matching

- ❑ Learn propensity score.
- ❑ Construct matched cohort.

➡

Inference for intervention effect

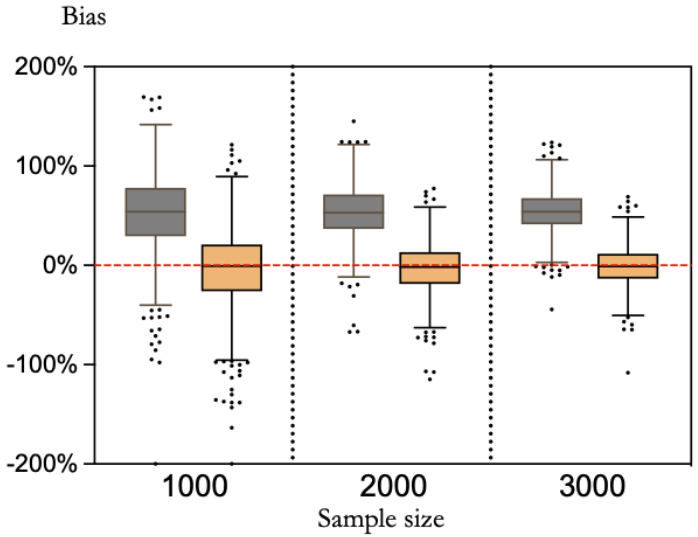
- ❑ Fit DiD model

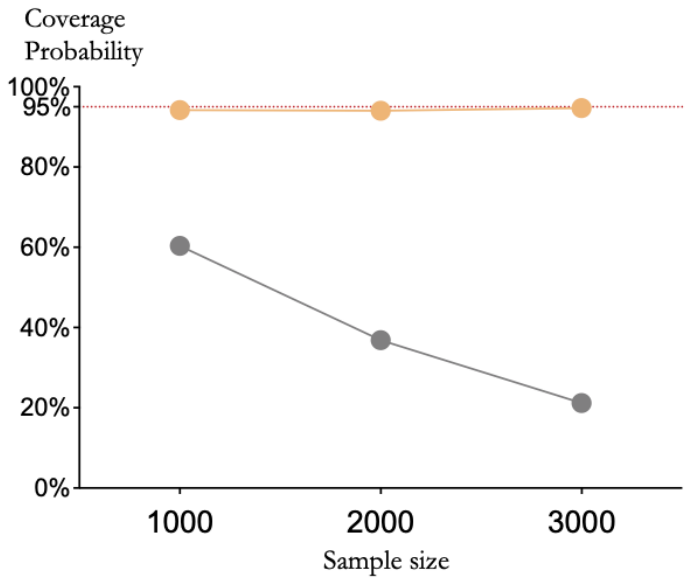
➡

Empirical Calibration

- ❑ Use NCOs to estimate systematic bias.
- ❑ Calibrate systematic bias.

Results from NCO-DiD





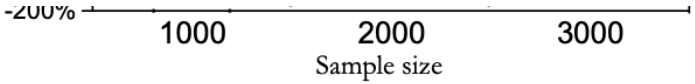
- Key message: across all scenarios, the proposed NCO-DiD model successfully calibrates the systematic bias.

Results from NCO-DiD

Coverage

Racial/Ethnic Differences in Long-COVID-Associated Symptoms among Pediatrics Population: Findings from Difference-in-differences Analyses in RECOVER Program

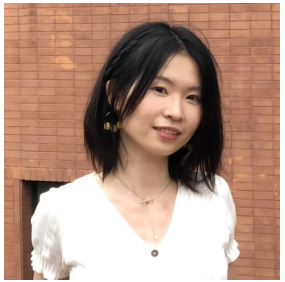
Author list:
Dazheng Zhang, MS[1^], Bingyu Zhang, MS[1,2^], Qiong Wu, PhD[1], Ting Zhou, PhD[1], Jiayi Tong, MS[1], Yiwen Lu[1,2], Jiajie Chen, PhD[1], Huiyuan Wang, PhD[1], Deena J Chisolm, PhD[3], Ravi Jhaveri, MD[4], Rachel C Kenney, PhD[5,6], Russell L Rothman, MD, MPP[7], Suchitra Rao, MD[8], David A Williams, MD[9], Mady Hornig, MA, MD[10], Linbo Wang, PhD[11], Jeffrey S Morris, PhD[12], Christopher B Forrest, MD, PhD[13], and Yong Chen, PhD[1,2] on behalf of the RECOVER consortium.



- Key message: across all scenarios, the proposed NCO-DiD model successfully calibrates the systematic bias.

Takeaway Messages

- Extend OHDSI empirical calibration framework.
- Control systematic bias from unmeasured confounding variables for DiD model.
- Help understand racial/ethnic differences in long COVID conditions after COVID-19 infection among children and adolescents.



Bingyu Zhang

Huiyuan Wang, Ph.D.

Charles J, Wolock, Ph.D.

Yiwen Lu

Yong Chen, Ph.D.

Acknowledgments:

Jiayi (Jessie) Tong, Ph.D.

Qiong Wu, Ph.D.,

Jiajie Chen, Ph.D.,

Lu Li,

Yuqing Lei,

and other lab members for your help with the project.

Poster: # 118

Oct 23, 3:00 pm - 5:00 pm

Code

Google scholar

Correspondence to:

Dazheng Zhang

Email:dazheng.zhang@pennmedicine.upenn.edu

Yong Chen, Ph.D.

Email: ychen123@upenn.edu

Health Trends Across Communities in Minnesota (HTAC-MN):

a Statewide Dashboard Leveraging the OMOP CDM to Monitor the Prevalence of Health Conditions

2024 OHDSI Symposium
October 23, 2024

Sam Patnoe, HealthPartners Institute
on behalf of the Minnesota EHR Consortium

HTAC-MN is funded through a Minnesota Public Health Infrastructure Grant from the Minnesota Department of Health.

Center for Community Health

MN EHR CONSORTIUM

Hennepin

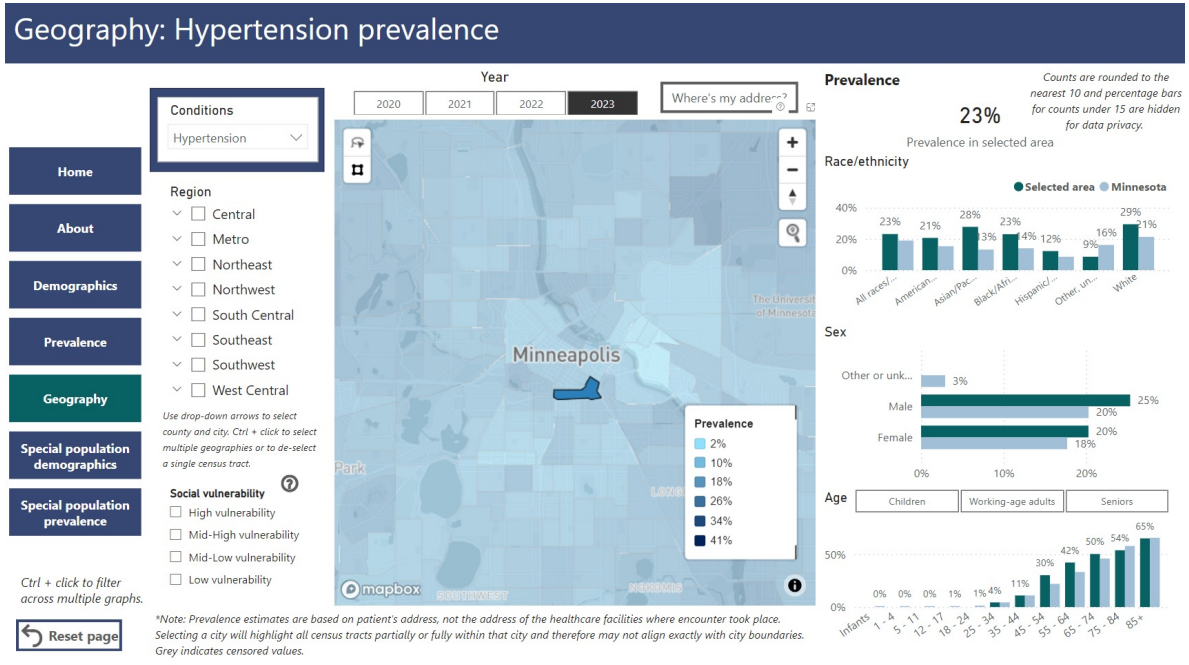
MN EHR Consortium

- Formed in March 2020 to address gaps in COVID-19 data sharing and communication
- 11 largest health systems in Minnesota
- >90% of residents in MN, all regions of state
- Federated data model - OMOP CDM (v5.3) adoption across all health systems from 2022-2023
- Health Trends Across Communities in Minnesota (HTAC-MN) began in 2023
  - Goal: Build a comprehensive statewide dashboard to support public health surveillance, inform community health assessments, and promote health equity

Allina Health
CentraCare
Children's Minnesota
Essentia Health
HealthPartners
Hennepin Healthcare
Mayo Clinic
Health Fairview
MN Community Measurement
Department of Health
North Memorial Health
Sanford Health

The HTAC-MN Dashboard

- Prevalence estimates for 30+ health conditions
  - 5.4M+ people (>90% of MN population)
- Race/ethnicity, age, sex
- State → census tract
- Data for 2020-2023
- Timely – refreshed annually



mnehrconsortium.org/htac

Selecting and defining health conditions

1. Identified and prioritized health conditions for dashboard:

- Public health significance
- Potential for action
- Lack of/limitations of existing data
- Emerging conditions
- Alignment with current public health priorities
- Detailed EHR data could support assessment work

MN EHR CONSORTIUM

Center for Community Health

2. Selected health conditions:

Chronic Conditions

- Asthma
- COPD
- Chronic kidney disease
- Diabetes, Type 2
- Heart failure
- Hyperlipidemia
- Hypertension
- Ischemic heart disease
- Obesity
- Peripheral vascular disease

Substance Use

- Alcohol
- Cannabis
- Cocaine
- Hallucinogens
- Inhalants
- Opioids
- Psychostimulants
- Sedatives

Mental Health

- Anxiety
- Bipolar disorder
- Depression
- PTSD
- Psychotic disorders
- Suicidal ideation or recent attempt

Maternal & Child Health

- Obstetrical deliveries
- Severe maternal morbidity
- Maternal opioid use

Other

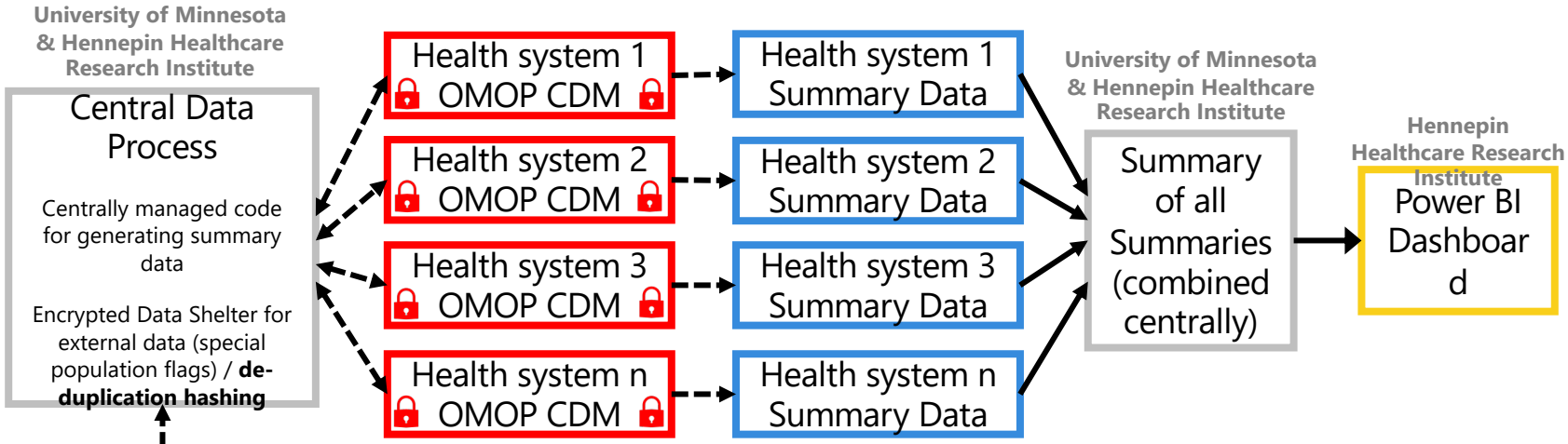
- Acute myocardial infarction
- Firearm injury
- Lung cancer
- Stroke

MN EHR CONSORTIUM

3. Developed standardized OMOP concept sets for each condition:

- Mapped existing ICD-10-CM diagnostic code sets to OMOP concepts (incl. SNOMED & ICD concepts)
- Accounts for metadata across MN EHR Consortium health systems
- Reviewed by clinicians
- Centrally managed
- Each condition is defined uniformly across all health systems

HTAC-MN Data Infrastructure



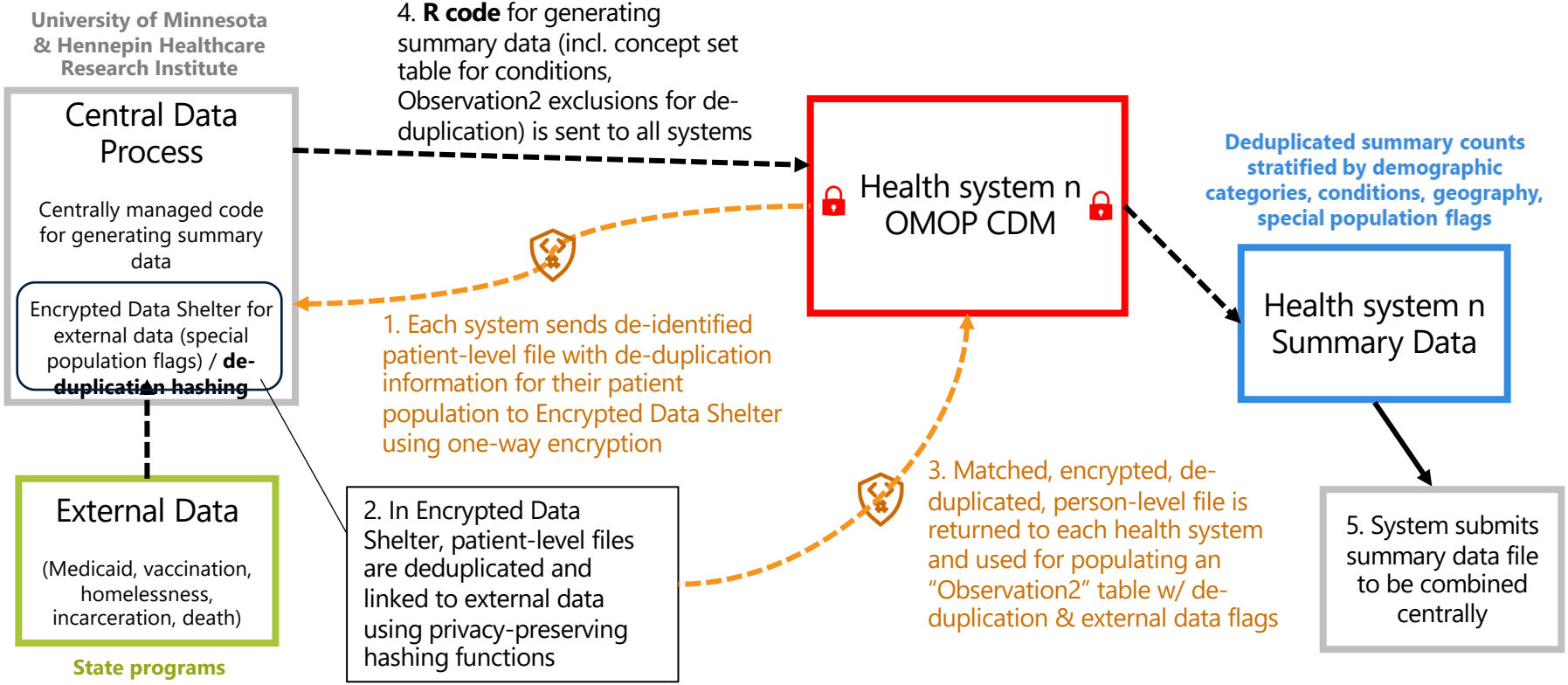
All 11 MN EHRC health systems have adopted the OMOP CDM v5.3

MN EHRC Customizations:

- Added "census_tract" column to LOCATION table
- Created "Observation2" table to store flags for external data and deduplication

Deduplicated summary counts stratified by demographic categories, conditions, geography, special population flags

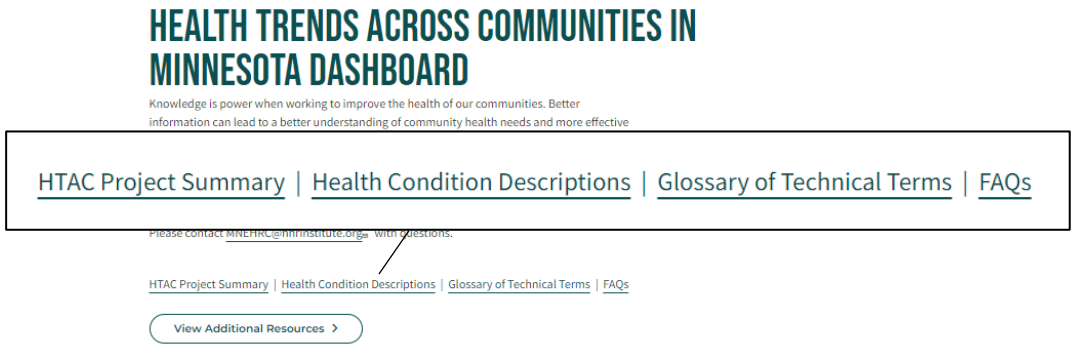
Data flow – health system perspective



Conclusion

- The HTAC-MN Dashboard demonstrates how the OMOP CDM can facilitate sharing summary EHR data across an entire state to monitor community health

- For more information, visit me at poster 119 and check out the website:



mnehrconsortium.org/htac

How Often: Characterizing Heterogeneity in Drug-Outcome Incidence Rate Estimates Attributed to Drug Indication

Results from the 2023 OHDSI Global Symposium

Hsin Yi Chen, BS, Christopher Knoll, BS, Elise Ruan, MD, MPH, Adam Black, BA, Sarah Seager, BA, Patrick Ryan, PhD, George Hripcsak, MD, MS

Incidence Rates

- Incidence rate calculations is one of the most common analyses in pharmacoepidemiology
  - Comparative background estimation
  - Drug Adverse Events

$$\text{Incidence Rate} = \frac{\begin{array}{c}\text{\# persons in the target cohort who have}\\ \text{new outcome occurance during the time-at-risk}\end{array}}{\begin{array}{c}\textbf{person-time-at-risk}\text{ for persons in the target cohort}\\ \text{with time at risk}\end{array}}$$

Why Incidence Rates?

- ”Simple” (fewer assumptions)
- Not causal, but still useful
  - If incidence is low and side effect is not serious, then we’re good
  - If incidence is high, then need to look out for it even if not caused by drug

…but incidence rate calculations can still be complex

- Heterogeneity in incidence rates estimates can be influenced by factors such as age, sex, calendar time, and differences in databases
  - Magnitude of potential impact… up to 1000 fold

Factors Influencing Background Incidence Rate Calculation: Systematic Empirical Evaluation Across an International Network of Observational Databases

Anna Ostropolets[1†], Xintong Li[2†], Rupa Makadia[3], Gowtham Rao[3], Peter R. Rijnbeek[4], Talita Duarte-Salles[5], Anthony G. Sena[3,4], Azza Shaoibi[3], Marc A. Suchard[6,7], Patrick B. Ryan[1,3], Daniel Prieto-Alhambra[2] and George Hripcsak[1,8]*

[1]Columbia University Medical Center, New York, NY, United States, [2]Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, United Kingdom, [3]Janssen Research and Development, Titusville, NJ, United States, [4]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands, [5]Fundacio Institut Universitari per a la Recerca a L'Atencio Primaria de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain, [6]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, United States, [7]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, CA, United States, [8]New York-Presbyterian Hospital, New York, NY, United States

OPEN ACCESS

Edited by:
Elisabetta Poluzzi, University of Bologna, Italy

Reviewed by:
Michele Fusaroli, University of Bologna, Italy
Angela Acosta, ICESI University, Colombia
Raquel Herrera Comoglio, National University of Cordoba, Argentina

***Correspondence:**
George Hripcsak
gh13@cumc.columbia.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Pharmacoepidemiology, a section of the journal Frontiers in Pharmacology

Received: *12 November 2021*
Accepted: *17 March 2022*
Published: *26 April 2022*

Citation:
Ostropolets A, Li X, Makadia R, Rao G, Rijnbeek PR, Duarte-Salles T, Sena AG, Shaoibi A, Suchard MA, Ryan PB, Prieto-Alhambra D and

Objective: Background incidence rates are routinely used in safety studies to evaluate an association of an exposure and outcome. Systematic research on sensitivity of rates to the choice of the study parameters is lacking.

Materials and Methods: We used 12 data sources to systematically examine the influence of age, race, sex, database, time-at-risk, season and year, prior observation and clean window on incidence rates using 15 adverse events of special interest for COVID-19 vaccines as an example. For binary comparisons we calculated incidence rate ratios and performed random-effect meta-analysis.

Results: We observed a wide variation of background rates that goes well beyond age and database effects previously observed. While rates vary up to a factor of 1,000 across age groups, even after adjusting for age and sex, the study showed residual bias due to the other parameters. Rates were highly influenced by the choice of anchoring (e.g., health visit, vaccination, or arbitrary date) for the time-at-risk start. Anchoring on a healthcare encounter yielded higher incidence comparing to a random date, especially for short time-at-risk. Incidence rates were highly influenced by the choice of the database (varying by up to a factor of 100), clean window choice and time-at-risk duration, and less so by secular or seasonal trends.

How does indication affect incidence rates?

- Drug indication is an obvious but usually ignored source of error
  - Beta blockers can be indicated for acute myocardial infarction (AMI) and hypertension, and of course, those taking a beta blocker for AMI are at higher risk of subsequent AMI
- We know incidence of different health outcomes differ by indication.
  - **What is the extent of the variation?**
  - **What is its relative contribution to heterogeneity compared to age, biological sex, and database?**

OHDSI Symposium October 2023

- How Often: Large scale characterization of incidence of outcomes following drug exposure
- Pre-Symposium
  - Draft protocol
  - Develop and evaluate phenotypes
  - Gathered research questions from OHDSI community
  - Release analysis package that has all the targets and outcomes of interest
- During Symposium (October 2023)
  - Execute How Often Analysis Package across OHDSI Network
  - Deploy viewer to allow for exploration of results
  - Collaborate on appropriate use of evidence
    - How to ensure reliability of results?
    - How to improve user interface to disseminate results?
    - What have we learned that can fill evidence gaps and improve decision making?

Method

- Analysis was conducted in October 2023 on 13 databases
- Study Design:
  - Target cohorts: First occurrence of drug exposure (12 different classes, stratified by indication)
  - Outcome cohorts: 73 different outcomes (defined in the OHDSI phenotype library)
  - Time at risk: 1 day to 365 day after cohort start (Intent to treat)
  - Stratifications: Age and gender

$$\text{Incidence Rate} = \frac{\text{\# persons in the target cohort who have new outcome occurance during the time-at-risk}}{\textbf{person-time-at-risk} \text{ for persons in the target cohort with time at risk}}$$

Method

Target cohorts:

12 Drug classes, nested by indication

	Indications
Beta Blockers	1) hypertension, 2) heart failure, 3) acute myocardial infarction
Cephalosporins	1) Urinary tract infection, 2) pneumonia
Calcium Channel Blockers	1) Hypertension
DPP-4 Inhibitors	1) Type 2 diabetes mellitus
Fluoroquinolones	1) Urinary tract infection, 2) pneumonia
GLP-1 antagonists	1) Type 2 diabetes mellitus, 2) obesity
IL-23 Inhibitors	1) Psoriasis
JAK inhibitors	1) Rheumatoid arthritis, 2) Ulcerative colitis
SGLT2 Inhibitors	1) Type 2 diabetes mellitus, 2) heart failure
Thiazide Diuretics	1) Hypertension
Trimethoprim	1) Urinary tract infection, 2) pneumonia
TNF-alpha inhibitors	1) Rheumatoid arthritis, 2) Psoriatic Arthritis, 3) Crohns disease, 4) Ulcerative colitis, 5) Psoriasis

Method

Outcomes Cohort examples (73 total)

Cardiovascular

- 3 and 4-point major adverse cardiovascular event (MACE) outcomes
- Cardiac death
- Torsades de Pointes
- Hospitalization with heart failure events

Neurologic

- Stroke
- Headache
- Guillen-Barre Syndrome (GBS)

Gastrointestinal

- Abdominal Pain
- Acute Liver Injury
- Diarrhea
- GI Bleed

Analysis

- Random effect meta-analysis of incidence rates across the 13 databases
- For drug classes with >1 indication: Variance components analysis to quantify relative heterogeneity between age, biological sex, database, and indication
- R `metafor` package (`rma`)

Results

- 77,631 total incidence rates calculated
- 8 different drug classes had at least 2 indications

Drug class	Indications	Median VC
Beta Blockers	1) Essential Hypertension, 2) Left Heart Failure, 3) Acute Myocardial Infarction	0.1013
SGLT2 Inhibitors	1) Type 2 Diabetes Mellitus, 2) Left Heart Failure	0.2642
GLP-1 antagonists	**1) Type 2 Diabetes Mellitus, 2) Obesity**	**<0.001**
Cephalosporins	1) Urinary Tract Infection, 2) Acute Typical Pneumonia	0.0397
Fluoroquinolones	1) Urinary Tract Infection, 2) Acute Typical Pneumonia	0.0983
Trimethoprim	**1) Urinary tract infection, 2) Pneumonia**	**0.4887**
JAK inhibitors	1) Rheumatoid Arthritis, 2) Ulcerative Colitis	0.0383
TNF-alpha inhibitors	1) Plaque Psoriasis, 2) Rheumatoid Arthritis, 3) Ulcerative Colitis, 4) Psoriatic Arthritis, 5) Crohn's Disease	0.0332

Relative Variance Components

Median Variance Components attributed to…

Drug class	Indications	Database	Age	Biological Sex
Beta Blockers	0.1013	0.1537	**0.3102**	0.0204
SGLT2 Inhibitors	0.2642	**0.3170**	0.2779	0.0155
GLP-1 antagonists	<0.001	**0.6117**	0.3678	0.0289
Cephalosporins	0.0397	0.7230	**1.4631**	0.0515
Fluoroquinolones	0.0983	**0.994**	0.8573	0.0696
Trimethoprim	0.4887	0.2772	**1.5228**	0.1219
JAK inhibitors	0.0383	0.1792	**0.2055**	0.0937
TNF-alpha inhibitors	0.0332	0.1675	**0.1815**	0.0221

Key Takeaways & Next Steps

- Among the drug classes we looked at, Trimethoprim is the drug class that is *most* sensitive to indications; GLP-1 the least
- Relative heterogeneity:
  - Database/Age > Indications > Biological Sex
- Next Steps: How Often All x All

Thank you! ☺