

Lightning Talks #1

Moderator: Davera Gabriel



Mapping of Critical Care EHR Flowsheet data to OMOP CDM via SSSOM

A Simple Standard for Sharing Ontology Mappings

Presenter: **Polina Talapova**, MD, PhD



sciforce



Tufts | CTSI Tufts Clinical and
Translational Science Institute

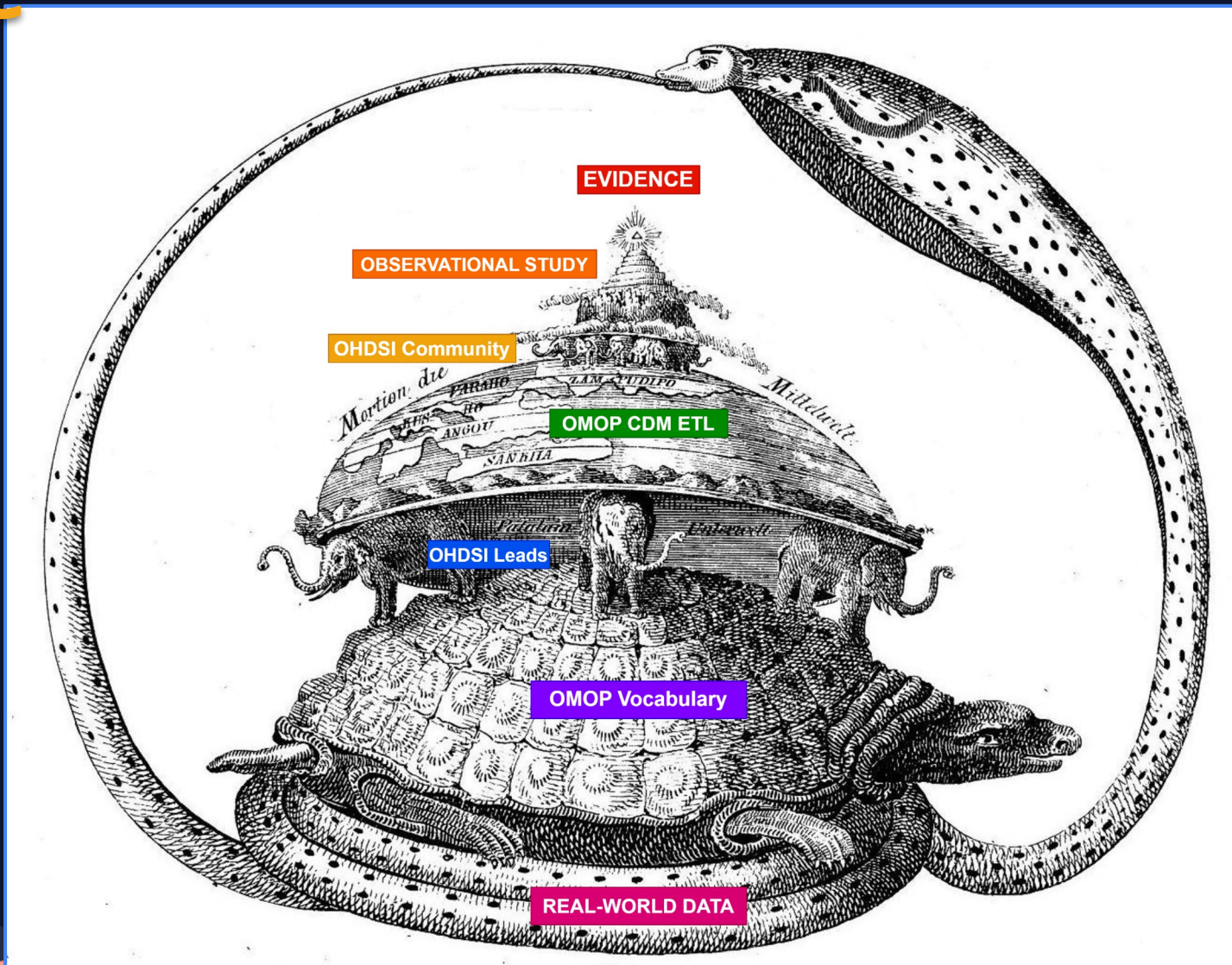


WE STAY



OHDSI

Prologue





Challenge [1]

Various Types of Critical / Intensive Care EHR Flowsheets

- Vital Signs
- Neurological Assessment
- Respiratory Assessment
- Cardiac Assessment
- Renal Assessment
- Intake and Output
- Gastrointestinal Assessment
- Nutritional Assessment
- Wound Care
- Pain Assessment
- Nursing

Semantic Domains

Measurements

Observations

Procedures

Conditions

Drugs

Devices





Challenge [2]

Health Data Mappings:

- Costly & use-case specific
- Essential for algorithm development and analytics
- Requires training & healthcare expertise

Open-Source Mappings:

- Lacking documentation & metadata
- Can lead to data inconsistencies

Adoption Challenges:

- Complicated by varied data sharing approaches





Challenge [3]

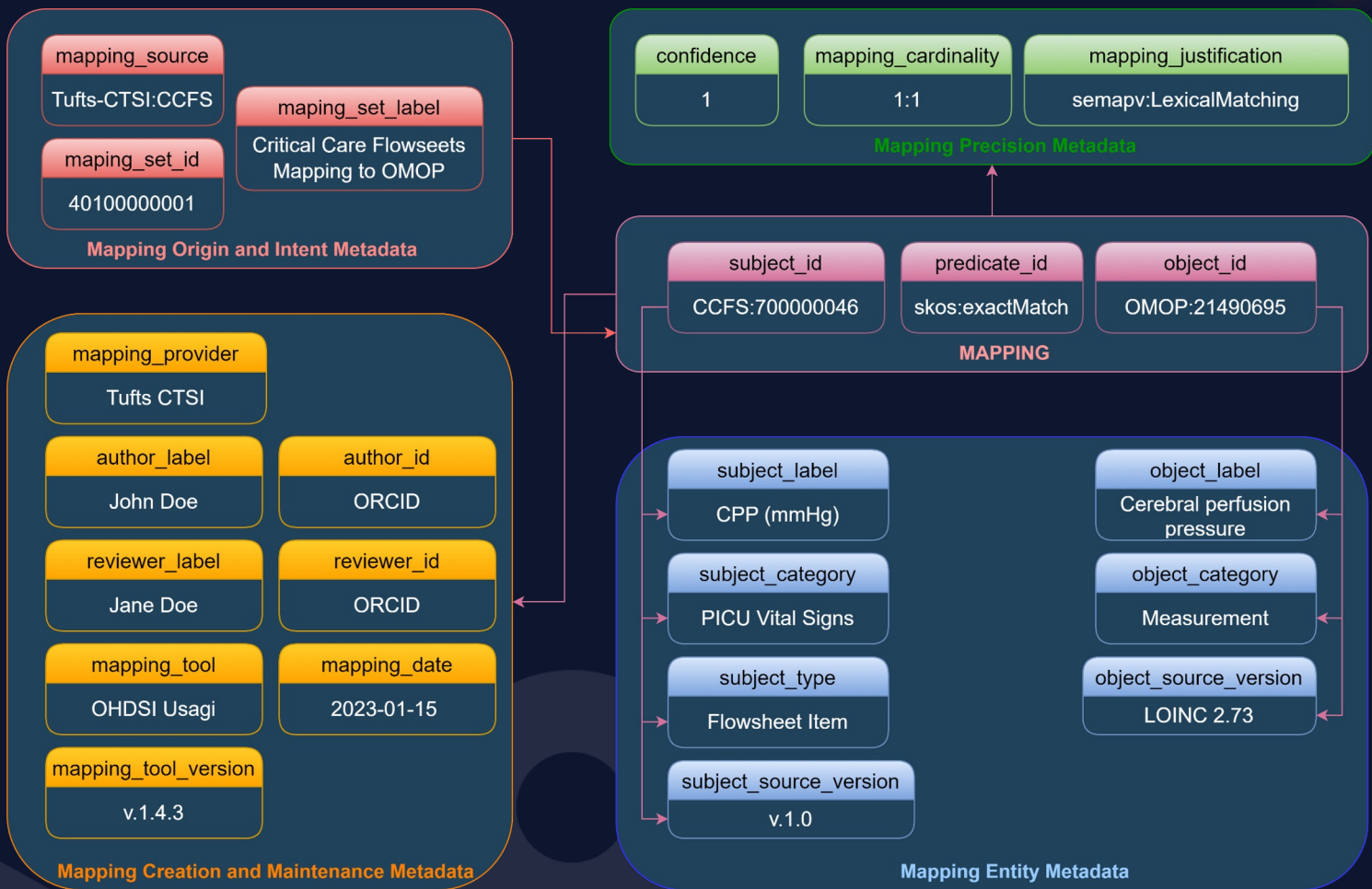
The OMOP Vocabularies, akin to a living organism, thrive with diligent care and stands to benefit from enhancements in areas such as:

- maintenance
- provenance
- precision
- mapping justification





Solution [1]: Generate SSSOM Metadata





Solution [2]: Use MAPPING_METADATA table

CDM Field	Datatype	Required
mapping_concept_id	integer	Yes
confidence	float	Yes
predicate_id	varchar	Yes
mapping_justification	varchar	Yes
mapping_provider	varchar	Yes
author_id	int	Yes
author_label	int	Yes
reviewer_id	int	Yes
reviewer_label	int	Yes
mapping_tool	varchar	No
mapping_tool_version	varchar	No
subject_category	varchar	No
subject_type	varchar	No





Solution [3]: Automation





Needs [1]: Integration with OHDSI tools



JACKALOPE



sciforce





Needs [2]: Community Contribution



sciforce





Visit our poster #501!



sciforce



SSSOM

SIMPLE STANDARD FOR SHARING
ONTOLOGY MAPPINGS



NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS,
RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES



Paving the way to estimate dose in OMOP CDM for Drug Utilisation Studies in DARWIN EU®

Theresa Burkard, PhD
Health Data Science Group – University of Oxford, UK

OHDSI US symposium - East Brunswick, USA
October 20, 2023



Is drug dosing valuable for pharmacoepidemiology studies?

YES

- as an inclusion criterion
- time trends of dosing
- high versus low dose

Background



WHO Collaborating Centre for
Drug Statistics Methodology

News

ATC/DDD Index

ATC code	Name	DDD	U	Adm.R
N02BE01	<u>paracetamol</u>	3	g	O
		3	g	P
		3	g	R

Daily Dose

Unit

Administration
Route



[WHOCC - ATC/DDD Index](#)

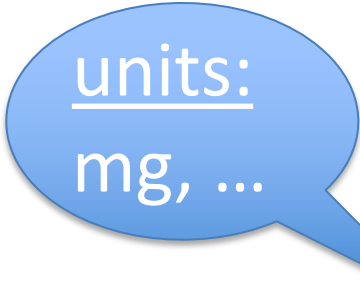
Objectives

Our aims were


- to introduce a uniform approach to develop dose formulas
- to validate suggested dose formulas

Objectives

Our aims were

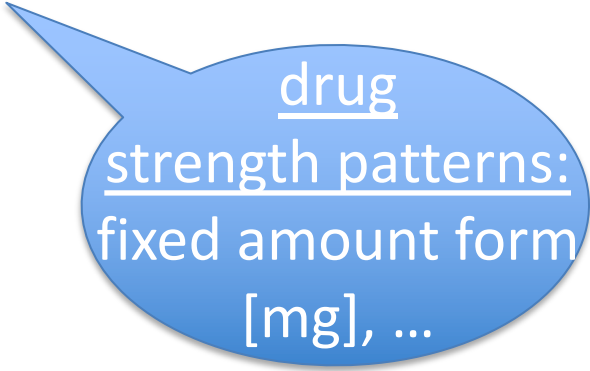


units:
mg, ...



routes:
oral, ...

- to introduce a uniform approach to develop dose formulas
- to validate suggested dose formulas



drug
strength patterns:
fixed amount form
[mg], ...

Drug strength patterns

31 patterns with clinically relevant units

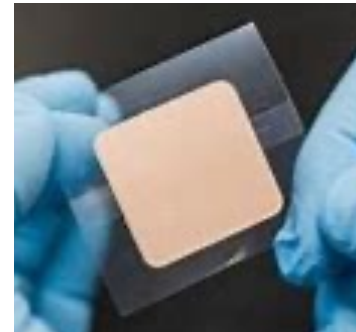
Fixed amount formulation patterns

e.g. pills, some injections, some inhalers



Time based formulation patterns

e.g. patches, extended release tablets



Concentration formulation patterns

e.g. mainly oral / injectable / inhalable solutions



Drug strength patterns

Examples of **Concentration formulation patterns**
e.g. mainly oral / injectable / inhalable solutions



DRUG STRENGTH TABLE

Concept name of drug concept id	Amount	Numerator	Concept name of Numerator unit	Denominator	Concept name of Denominator unit
2 ML ibuprofen 10 MG/ML Injection [Neoprofen]	NA	20	milligram	2	milliliter
itraconazole 10 MG/ML Oral Solution [Sporanox]	NA	10	milligram	NA	milliliter

Patterns

22 patterns with clinically relevant units

Drug strength patterns

Daily dose formulas (to be calculated per pattern):

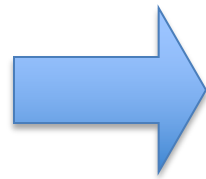
$$\frac{\text{Numerator value} * \text{quantity \{drug exposure table\}}}{\text{duration \{drug exposure table\}}}$$

clinical review
in CPRD AURUM/
GOLD (UK), IPCI (NL),
PharMetrics® Plus for
Academics (US)

pattern name	Oral route	Injectable route	Inhalable route
milliequivalent per milliliter	NA	NA	NA
milliequivalent per milliliter <i>missing denominator</i>	YES	NO	NA
milligram per actuation	NA	YES	YES
milligram per actuation <i>missing denominator</i>	YES	NO	YES
milligram per milligram	NO	NO	NO
milligram per milligram <i>missing denominator</i>	YES	NO	NO
milligram per milliliter	YES	YES	YES
milligram per milliliter <i>missing denominator</i>	YES	YES	YES
milliliter per milliliter	YES	YES	NA
milliliter per milliliter <i>missing denominator</i>	YES	NO	NA

Route through dose form: Poster 30
today: 4:15 – 5 pm
Sunday: noon – 1 pm

Validation of dose formulas



We estimated doses from 5 different ingredients in 5 different databases and compared them with the WHO Daily Dose

Ingredient list:

Concept Name	WHO DDD	Unit	Administration Route
furosemide	40	milligram	oral
	40	milligram	injectable
tiotropium	10	microgram	inhalable (powder)
	5	microgram	inhalable (solution)
metformin	2	gram	oral
enoxaparin	2	1000 IU	injectable
salmeterol	0.1	milligram	inhalable

WHO : World Health Organisation

DDD : Dispensed Daily Dose

IU : international unit

Dose finding and validation: Poster 502
today: 2:45 – 3:30 pm
Sunday: noon – 1 pm

Validation – Furosemide (WHO DDD: 40 mg oral / injectable)

	Unit (%), DD (median, IQR)	DD (median, IQR)	DD (median, IQR)
IQVIA Germany N = 1'375'495	[mg]: 93.3%, 40 mg (40-40) NA : 6.7%	oral and [mg]: 92.6%, 40 mg (40-40) inj. and [mg]: 0.6%, 40 mg (39-40) NA: 6.7%	“mg” [fixed] and oral: 92.3%, 40 mg (40-40) “mg/ml” [conc.] and oral: 0.3%, 10 mg (10-10) “mg/ml” [conc.] and inj.: 0.6%, 40 mg (39-40) NA : 6.7%
IPCI (NL) N = 2'694'879	[mg]: 99.8%, 40 mg (20-40) NA : 0.2%	oral and [mg]: 99.7%, 40 mg (20-40) inj. and [mg]: 0.2%, 1 mg (2-20) NA : 0.2%	“mg” [fixed] and oral: 99.6%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.1%, 0 mg (0-0) “mg/ml” [conc.] and oral: 0.0%, 20 mg (10-20) “mg/ml” [conc.] and inj.: 0.1%, 1 mg (2-20) “mg/ml*” [conc.] and inj.: 0.0%, 0 mg (0-0) NA : 0.2%
PharMetrics® Plus for Academics (US) N = 4'561'608	[mg] : 100%, 40 mg (20-40)	oral and [mg]: 93.3%, 40 mg (20-40) inj. and [mg]: 6.7%, 40 mg (20-80)	“mg” [fixed] and oral: 93.1%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.2%, 20 mg (12-30) “mg” [fixed] and inj.: 3.9%, 40 mg (20-40) “mg/ml” [conc.] and inj.: 2.8%, 80 mg (40-80) “mg/ml*” [conc.] and inj.: 0.0%, 20 mg (10-20)

* *Pattern with missing denominator*

Validation – Furosemide (WHO DDD: 40 mg oral / injectable)

	Unit (%), DD (median, IQR)	Route and unit (%) DD (median, IQR)	DD (median, IQR)
IQVIA Germany N = 1'375'495	[mg]: 93.3%, 40 mg (40-40) NA : 6.7%	oral and [mg]: 92.6%, 40 mg (40-40) inj. and [mg]: 0.6%, 40 mg (39-40) NA: 6.7%	“mg” [fixed] and oral: 92.3%, 40 mg (40-40) “mg/ml” [conc.] and oral: 0.3%, 10 mg (10-10) “mg/ml” [conc.] and inj.: 0.6%, 40 mg (39-40) NA : 6.7%
IPCI (NL) N = 2'694'879	[mg]: 99.8%, 40 mg (20-40) NA : 0.2%	oral and [mg]: 99.7%, 40 mg (20-40) inj. and [mg]: 0.2%, 1 mg (2-20) NA : 0.2%	“mg” [fixed] and oral: 99.6%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.1%, 0 mg (0-0) “mg/ml” [conc.] and oral: 0.0%, 20 mg (10-20) “mg/ml” [conc.] and inj.: 0.1%, 1 mg (2-20) “mg/ml*” [conc.] and inj.: 0.0%, 0 mg (0-0) NA : 0.2%
PharMetrics® Plus for Academics (US) N = 4'561'608	[mg] : 100%, 40 mg (20-40)	oral and [mg]: 93.3%, 40 mg (20-40) inj. and [mg]: 6.7%, 40 mg (20-80)	“mg” [fixed] and oral: 93.1%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.2%, 20 mg (12-30) “mg” [fixed] and inj.: 3.9%, 40 mg (20-40) “mg/ml” [conc.] and inj.: 2.8%, 80 mg (40-80) “mg/ml*” [conc.] and inj.: 0.0%, 20 mg (10-20)

* *Pattern with missing denominator*

Validation – Furosemide (WHO DDD: 40 mg oral / injectable)

	Unit (%), DD (median, IQR)	Route and unit (%) DD (median, IQR)	Pattern and route (%) DD (median, IQR)
IQVIA Germany N = 1'375'495	[mg]: 93.3%, 40 mg (40-40) NA : 6.7%	oral and [mg]: 92.6%, 40 mg (40-40) inj. and [mg]: 0.6%, 40 mg (39-40) NA: 6.7%	“mg” [fixed] and oral: 92.3%, 40 mg (40-40) “mg/ml” [conc.] and oral: 0.3%, 10 mg (10-10) “mg/ml” [conc.] and inj.: 0.6%, 40 mg (39-40) NA : 6.7%
IPCI (NL) N = 2'694'879	[mg]: 99.8%, 40 mg (20-40) NA : 0.2%	oral and [mg]: 99.7%, 40 mg (20-40) inj. and [mg]: 0.2%, 1 mg (2-20) NA : 0.2%	“mg” [fixed] and oral : 99.6%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.1%, 0 mg (0-0) “mg/ml” [conc.] and oral: 0.0%, 20 mg (10-20) “mg/ml” [conc.] and inj.: 0.1%, 1 mg (2-20) “mg/ml*” [conc.] and inj.: 0.0%, 0 mg (0-0) NA : 0.2%
PharMetrics® Plus for Academics (US) N = 4'561'608	[mg] : 100%, 40 mg (20-40)	oral and [mg]: 93.3%, 40 mg (20-40) inj. and [mg]: 6.7%, 40 mg (20-80)	“mg” [fixed] and oral: 93.1%, 40 mg (20-40) “mg/ml*” [conc.] and oral: 0.2%, 20 mg (12-30) “mg” [fixed] and inj.: 3.9%, 40 mg (20-40) “mg/ml” [conc.] and inj.: 2.8%, 80 mg (40-80) “mg/ml*” [conc.] and inj.: 0.0%, 20 mg (10-20)

* *Pattern with missing denominator*

Validation – Tiotropium (WHO DDD:
10 mcg powder inhalable / 5 mcg solution inhalable)

	Unit (%) DD (median, IQR)	Route and unit (%) DD (median, IQR)	Pattern and route (%) DD (median, IQR)
IQVIA Germany N = 1'016'219	mg : 87.0%, 0.018 (0.01 NA : 13.0%	inh. and [mg] : 87.0%, -0.054)	“mg” [fixed] and inh.: 58.4%, 0.036 (0.018-0.054) “mg/act” [conc.] and inh.: 20.7%, 0.0100 (0.005-0.015) “mg/ml” [conc.] and inh.: 7.8%, 0.000 (0.000-0.000) NA : 13.0%
IPCI (NL) N = 1'370'631	mg : 100.0% 0.018 (0.00 NA : 0.0%] : 100.0%, -0.018)	“mg” [fixed] and inh.: 60.7%, 0.018 (0.018-0.18) “mg/act” [conc.] and inh.: 39.3%, 0.005 (0.005-0.005) “mg/act*” [conc.] and inh.: 0.0%, 0.000 (0.000-0.003) NA : 0.0%
PharMetrics® Plus for Academics (US) N = 950'129	mg : 100%, 0.018 (0.01 0.020)] : 100%, -0.020)	“mg” [fixed] and inh.: 51.7%, 0.018 (0.018-0.18) “mg/act” [conc.] and inh.: 48.3%, 0.020 (0.020 - 0.020)

Not
applicable

* Pattern with missing denominator

Strength and Limitations



Demonstration of a uniform approach towards dose finding

Validation of dose formulas

Strength and Limitations



Demonstration of a uniform approach towards dose finding

Validation of dose formulas



This dose finding process is slow due to extensive clinical reviews.

Major obstacles is the “quantity” field which varies a lot depending on databases and makes it hard to suggest a uniform dose formula

Conclusion

Depending on the setting of the data (hospital, primary care, claims, electronic health record), the dosing estimation worked better or worse for different formulations and routes.

-> Thorough diagnostic investigations are needed before estimating dose in an individual data base.

Conclusion

The dose estimation is available in the DrugUtilisation R Package developed under DARWIN EU



THANK YOU!



Erasmus MC



Mees Mosseveld



CENTRE
HOSPITALIER
UNIVERSITAIRE
BORDEAUX

Romain Griffier



Christian Reich
Jasmine Gratton

Marti Catala Sabate
Edward Burn
Lucia Bellas
Kim Lopez Guell
Albert Prats Uribe
Annika Joedicke



Health Data Science Group,
University of Oxford, UK
Prof. Dani Prieto-Alhambra



ODYSSEUS
DATA SERVICES INC

Artem Gorbachev
Asieh Golozar



Poster 502: today 2:45 – 3:30pm, Sunday noon – 1 pm

Generating Synthetic Electronic Health Records in OMOP using GPT

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar
Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason
Patterson, Karthik Natarajan

Department of Biomedical Informatics
Columbia University



OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



Motivations for synthetic EHR data

Machine Learning

- Prediction research
- External validation

Phenotype algorithm validation

Tool development

Training and education

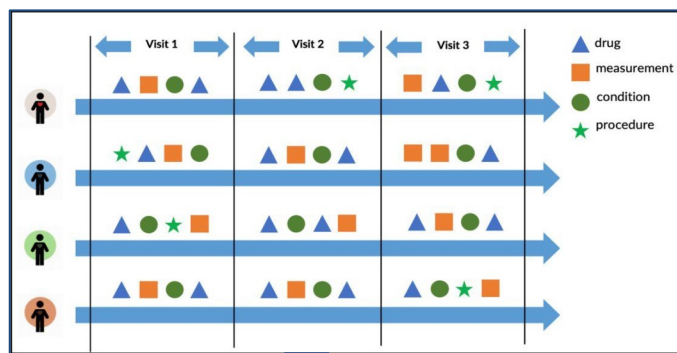
Fairness and Bias

- Debiasing the source data
- Counterfactual dataset



Common Approach: Bag of Word (BOW) + GAN

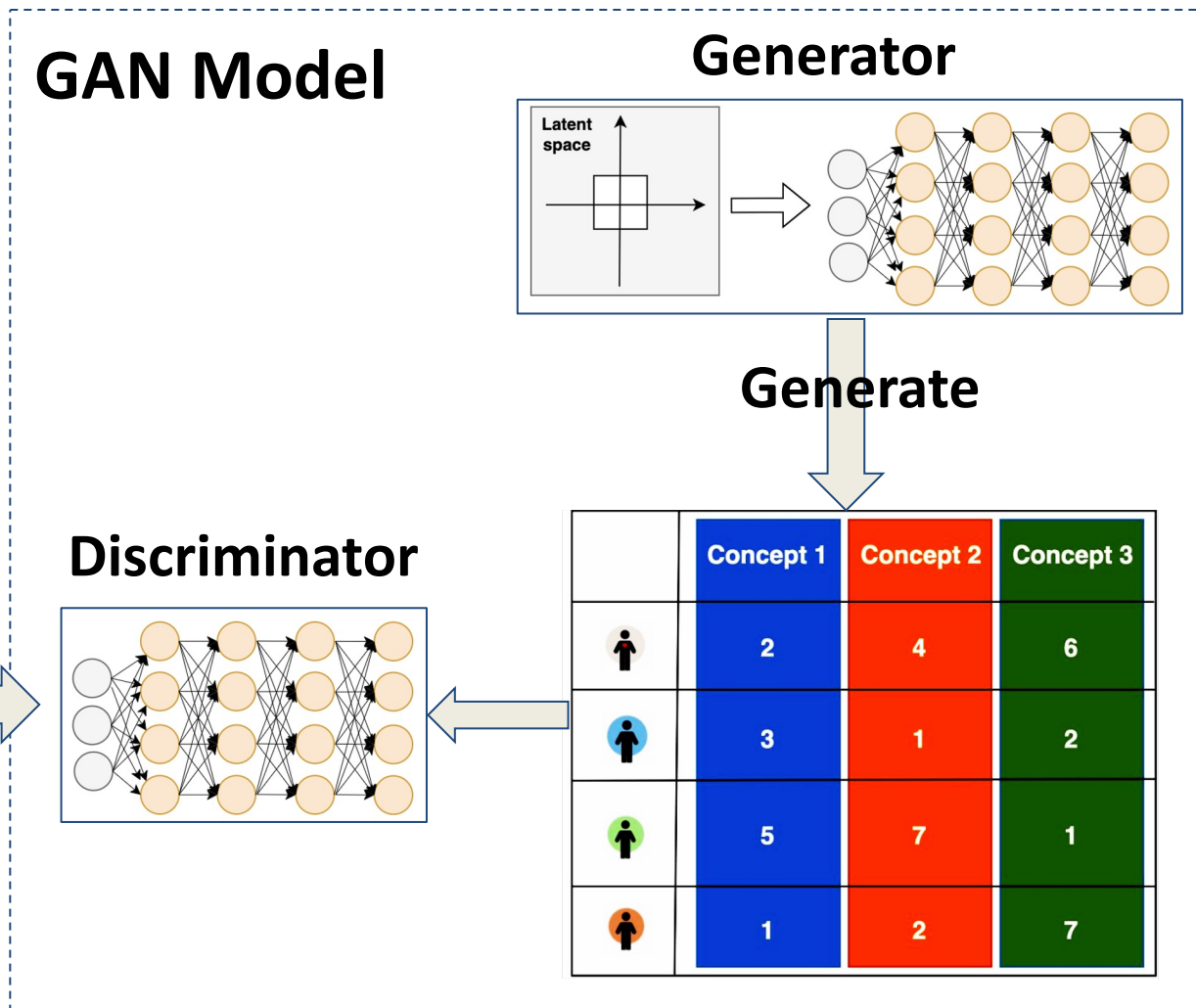
EHR Data



BOW Processing

	Concept 1	Concept 2	Concept 3
1	2	4	6
2	3	1	2
3	5	7	1
4	1	2	7

GAN Model



	Concept 1	Concept 2	Concept 3
1	2	4	6
2	3	1	2
3	5	7	1
4	1	2	7



JOURNAL ARTICLE

SynTEG: a framework for temporal structured electronic health data simulation FREE

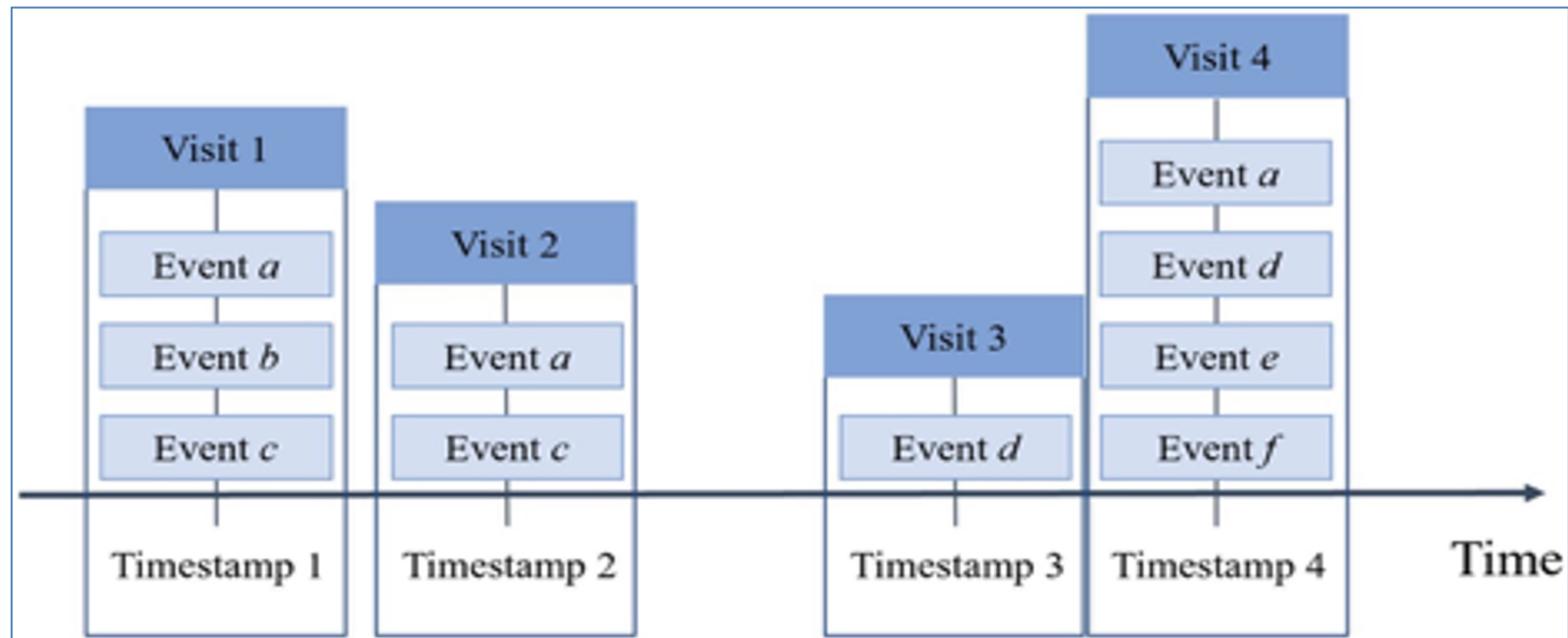
Ziqi Zhang, Chao Yan ✉, Thomas A Lasko, Jimeng Sun, Bradley A Malin

Journal of the American Medical Informatics Association, Volume 28, Issue 3, March 2021, Pages 596–604,

<https://doi.org/10.1093/jamia/ocaa262>

Published: 23 November 2020 **Article history** ▼

PDF Split View Cite Permissions Share ▼





JOURNAL ARTICLE

SynTEG: a framework for temporal structured electronic health

Ziqi Zhang, Chao Yan ✉, Thomas A Lasko, Jimeng Sun, Bradley A Malin

Journal of the American Medical Association

<https://doi.org/10.1093/jamia/ocaa262>

Published: 23 November 2020 Article history ▾

PDF Split View Cite Permissions Share ▾

- All visits assume to end on the same day as the visit start (Not true for inpatient visits)

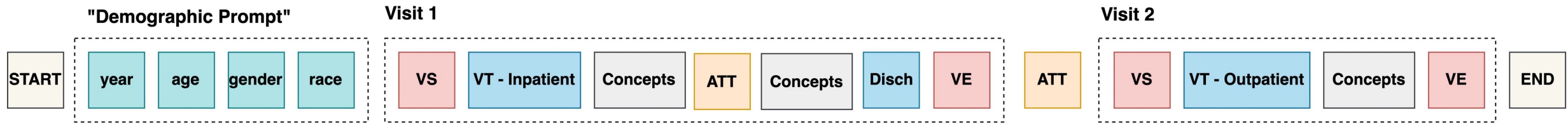
- Visit type is missing
- Discharge type is missing

- Not easily disseminated for use





Patient Representation



year Year at first visit

age Age at first visit

gender Gender

race Race

VS Visit Start

VE Visit End

VT Visit Type

Disch Discharge type

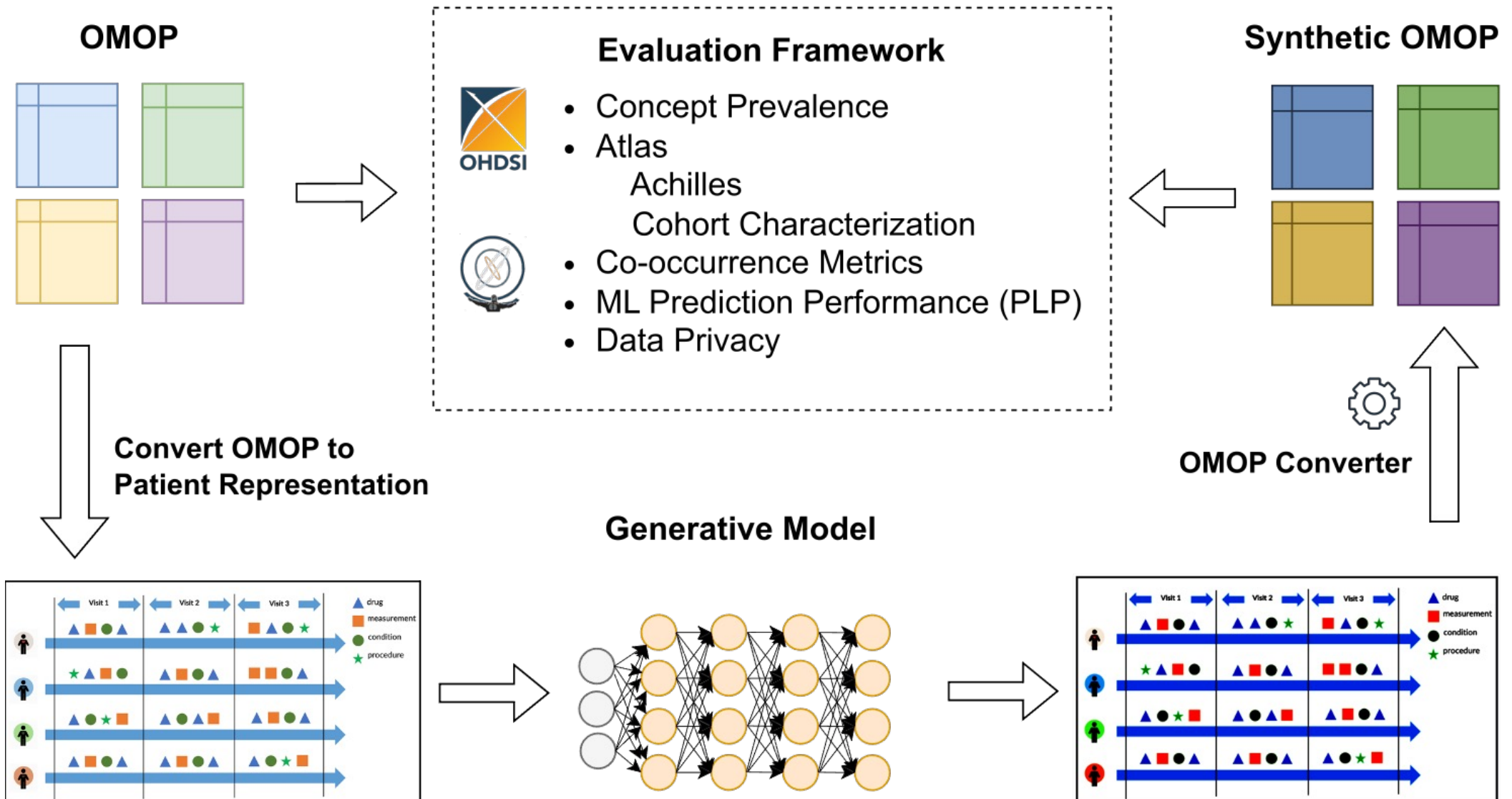
ATT Artificial Time Token
Day token

Concepts Condition, Drug
Ingredient, Procedure

CEHR-BERT <https://proceedings.mlr.press/v158/pang21a/pang21a.pdf>

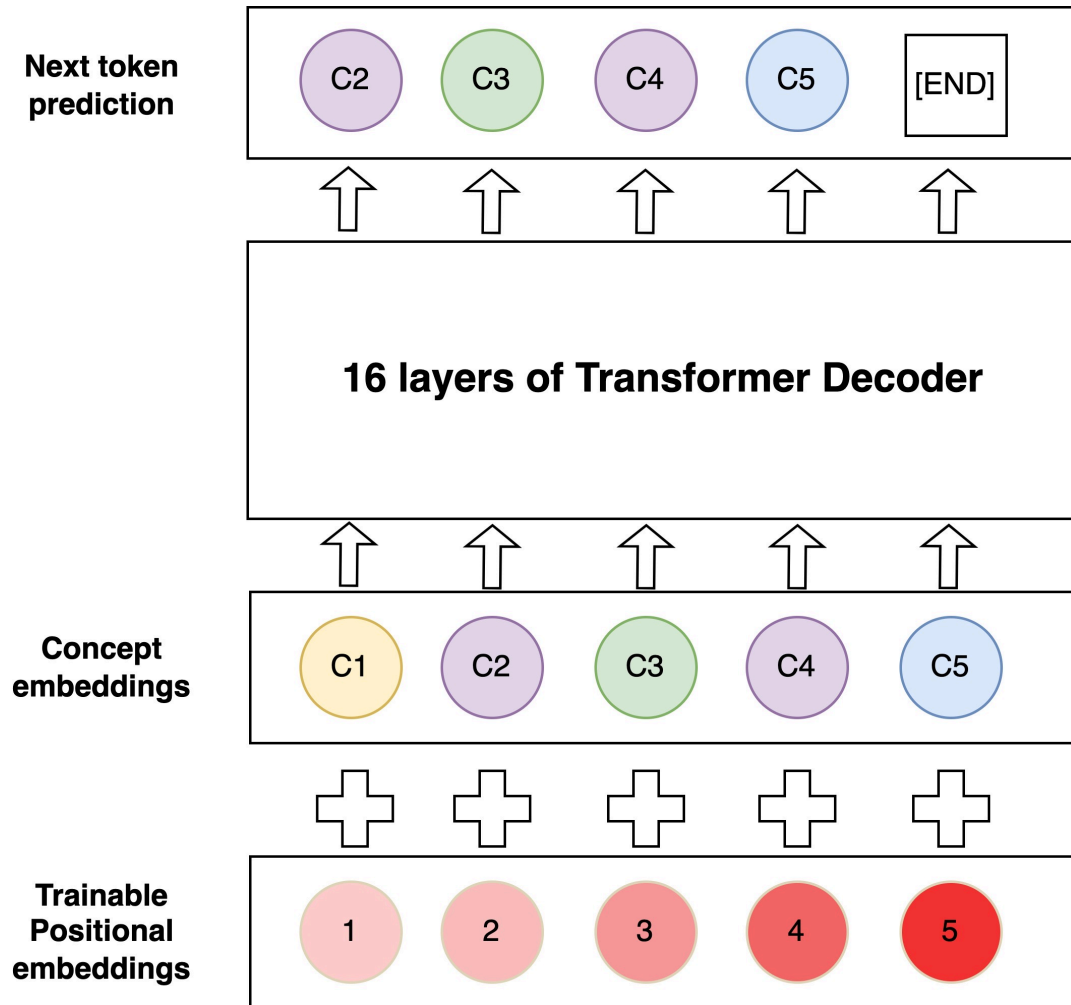


Proposed Synthetic Data Framework





Training a Generative Model



Data Preprocessing

- Condition, drug, procedure
- Context window 512
- Min number of concepts 20
- Truncate the long sequences
- 3 million patients after filtering

Training parameters

- Batch size 32
- Learning rate 1e-5
- Adam optimizer
- 2 epochs
- Save every 10000 steps

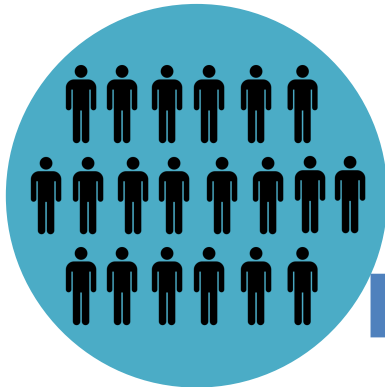


Generate new patient sequences

Inference model

- Top k=100, 200, 300
- Top p=95%, 100%
- Generated 500K for each sampling strategy

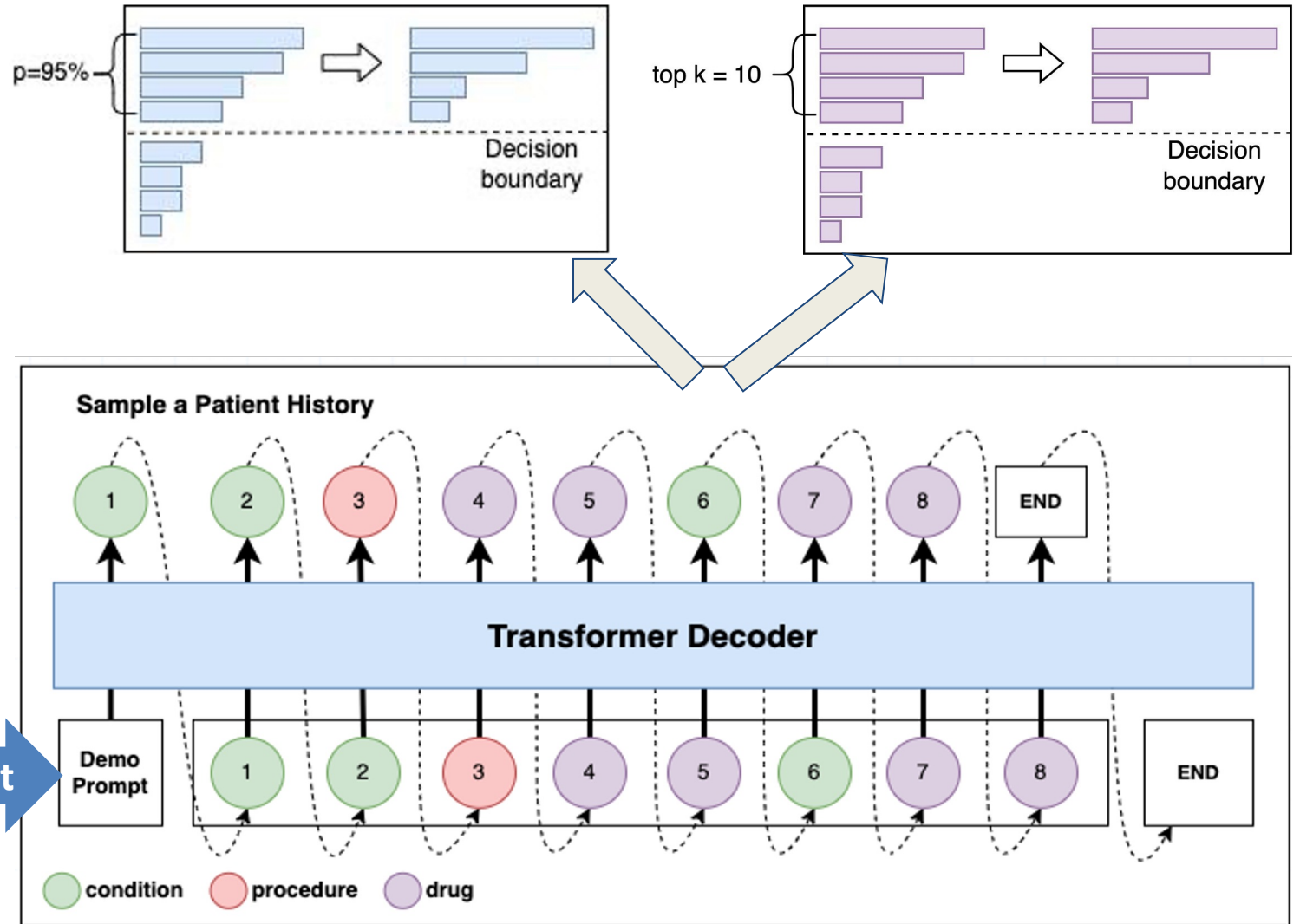
Patient Population



Sample



Prompt

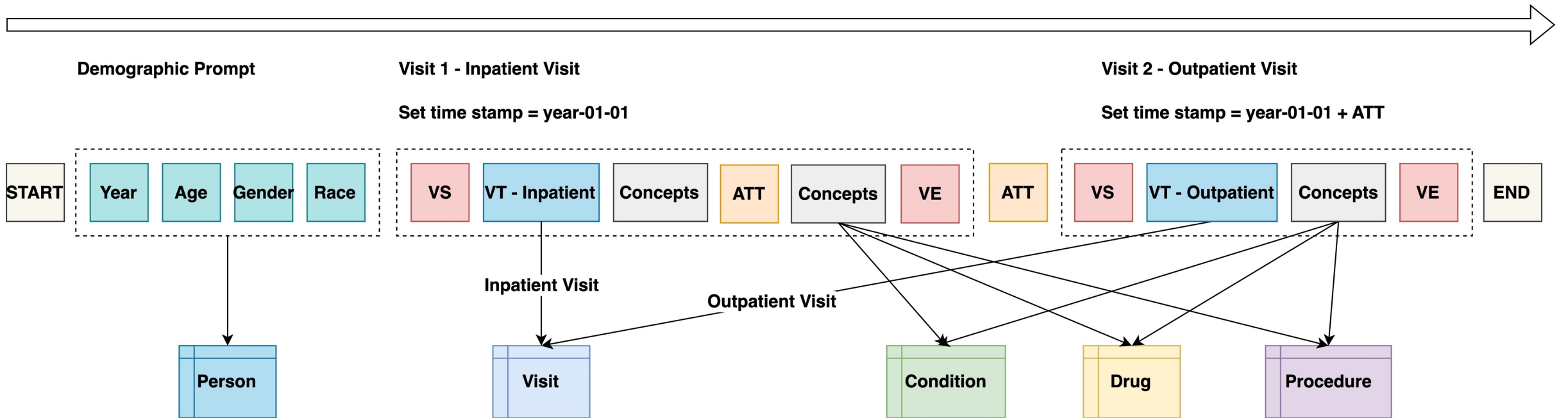
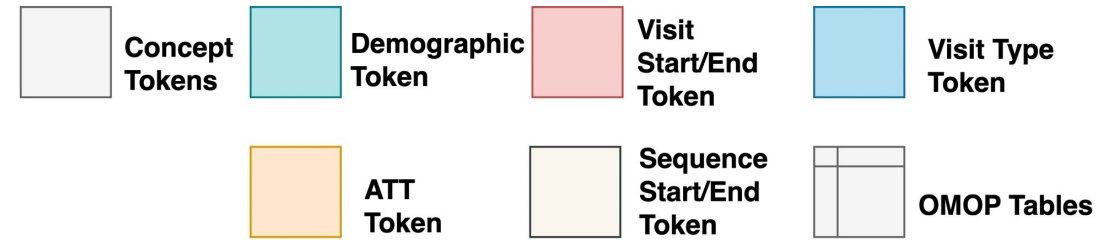




OMOP Converter

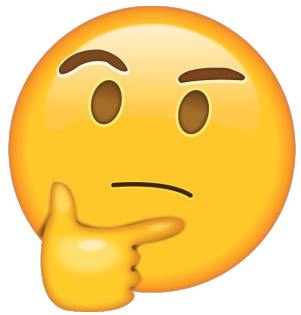
Patient Representation and OMOP Converter

Convert each patient sequence into a set of records in OMOP tables chronologically





How do you measure the similarity of two OMOP instances?



$$fx\left(\begin{array}{|c|c|} \hline \text{blue} & \text{green} \\ \hline \text{yellow} & \text{purple} \\ \hline \end{array}, \begin{array}{|c|c|} \hline \text{blue} & \text{green} \\ \hline \text{yellow} & \text{purple} \\ \hline \end{array}\right) = ?$$



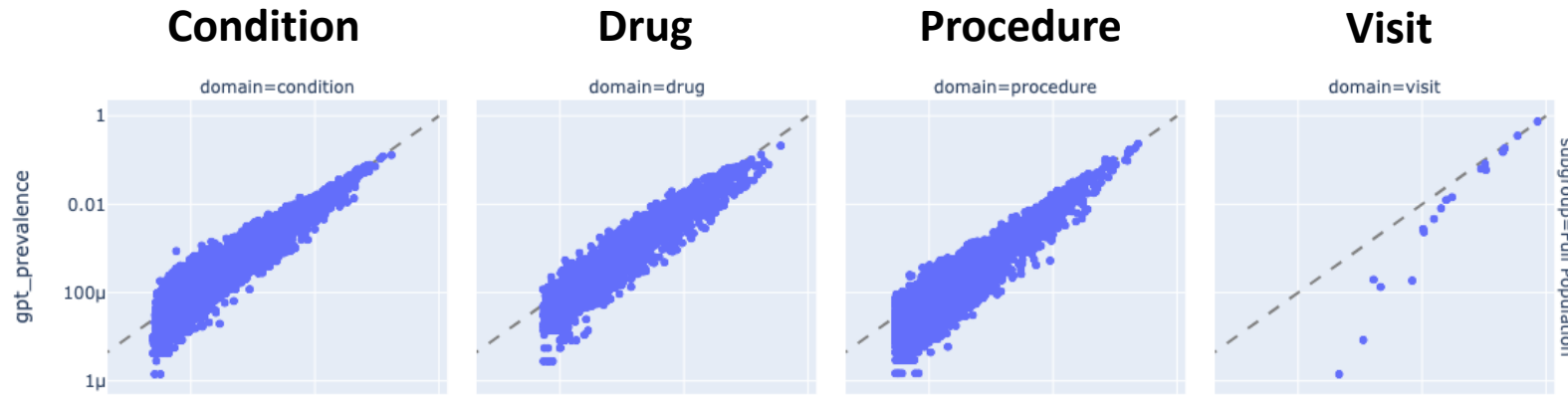
Evaluation framework

- **Level 1: Concept distributions at the full population, subgroups, cohorts. Marginal distribution e.g. $P(a; \text{group})$**
- **Level 2: Similarity of co-occurrence matrices at the full population. Conditional distribution e.g. $P(a | b)$**
- **Level 3: Logistic regression performance on synthetic cohorts. Proxy for joint distribution e.g. $P(a, b, c, d ; \text{group})$**

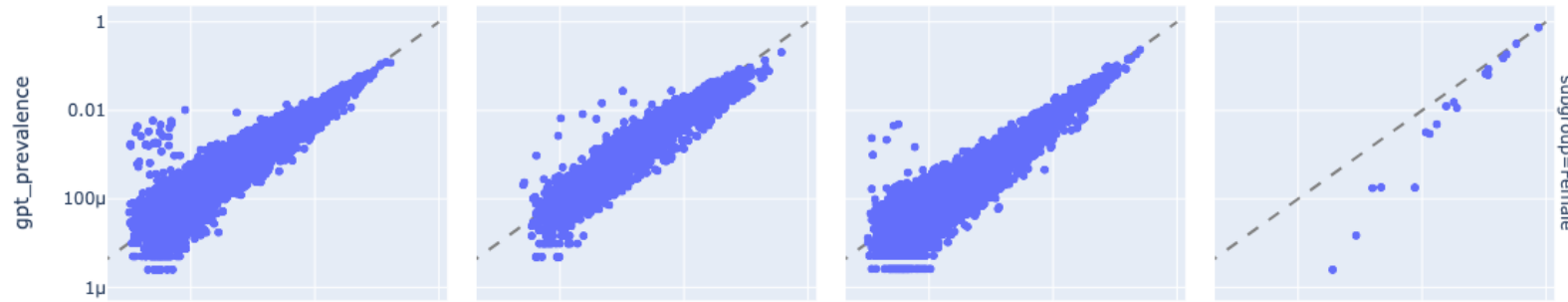


Level 1: Concept distributions

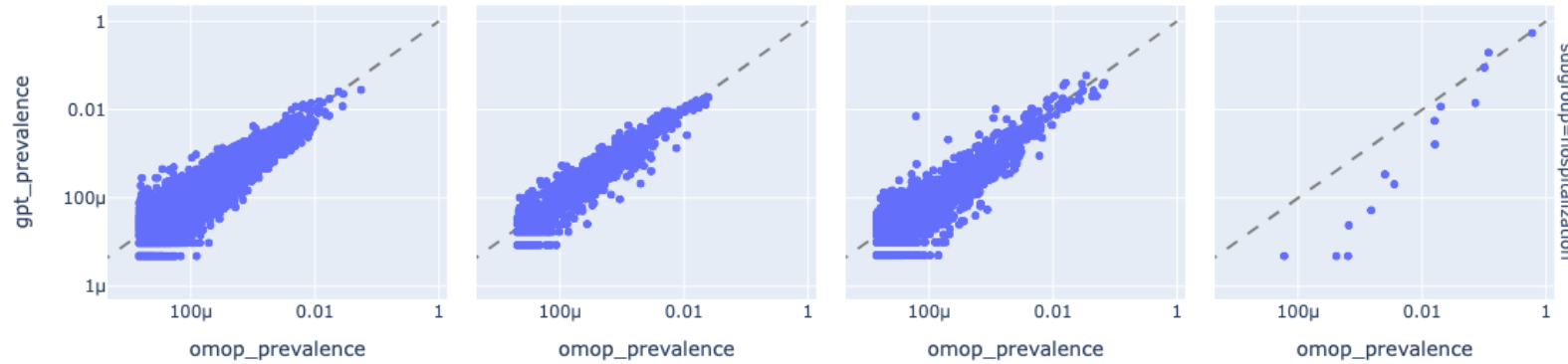
Full Population



Female Population



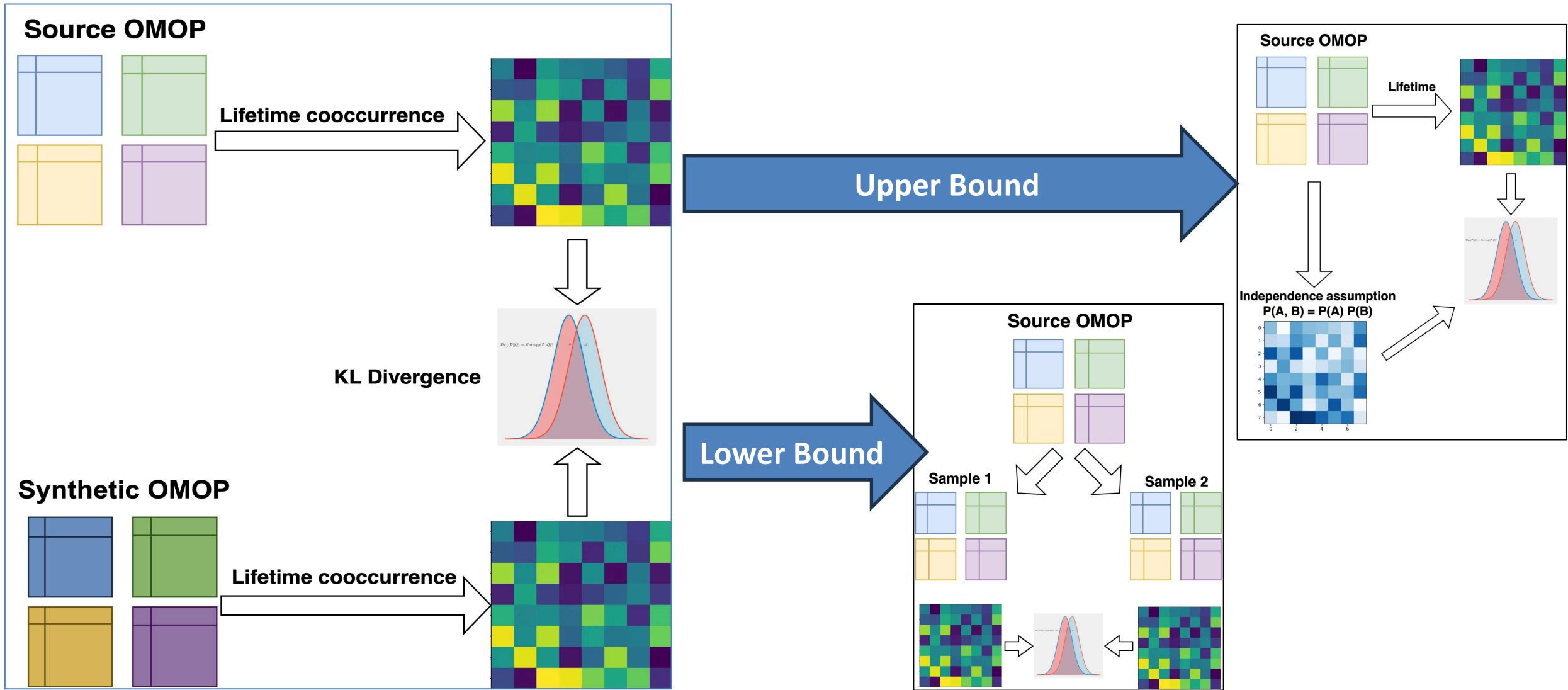
Hospitalization cohort



- Synthetic data: Top P=95%
- X: source data
- Y: synthetic data

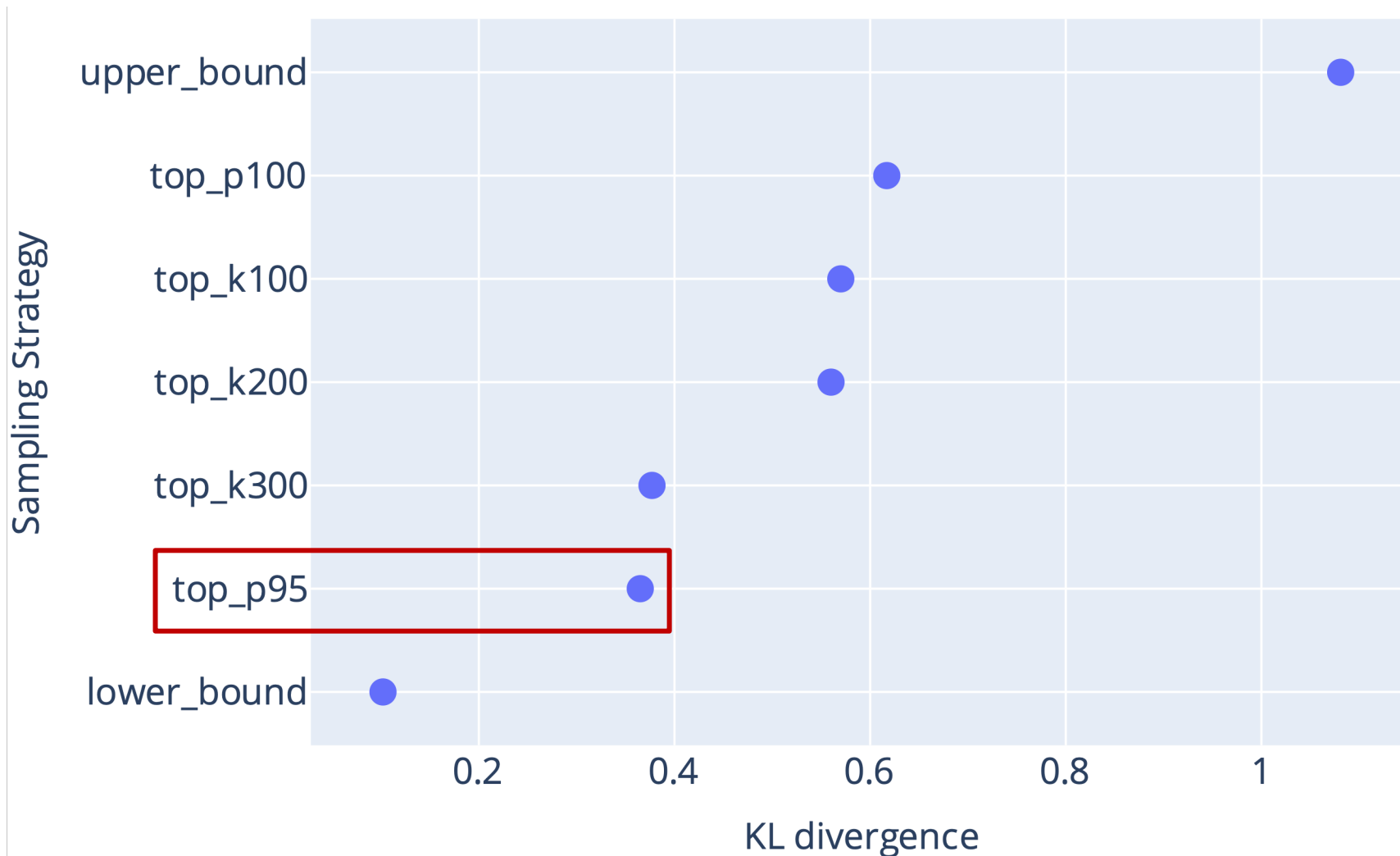


Level 2: Similarity of co-occurrence matrices





Level 2: Similarity of co-occurrence matrices



- Top k=100, 200, 300
- Top p=95%, 100%
- Sampling strategies affect results.
- Top p=95% has the best KL-divergence



Level 3: Logistic Regression model performance

	Cohort Definition used in CEHR-BERT
HF readmission	HF patients who have a 30-day all-cause readmission. Observation window: 360 days, Prediction windows 30 days
Hospitalization	2-year risk of hospitalization starting from the 3rd year since the initial entry into the EHR system Observation window: 540 days, hold-off window: 180 days, Prediction windows 720 days
COPD readmission	COPD patients who have a 30-day all-cause readmission. Observation window: 360 days, Prediction windows 30 days
Afib ischemic stroke	Afib patients with 1 year risk since the initial diagnosis of afib ischemic stroke Observation window: 720 days, Prediction windows 360 day
CAD CABG	Patients initially diagnosed with Coronary Arterial Disease (CAD) without any prior stent graft that will receive the Coronary artery bypass surgery (CABG) treatment Observation window: 720 days, Prediction windows 360 day

Level 3: Logistic Regression model performance

	Real data	Top P=95%	Top P=100%	Top K=100	Top K=200	TOP K=300
HF readmission	Pre = 25.7 AUC = 65.7 PR = 39.3	Pre = 27.6 AUC = 69.2 PR = 45.7	Pre = 28.4 AUC = 65.9 PR = 41.8	Pre = 30.7 AUC = 68.1 PR = 47.8	Pre = 29.3 AUC = 54.0 PR = 32.9	Pre = 26.5 AUC = 61.1 PR = 33.8
Hospitalization	Pre = 5.6 AUC = 75.3 PR = 19.5	Pre = 5.2 AUC = 77.1 PR = 21.4	Pre = 7.3 AUC = 68.3 PR = 16.5	Pre = 2.8 AUC = 87.0 PR = 22.1	Pre = 5.2 AUC = 84.2 PR = 20.8	Pre = 6.3 AUC = 78.7 PR = 24.6
COPD readmission	Pre = 34.5 AUC = 74.2 PR = 83.8	Pre = 37.8 AUC = 76.4 PR = 84.4	Pre = 47.2 AUC = 74.1 PR = 67.2	Pre = 26.4 AUC = 75.9 PR = 90.3	Pre = 28.3 AUC = 70.1 PR = 82.8	Pre = 34.5 AUC = 68.8 PR = 80.2
Afib ischemic stroke	Pre = 8.7 AUC = 84.0 PR = 48.5	Pre = 10.2 AUC = 78.9 PR = 41.2	Pre = 10.4 AUC = 70.7 PR = 39.1	Pre = 16.6 AUC = 77.1 PR = 50.5	Pre = 15.8 AUC = 68.9 PR = 36.6	Pre = 10.8 AUC = 76.8 PR = 38.5
CAD CABG	Pre = 7.1 AUC = 88.4 PR = 55.9	Pre = 4.1 AUC = 81.5 PR = 25.2	Pre = 4.4 AUC = 52.9 PR = 4.3	Pre = 7.2 AUC = 75.6 PR = 38.5	Pre = 4.9 AUC = 73.5 PR = 24.3	Pre = 4.0 AUC = 79.0 PR = 24.1



Conclusion

- First deep learning framework generated longitudinal synthetic EHR data using OMOP CDM.
- Designed an innovative patient representation, which allowed the reconstruction of patient medical timeline without loss of temporal information.
- Comprehensive evaluation procedures showed that the synthetic data preserved the underlying characteristics of the real patient population.



Acknowledgement

Team

Xinzhuo (Zoey) Jiang
Nishanth Parameshwar
Pavinkurve
Krishna S. Kalluri
Elise L. Minto
Jason Patterson
Karthik Natarajan

OHDSI (APOLLO)

Martijn Schuemie
Yong Chen
Egill Fridgeirsson
Chungsoo Kim
Jenna Reps
Marc Suchard
Xiaoyu Wang

Columbia DBMI

George Hripcsak
Lingying Zhang
Harry Reyes
Tara Anand
Maura Beaton
Nripendra Acharya

Grants

This project is partially supported by
5U01TR002062 and 5U2COD023196

COMPARING EXTRACTED CONCEPTS FROM TEXT TO STRUCTURED CONDITIONS

Tom Seinen

PhD Student – Erasmus MC



This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.



Erasmus MC
University Medical Center Rotterdam



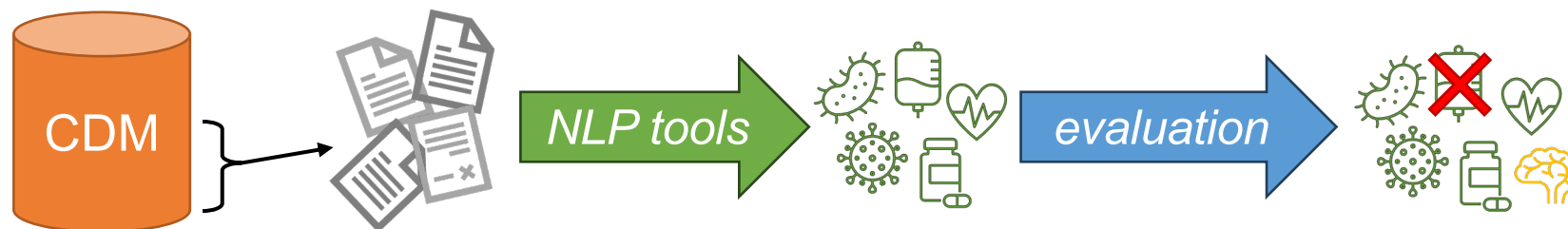
Context & Problem

Dutch general practitioner database:



- 2.5 million patients
- 8% population of the Netherlands

Unstructured data: free text

- CDM **notes** table
- **35%** physical space of the **database**
- Potential **information** currently **unused**



Extracting clinical concepts

- **Many** tools for **English** 
- **Not** for **Dutch** 
- So.. we **created a framework** for **Dutch**

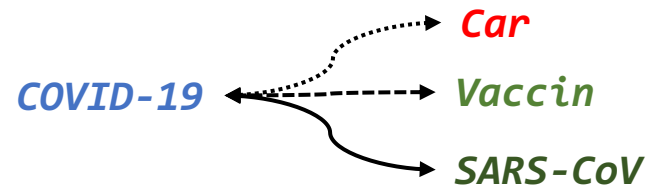
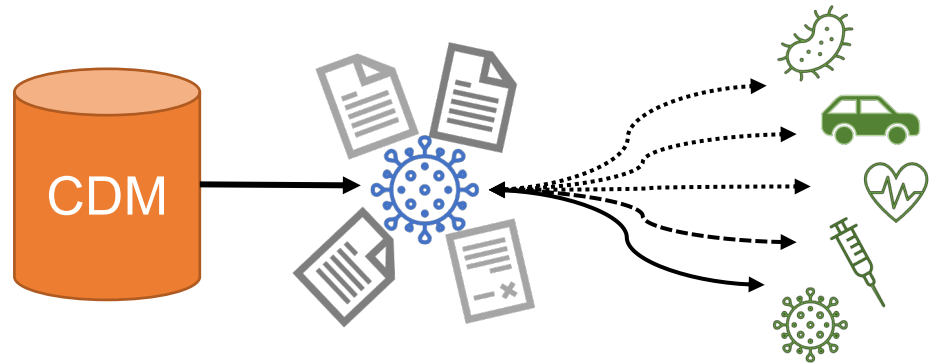
Concept extraction evaluation

- Requires an **annotated dataset** (ground truth)
- **None** exists for **Dutch**

Research objective

Possible solution:

- Notes do not exist by **themselves**
- They often occur **together** with a **condition code**



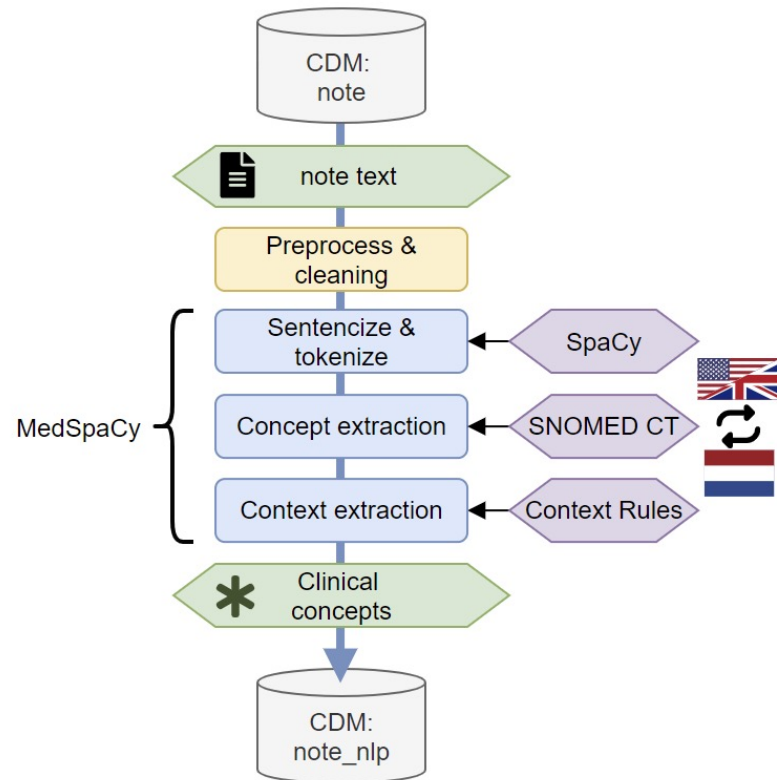
Can we use the **structured codes** for **evaluation**?:

- Surrogate annotations
- Compare the **extracted codes** with the **structured code**
- Can we find **similar** or **related concepts** in the **text**?

We find similar concepts, then the extraction works!

Methods

Dutch concept extraction framework:



Experimental setup:

- Most frequent conditions in the database
- Take all notes within a 3-day window
- Extract clinical concepts from these notes

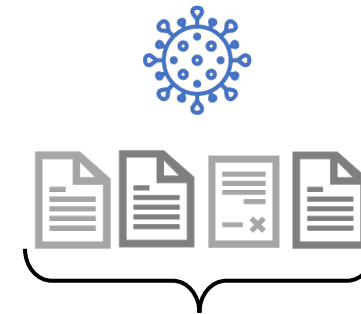
Methods

However, the **assumption** that the **text is related to the coded condition** might **not** always hold.

Ground truth is still needed

We **annotated** a set of 2000 code observations

- 200 different codes
- **Slow**: annotate every clinical concept in the text.
- **Fast**: does the **text** describe:
 - A **similar** concept or
 - A **related** concept to the **recorded condition**?
 - Two **yes/no** questions



Annotate:

Similar to condition?
Related to condition?

Methods

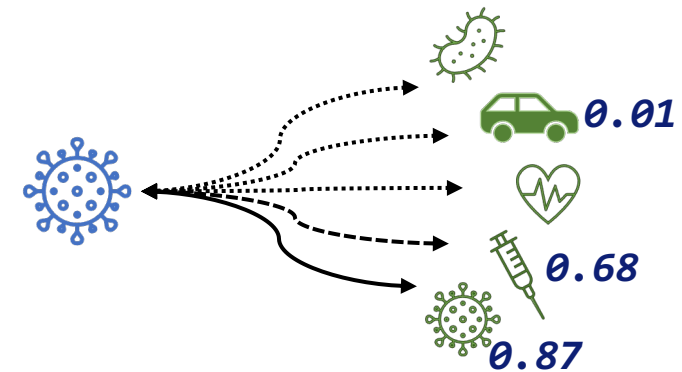
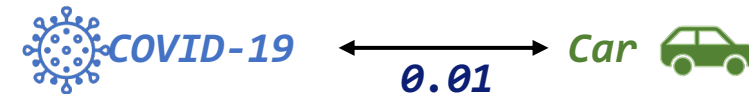
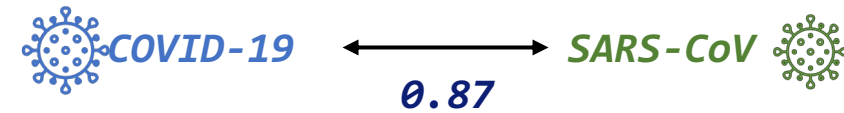
Concept similarity

- Pretrained Concept **embeddings** (SNOMED CT)
 - Numerical representations of the concept
 - Generated using a neural network
- Cosine **distance** between **embeddings** = **semantic similarity**

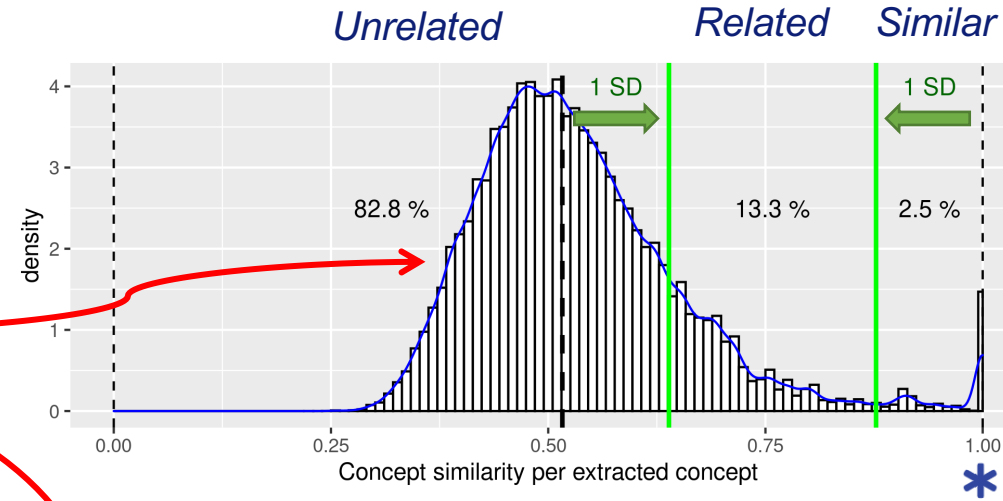
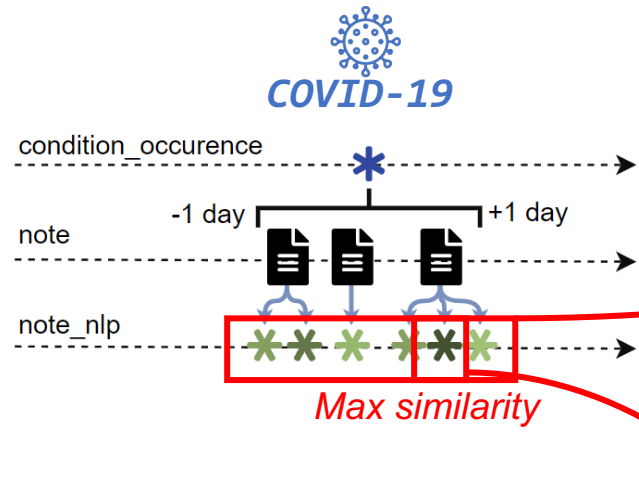
Is the condition mentioned in the text?

- Find the **most similar** concept
 - Concept with **maximum** similarity
- When is the concept the same? Or related?
 - Set **thresholds** on similarity...

SNOMED CT
The global language of healthcare



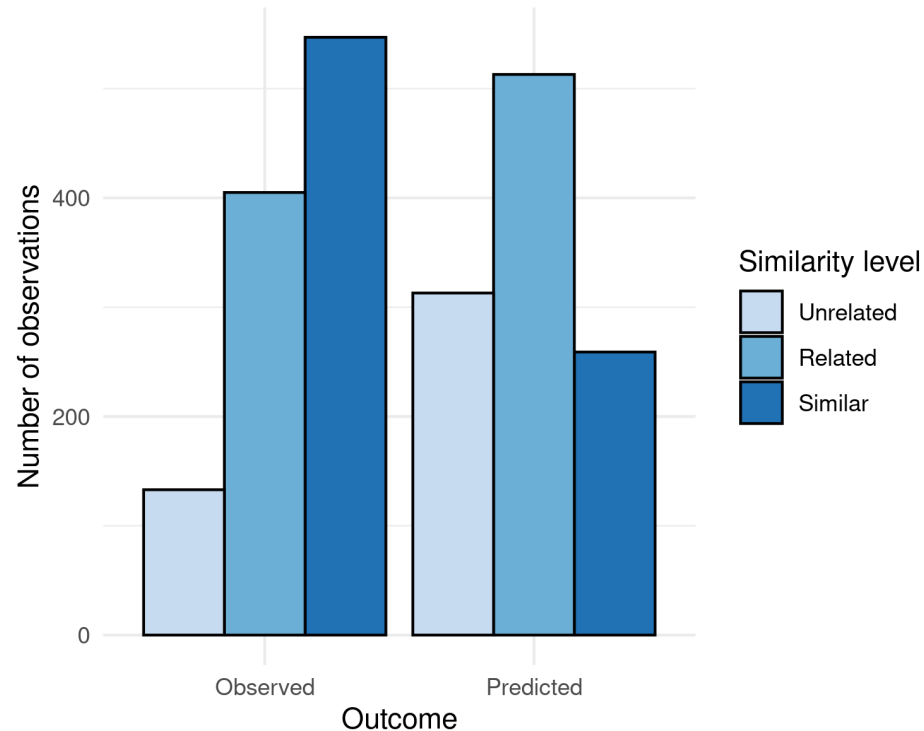
Results



For **29 million** condition occurrences:

- in **27%** we find a **similar** concept
- in **47%** we find a **related** concept
- in **27%** we find **only unrelated concepts**

Results – evaluate on annotated set



	F1	Recall	Prec.	Acc.
Similar	.61	.47	.99	.73
Related	.76	.63	.94	.70
Similar or related	.88	.80	.98	.81

In 2000 occurrences:

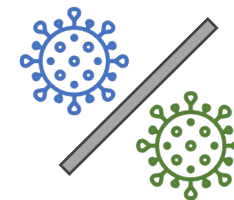
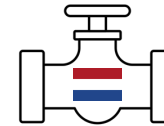
- Found **less similar** concepts than expected
- **More related** concepts than expected
- Slightly **less similar or related** than expected
- If no **similar** concept was found, then usually a **related** concept was identified

Conclusion

1. We created a **non-English concept extraction** framework using public resources
2. We **evaluated** the **framework** using the **structure data** as **surrogate labels**
 - **Limitation**: Only **tests** whether we can **extract** the **information** that is **expected**
 - **Language agnostic**
3. Our **framework** performs **relatively well**, but it can be **improved**
 - **Limitation**: Currently uses only **SNOMED synonyms**
4. **Most conditions** have **related** or **similar** concepts in the **surrounding text**

More info?

Meet me at my **poster: 504**





Finding a constrained number of predictor phenotypes for multiple outcome prediction

Jenna M Reps, Jenna Wong, Egill A. Fridgeirsson, Chungsoo Kim, Luis H. John, Ross D. Williams, Patrick Ryan



A Team Effort Made This Possible





Motivation

Aim: Can we find a constrained set of predictors that can be used for many health outcome prediction tasks and lead to good performance?

← → ↻ mdcalc.com/calc/801/cha2ds2-vasc-score-atrial-fibrillation-stroke-risk

☰ MD+ CALC 🔍 Search "QT interval" or "QT" or "EKG"

CHA₂DS₂-VASc Score for Atrial Fibrillation Stroke Risk ☆

Calculates stroke risk for patients with atrial fibrillation, possibly better than the [CHADS₂ Score](#).

When to Use ▾ Pearls/Pitfalls ▾ Why Use ▾

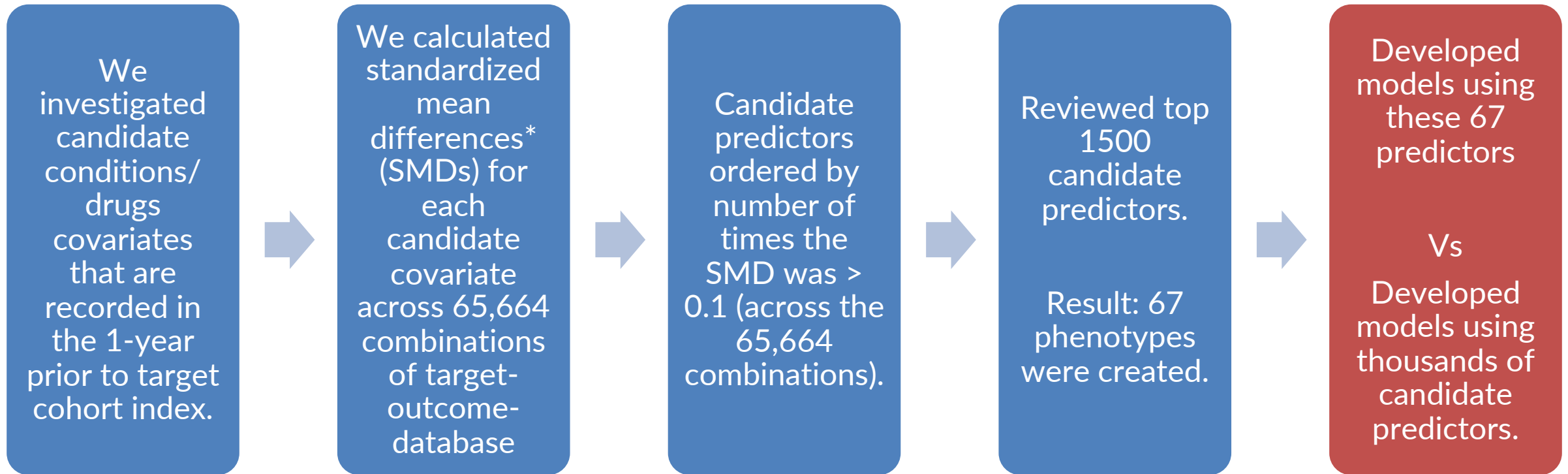
Age	<65 0	65-74 +1	≥75 +2
Sex	Female +1	Male 0	
CHF history	No 0	Yes +1	
Hypertension history	No 0	Yes +1	
Stroke/TIA/thromboembolism history	No 0	Yes +2	
Vascular disease history (prior MI, peripheral	No 0	Yes +1	

Ideal output: a website with one form and thousands of models



Methodology

We developed a process to learn conditions/drugs that are generally predictive across many target cohorts and outcomes...



*SMD compares baseline prevalence of the candidate covariate between cases and non-cases

Results: Our constrained predictor set

Predictor	Disorder classification
Alcoholism	Behavioral
Smoking	Behavioral
Anemia	Blood
Osteoarthritis	Bone
Osteoporosis	Bone
Cancer	Cancer
Atrial fibrillation	Cardiovascular
Congestive heart failure	Cardiovascular
Coronary artery disease	Cardiovascular
Heart valve disorder	Cardiovascular
Hyperlipidemia	Cardiovascular
Hypertension	Cardiovascular
Angina	Cardiovascular
Skin ulcer	Debility
Diabetes type 1	Endocrine
Diabetes type 2	Endocrine
Hypothyroidism	Endocrine
Obesity	Endocrine
Gastroesophageal reflux disease (GERD)	GI
Gastrointestinal (GI) bleed	GI
Inflammatory bowel disorder	GI/Rheumatology

Predictor	Disorder classification
Hormonal contraceptives	Gynecologic
Antibiotic use (separated by family)	Infection
Pneumonia	Infection/Respiratory
Sepsis	Infection
Urinary tract infection (UTI)	Infection
Hepatitis	Liver
Anxiety	Mood
Depression	Mood
Psychotic disorder	Mood
Antiepileptics (pain)	Neurology/Pain
Seizure	Neurology
Hemorrhagic stroke	Neurology/Vascular
Non-hemorrhagic stroke	Neurology/Vascular
Acetaminophen prescription	Pain/Infection
Low back pain	Pain
Neuropathy	Pain/Neurology
Opioids	Pain
Acute kidney injury	Kidney
Chronic kidney disease	Kidney

Predictor	Disorder classification
Asthma	Respiratory
Chronic obstructive pulmonary disorder (COPD)	Respiratory
Dyspnea	Respiratory
Respiratory failure	Respiratory
Sleep apnea	Respiratory
Rheumatoid arthritis	Rheumatology
Steroids	Rheumatology/Pain/Pulmonary
Peripheral vascular disease	Vascular
Aspirin	Vascular
Deep vein thrombosis (DVT)	Vascular
Edema	Vascular
Inpatient visit	Inpatient Visit

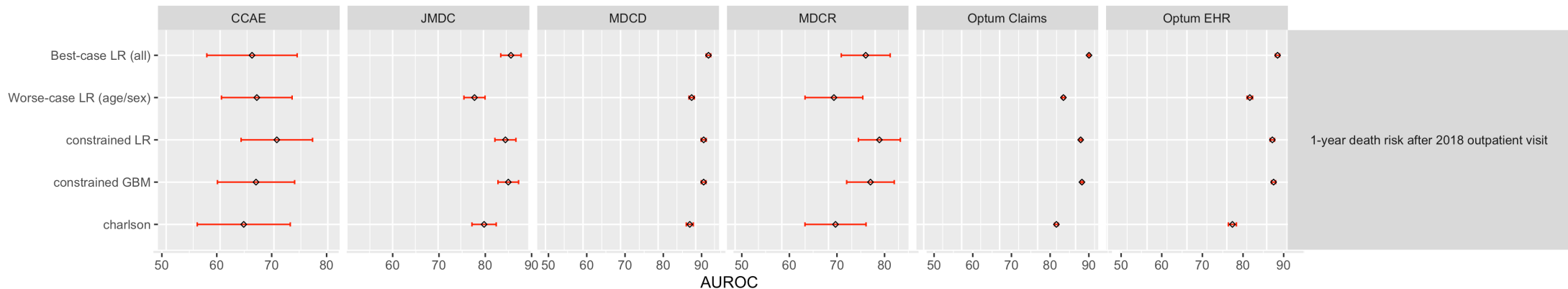
These phenotypes are available in the OHDSI phenotype library

Results: evaluation of our constrained predictor set

For many prediction tasks we developed four models:

- Logistic regression using >10,000 SNOMED/RxNorm codes plus age/sex (best-case LR)
- Logistic regression using only age/sex predictors (worse-case LR)
- Logistic regression using our 67 predictors plus age/sex (constrained LR)
- Gradient Boosting Machine using our 67 predictors plus age/sex (constrained GBM)

Results for the task of predicting 1-year death after an outpatient visit in 2018



*Charlson – an existing model for this prediction task



What are your risks?

The constrained predictors led to good models.

Try it out yourselves:

www.WhatIHappenToMe.org

Predicted Risks	
<u>Outcome</u>	<u>Risk</u> ↓
<input type="text"/>	
Coronary artery disease (CAD)	5.42%
arrhythmia, condition, procedure, devise or drug	5.17%
Type 2 Diabetes Mellitus (DM), with no type 1 or secondary DM	2.73%
Heart failure	2.41%
Major depressive disorder, with NO occurrence of certain psychiatric disorder	2.15%
Muscle weakness or injury	2.1%
Ulcerative colitis	2.07%
Atrial Fibrillation	1.95%
Crohns disease	1.84%
Urinary tract infections (UTI)	1.64%