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
Prologue
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
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

Mapping Origin and Intent Metadata
- mapping_source: Tufts-CTSI:CCFS
- mapping_set_id: 40100000001
- mapping_set_label: Critical Care Flowseets Mapping to OMOP

Mapping Creation and Maintenance Metadata
- mapping_provider: Tufts-CTSI
- author_id: ORCID
- reviewer_id: ORCID
- mapping_tool: OHDSI Usagi
- mapping_date: 2023-01-15
- mapping_tool_version: v.1.4.3

Mapping Definition Metadata
- confidence: 1
- mapping_cardinality: 1:1
- mapping_justification: semapv:LexicalMatching

Subject
- subject_id: CCFS:700000046
- subject_category: PICU Vital Signs
- subject_type: Flowsheet Item
- subject_source_version: v.1.0
- subject_label: CPP (mmHg)

Predicate
- predicate_id: skos:exactMatch
- predicate_label: measurement

Object
- object_id: OMOP:21490695
- object_category: LOINC 2.73
Solution [2]: Use MAPPING_METADATA table
Solution [3]: Automation

- SSSOM MAPPING TABLE
- STAGING TABLES
- MAPPING_METADATA TABLE
- BASIC VOCABULARY TABLES
- EVENT TABLES
Needs [1]: Integration with OHDSI tools
Needs [2]: Community Contribution
Visit our poster #501!
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
### ATC/DDD Index

<table>
<thead>
<tr>
<th>ATC code</th>
<th>Name</th>
<th>Daily Dose</th>
<th>Unit</th>
<th>Administration Route</th>
</tr>
</thead>
<tbody>
<tr>
<td>N02BE01</td>
<td>paracetamol</td>
<td>3</td>
<td>g</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>g</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>g</td>
<td>R</td>
</tr>
</tbody>
</table>
Objectives

Our aims were

• to introduce a uniform approach to develop dose formulas

• to validate suggested dose formulas
Objectives

Our aims were

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

<table>
<thead>
<tr>
<th>Concept name of drug concept id</th>
<th>Amount</th>
<th>Numerator</th>
<th>Concept name of Numerator unit</th>
<th>Denominator</th>
<th>Concept name of Denominator unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 ML ibuprofen 10 MG/ML Injection [Neoprofen]</td>
<td>NA</td>
<td>20</td>
<td>milligram</td>
<td>2</td>
<td>milliliter</td>
</tr>
<tr>
<td>itraconazole 10 MG/ML Oral Solution [Sporanox]</td>
<td>NA</td>
<td>10</td>
<td>milligram</td>
<td>NA</td>
<td>milliliter</td>
</tr>
</tbody>
</table>

Patterns

22 patterns with clinically relevant units
Drug strength patterns

Daily dose formulas (to be calculated per pattern):
Numerator value * quantity \{drug exposure table\}
duration \{drug exposure table\}

<table>
<thead>
<tr>
<th>pattern name</th>
<th>Oral route</th>
<th>Injectable route</th>
<th>Inhalable route</th>
</tr>
</thead>
<tbody>
<tr>
<td>milliequivalent per milliliter</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>milliequivalent per milliliter * missing denominator</td>
<td>YES</td>
<td>NO</td>
<td>NA</td>
</tr>
<tr>
<td>milligram per actuation</td>
<td>NA</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>millgram per actuation * missing denominator</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>milligram per milligram</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>milligram per milligram * missing denominator</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>milliliter per milliliter</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>milliliter per milliliter * missing denominator</td>
<td>YES</td>
<td>YES</td>
<td>NA</td>
</tr>
<tr>
<td>milliliter per milliliter * missing denominator</td>
<td>YES</td>
<td>NO</td>
<td>NA</td>
</tr>
</tbody>
</table>

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

clinical review in CPRD AURUM/GOLD (UK), IPCI (NL), PharMetrics® Plus for Academics (US)
We estimated doses from 5 different ingredients in 5 different databases and compared them with the WHO Daily Dose.

<table>
<thead>
<tr>
<th>Concept Name</th>
<th>WHO DDD</th>
<th>Unit</th>
<th>Administration Route</th>
</tr>
</thead>
<tbody>
<tr>
<td>furosemide</td>
<td>40</td>
<td>milligram</td>
<td>oral injectable</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>milligram</td>
<td></td>
</tr>
<tr>
<td>tiotropium</td>
<td>10</td>
<td>microgram</td>
<td>inhalable (powder) inhalable (solution)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>microgram</td>
<td></td>
</tr>
<tr>
<td>metformin</td>
<td>2</td>
<td>gram</td>
<td>oral</td>
</tr>
<tr>
<td>enoxaparin</td>
<td>2</td>
<td>1000 IU</td>
<td>injectable</td>
</tr>
<tr>
<td>salmeterol</td>
<td>0.1</td>
<td>milligram</td>
<td>inhalable</td>
</tr>
</tbody>
</table>

WHO: World Health Organisation
DDD: Dispensed Daily Dose
IU: International unit
### Validation – Furosemide (WHO DDD: 40 mg oral / injectable)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Unit (%)</th>
<th>DD (median, IQR)</th>
<th>DD (median, IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IQVIA Germany</strong></td>
<td>[mg]: 93.3%, 40 mg (40-40) NA : 6.7%</td>
<td>oral and [mg]: 92.6%, 40 mg (40-40) inj. and [mg]: 0.6%, 40 mg (39-40) NA: 6.7%</td>
<td>“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%</td>
</tr>
<tr>
<td>N = 1‘375’495</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IPCI (NL)</strong></td>
<td>[mg]: 99.8%, 40 mg (20-40) NA : 0.2%</td>
<td>oral and [mg]: 99.7%, 40 mg (20-40) inj. and [mg]: 0.2%, 1 mg (2-20) NA : 0.2%</td>
<td>“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%</td>
</tr>
<tr>
<td>N = 2‘694’879</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PharMetrics® Plus for Academics (US)</strong></td>
<td>[mg] : 100%, 40 mg (20-40)</td>
<td>oral and [mg]: 93.3%, 40 mg (20-40) inj. and [mg]: 6.7%, 40 mg (20-80)</td>
<td>“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)</td>
</tr>
<tr>
<td>N = 4‘561’608</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* **Pattern with missing denominator**
## Validation – Furosemide (WHO DDD: 40 mg oral / injectable)

<table>
<thead>
<tr>
<th>Source</th>
<th>Unit (%)</th>
<th>Route and unit (%)</th>
<th>DD (median, IQR)</th>
<th>Pattern and route (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IQVIA</strong></td>
<td>[mg]: 93.3%, 40 mg (40-40) NA: 6.7%</td>
<td>oral and [mg]: 92.6%, 40 mg (40-40) 40 mg (40-40) inj. and [mg]: 0.6%, 40 mg (39-40) NA: 6.7%</td>
<td>“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%</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 1’375’495</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IPCI (NL)</strong></td>
<td>[mg]: 99.8%, 40 mg (20-40) NA: 0.2%</td>
<td>oral and [mg]: 99.7%, 40 mg (20-40) 40 mg (20-40) inj. and [mg]: 0.2%, 1 mg (2-20) NA: 0.2%</td>
<td>“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 inj.: 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%</td>
<td></td>
</tr>
<tr>
<td>N = 2’694’879</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PharMetrics® Plus for Academics (US)</strong></td>
<td>[mg]: 100%, 40 mg (20-40)</td>
<td>oral and [mg]: 93.3%, 40 mg (20-40) 40 mg (20-40) inj. and [mg]: 6.7%, 40 mg (20-80)</td>
<td>“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.: 0.2%, 80 mg (40-80) “mg/ml” [conc.] and inj.: 0.0%, 20 mg (10-20)</td>
<td></td>
</tr>
<tr>
<td>N = 4’561’608</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Pattern with missing denominator
<table>
<thead>
<tr>
<th>Source</th>
<th>Unit (%, DD (median, IQR))</th>
<th>Route and unit (%) DD (median, IQR)</th>
<th>Pattern and route (%) DD (median, IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IQVIA Germany</strong>&lt;br&gt;N = 1′375′495</td>
<td>[mg]: 93.3%, 40 mg (40-40)&lt;br&gt;NA : 6.7%</td>
<td>oral and [mg]: 92.6%, 40 mg (40-40)&lt;br&gt;inj. and [mg]: 0.6%, 40 mg (39-40)&lt;br&gt;NA : 6.7%</td>
<td>“mg” [fixed] and oral: 92.3%, 40 mg (40-40)&lt;br&gt;“mg/ml” [conc.] and oral: 0.3%, 10 mg (10-10)&lt;br&gt;“mg/ml” [conc.] and inj.: 0.6%, 40 mg (39-40)&lt;br&gt;NA : 6.7%</td>
</tr>
<tr>
<td><strong>IPCI (NL)</strong>&lt;br&gt;N = 2′694′879</td>
<td>[mg]: 99.8%, 40 mg (20-40)&lt;br&gt;NA : 0.2%</td>
<td>oral and [mg]: 99.7%, 40 mg (20-40)&lt;br&gt;inj. and [mg]: 0.2%, 1 mg (2-20)&lt;br&gt;NA : 0.2%</td>
<td>“mg” [fixed] and oral: 99.6%, 40 mg (20-40)&lt;br&gt;“mg/ml” [conc.] and oral: 0.1%, 0 mg (0-0)&lt;br&gt;“mg/ml” [conc.] and oral: 0.0%, 20 mg (10-20)&lt;br&gt;“mg/ml” [conc.] and inj.: 0.1%, 1 mg (2-20)&lt;br&gt;“mg/ml” [conc.] and inj.: 0.0%, 0 mg (0-0)&lt;br&gt;NA : 0.2%</td>
</tr>
<tr>
<td><strong>PharMetrics® Plus for Academics (US)</strong>&lt;br&gt;N = 4′561′608</td>
<td>[mg] : 100%, 40 mg (20-40)</td>
<td>oral and [mg]: 93.3%, 40 mg (20-40)&lt;br&gt;inj. and [mg]: 6.7%, 40 mg (20-80)</td>
<td>“mg” [fixed] and oral: 93.1%, 40 mg (20-40)&lt;br&gt;“mg/ml” [conc.] and oral: 0.2%, 20 mg (12-30)&lt;br&gt;“mg” [fixed] and inj.: 3.9%, 40 mg (20-40)&lt;br&gt;“mg/ml” [conc.] and inj.: 2.8%, 80 mg (40-80)&lt;br&gt;“mg/ml” [conc.] and inj.: 0.0%, 20 mg (10-20)</td>
</tr>
</tbody>
</table>

* Pattern with missing denominator
Validation – Tiotropium (WHO DDD: 10 mcg powder inhalable / 5 mcg solution inhalable)

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

* 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
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!

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

Mees Mosseveld
Romain Griffier
Christian Reich
Jasmine Gratton

Artem Gorbachev
Asieh Golozar

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

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
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**

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

**GAN Model**

**Generator**

**Discriminator**

**Generate**

- Latent space
- Concept 1
- Concept 2
- Concept 3
- Values:
  - Concept 1: 2, 3, 5, 1
  - Concept 2: 4, 1, 7, 2
  - Concept 3: 6, 2, 1, 7

www.ohdsi.org

#JoinTheJourney
SynTEG: a framework for temporal structured electronic health data simulation

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


Published: 23 November 2020   Article history ▼
• 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

"Demographic Prompt"

START

year age gender race

Visit 1

VS VT - Inpatient Concepts ATT Concepts Disch VE

Visit 2

ATT VS VT - Outpatient Concepts VE

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


@OHDSI
www.ohdsi.org
#JoinTheJourney

ohdsi
Proposed Synthetic Data Framework

OMOP

Evaluation Framework
- Concept Prevalence
- Atlas
  - Achilles
  - Cohort Characterization
- Co-occurrence Metrics
- ML Prediction Performance (PLP)
- Data Privacy

Synthetic OMOP

Convert OMOP to Patient Representation

Generative Model

OMOP Converter
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

Demographic Prompt

Visit 1 - Inpatient Visit
Set time stamp = year-01-01

Visit 2 - Outpatient Visit
Set time stamp = year-01-01 + ATT

START Year Age Gender Race VS VT - Inpatient Concepts ATT Concepts VE ATT VS VT - Outpatient Concepts VE END

Inpatient Visit

Outpatient Visit

Person Visit Condition Drug Procedure

OMOP Tables
How do you measure the similarity of two OMOP instances?
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

- **Condition**
- **Drug**
- **Procedure**
- **Visit**

**Full Population**

- **Female Population**

**Hospitalization cohort**

- Synthetic data: Top P=95%
- X: source data
- Y: synthetic data
Level 2: Similarity of co-occurrence matrices

Source OMOP

Lifetime cooccurrence

KL Divergence

Synthetic OMOP

Lifetime cooccurrence

Upper Bound

Source OMOP

Lifetime

Independence assumption

P(A, B) = P(A) P(B)

Lower Bound

Sample 1

Sample 2
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

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

<table>
<thead>
<tr>
<th></th>
<th>Real data</th>
<th>Top P=95%</th>
<th>Top P=100%</th>
<th>Top K=100</th>
<th>Top K=200</th>
<th>TOP K=300</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HF readmission</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>25.7</td>
<td>27.6</td>
<td>28.4</td>
<td>30.7</td>
<td>29.3</td>
<td>26.5</td>
</tr>
<tr>
<td>AUC</td>
<td>65.7</td>
<td>69.2</td>
<td>65.9</td>
<td>68.1</td>
<td>54.0</td>
<td>61.1</td>
</tr>
<tr>
<td>PR</td>
<td>39.3</td>
<td>45.7</td>
<td>41.8</td>
<td>47.8</td>
<td>32.9</td>
<td>33.8</td>
</tr>
<tr>
<td><strong>Hospitalization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>5.6</td>
<td>5.2</td>
<td>7.3</td>
<td>2.8</td>
<td>5.2</td>
<td>6.3</td>
</tr>
<tr>
<td>AUC</td>
<td>75.3</td>
<td>77.1</td>
<td>68.3</td>
<td>87.0</td>
<td>84.2</td>
<td>78.7</td>
</tr>
<tr>
<td>PR</td>
<td>19.5</td>
<td>21.4</td>
<td>16.5</td>
<td>22.1</td>
<td>20.8</td>
<td>24.6</td>
</tr>
<tr>
<td><strong>COPD readmission</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>34.5</td>
<td>37.8</td>
<td>47.2</td>
<td>26.4</td>
<td>28.3</td>
<td>34.5</td>
</tr>
<tr>
<td>AUC</td>
<td>74.2</td>
<td>76.4</td>
<td>74.1</td>
<td>75.9</td>
<td>70.1</td>
<td>68.8</td>
</tr>
<tr>
<td>PR</td>
<td>83.8</td>
<td>84.4</td>
<td>67.2</td>
<td>90.3</td>
<td>82.8</td>
<td>80.2</td>
</tr>
<tr>
<td><strong>Afib ischemic stroke</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>8.7</td>
<td>10.2</td>
<td>10.4</td>
<td>16.6</td>
<td>15.8</td>
<td>10.8</td>
</tr>
<tr>
<td>AUC</td>
<td>84.0</td>
<td>78.9</td>
<td>70.7</td>
<td>77.1</td>
<td>68.9</td>
<td>76.8</td>
</tr>
<tr>
<td>PR</td>
<td>48.5</td>
<td>41.2</td>
<td>39.1</td>
<td>50.5</td>
<td>36.6</td>
<td>38.5</td>
</tr>
<tr>
<td><strong>CAD CABG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>7.1</td>
<td>4.1</td>
<td>4.4</td>
<td>7.2</td>
<td>4.9</td>
<td>4.0</td>
</tr>
<tr>
<td>AUC</td>
<td>88.4</td>
<td>81.5</td>
<td>52.9</td>
<td>75.6</td>
<td>73.5</td>
<td>79.0</td>
</tr>
<tr>
<td>PR</td>
<td>55.9</td>
<td>25.2</td>
<td>4.3</td>
<td>38.5</td>
<td>24.3</td>
<td>24.1</td>
</tr>
</tbody>
</table>
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 5U2C0D023196
COMPARING extrAacted concepts FROM tEXT TO sTRUCTURED conditions

Tom Seinen
PhD Student – Erasmus MC
**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
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…

Example:

- COVID-19 to SARS-CoV: $\theta = 0.87$
- COVID-19 to Car: $\theta = 0.01$
- COVID-19 to other conditions: $\theta = 0.68$
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**

<table>
<thead>
<tr>
<th>Similarity level</th>
<th>F1</th>
<th>Recall</th>
<th>Prec.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar</td>
<td>.61</td>
<td>.47</td>
<td>.99</td>
<td>.73</td>
</tr>
<tr>
<td>Related</td>
<td>.76</td>
<td>.63</td>
<td>.94</td>
<td>.70</td>
</tr>
<tr>
<td>Similar or related</td>
<td>.88</td>
<td>.80</td>
<td>.98</td>
<td>.81</td>
</tr>
</tbody>
</table>

In 2000 occurrences:
- Found **less similar** concepts than expected
- **More related** concepts than expected
- Slightly **less similar or related** then 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?

**Ideal output:** a website with one form and thousands of models
We investigated candidate conditions/drugs covariates that are recorded in the 1-year prior to target cohort index.

We calculated standardized mean differences* (SMDs) for each candidate covariate across 65,664 combinations of target-outcome-database.

Candidate predictors ordered by number of times the SMD was > 0.1 (across the 65,664 combinations).

Reviewed top 1500 candidate predictors.

Result: 67 phenotypes were created.

Developed models using these 67 predictors.

Vs Developed models using thousands of candidate predictors.

* SMD compares baseline prevalence of the candidate covariate between cases and non-cases.
Results: Our constrained predictor set

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Disorder classification</th>
<th>Predictor</th>
<th>Disorder classification</th>
<th>Predictor</th>
<th>Disorder classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcoholism</td>
<td>Behavioral</td>
<td>Hormonal contraceptives</td>
<td>Gynecologic</td>
<td>Asthma</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Smoking</td>
<td>Behavioral</td>
<td>Antibiotic use (separated by family)</td>
<td>Infection</td>
<td>Chronic obstructive pulmonary disorder (COPD)</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Anemia</td>
<td>Blood</td>
<td>Pneumonia</td>
<td>Infection/Respiratory</td>
<td>Dyspnea</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>Bone</td>
<td>Sepsis</td>
<td>Infection</td>
<td>Respiratory failure</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>Bone</td>
<td>Urinary tract infection (UTI)</td>
<td>Infection</td>
<td>Sleep apnea</td>
<td>Respiratory</td>
</tr>
<tr>
<td>Cancer</td>
<td>Cancer</td>
<td>Hepatitis</td>
<td>Liver</td>
<td>Rheumatoid arthritis</td>
<td>Rheumatology</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>Cardiovascular</td>
<td>Anxiety</td>
<td>Mood</td>
<td>Steroids</td>
<td>Rheumatology/Pulmonary</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>Cardiovascular</td>
<td>Depression</td>
<td>Mood</td>
<td>Peripheral vascular disease</td>
<td>Vascular</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>Cardiovascular</td>
<td>Psychotic disorder</td>
<td>Mood</td>
<td>Aspirin</td>
<td>Vascular</td>
</tr>
<tr>
<td>Heart valve disorder</td>
<td>Cardiovascular</td>
<td>Antiepileptics (pain)</td>
<td>Neurology/Pain</td>
<td>Deep vein thrombosis (DVT)</td>
<td>Vascular</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>Cardiovascular</td>
<td>Seizure</td>
<td>Neurology</td>
<td>Edema</td>
<td>Vascular</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Cardiovascular</td>
<td>Hemorrhagic stroke</td>
<td>Neurology/Vascular</td>
<td>Inpatient visit</td>
<td>Inpatient Visit</td>
</tr>
<tr>
<td>Angina</td>
<td>Cardiovascular</td>
<td>Non-hemorrhagic stroke</td>
<td>Neurology/Vascular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin ulcer</td>
<td>Debility</td>
<td>Acetaminophen prescription</td>
<td>Pain/Infection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes type 1</td>
<td>Endocrine</td>
<td>Low back pain</td>
<td>Pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes type 2</td>
<td>Endocrine</td>
<td>Neuropathy</td>
<td>Pain/Neurology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothyroid</td>
<td>Endocrine</td>
<td>Opioids</td>
<td>Pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>Endocrine</td>
<td>Acute kidney injury</td>
<td>Kidney</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastroesophageal reflux disease (GERD)</td>
<td>GI</td>
<td>Chronic kidney disease</td>
<td>Kidney</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastrointestinal (GI) bleed</td>
<td>GI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflammatory bowel disorder</td>
<td>GI/Rheumatology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.WhatllHappenToMe.org