Agenda

1. Welcome and Sentinel overview
2. FDA opening remarks
3. DI2: Representation of unstructured data across Common Data Models
4. DI3: Identification and mitigation of structured EHR source data mapping issues
5. FE1: Computable phenotyping framework
6. FE2: NLP tools for cohort identification, exposure assessment, covariate ascertainment
7. FE3: Improving probabilistic phenotyping of incident outcomes
8. CI1: Enhancing Causal Inference in the Sentinel System
9. CI2: A causal inference framework for Sentinel
10. Closing remarks
Overview
FDA’s Sentinel system

2007 FDA Amendments Act mandates FDA to establish **active surveillance system** for monitoring drugs using electronic healthcare data.

Through the Sentinel Initiative, FDA aims to assess the post-marketing safety of approved medical products.
Inability to identify certain study populations of interest from insurance claims

Inability to identify certain outcomes of interest from insurance claims

Other limitations (inadequate duration of follow-up, the need for additional signal identification tools)

---

### Current Sentinel system limitations

- **Inability to identify certain study populations of interest from insurance claims**
- **Inability to identify certain outcomes of interest from insurance claims**
- **Other limitations** (inadequate duration of follow-up, the need for additional signal identification tools)

### Sentinel Innovation Center Initiatives

#### Data infrastructure (DI)

- **10+ million people**
- **EHR + Claims**

#### Feature engineering (FE)

- Emerging methods including machine learning and scalable automated natural language processing (NLP) approaches to enable computable phenotyping from unstructured EHR data

#### Causal inference (CI)

- Methodologic research to address specific challenges when using EHRs such as approaches to handle missing data, calibration methods for enhanced confounding adjustment

#### Detection analytics (DA)

- Development of signal detection approaches to account for and leverage differences in data content and structure of EHRs

---

### Sentinel Innovation Center vision

A query-ready, quality-checked distributed data network containing EHR for at least 10 million lives with reusable analysis tools

---

**2020**

---

**2024**
<table>
<thead>
<tr>
<th>Priorities</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data infrastructure</td>
<td>Identification and queries of potential EHR data partners (Horizon Scan: DI1)</td>
<td>Adding unstructured data and necessary data elements (DI2)</td>
<td>Source data mapping (DI3)</td>
<td>Data quality metrics and quality assurance strategy</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Uploading CDM to include EHR data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Harmonizing EHRs (DI4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Data harmonization strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FHIR preparedness (DI7)</td>
</tr>
<tr>
<td>Feature engineering</td>
<td></td>
<td>Computed phenotyping framework (FE1)</td>
<td>Increasing automation in computed phenotyping</td>
<td>Enhancing transportability of phenotypes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NLP tools for cohort identification, exposure assessment, covariate</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ascertaining (Scalable NLP: FE2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Improving probabilistic phenotyping of incident outcomes (FE3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Developing NLP-assisted chart abstraction tool (FE4)</td>
<td>Implementing NLP-assisted chart abstraction tool</td>
<td></td>
</tr>
<tr>
<td>Causal inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Causal inference framework (CI2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Innovation incubator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Priorities

<table>
<thead>
<tr>
<th>Data infrastructure</th>
<th>Feature engineering</th>
<th>Causal inference</th>
<th>Detection analytics</th>
<th>Innovation incubator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>Year 2</td>
<td>Year 3</td>
<td>Year 4</td>
<td>Year 5</td>
</tr>
<tr>
<td>Master plan</td>
<td>Evaluation</td>
<td>Causal</td>
<td>Detection</td>
<td>Data Sandbox</td>
</tr>
<tr>
<td></td>
<td>and queries</td>
<td>inference</td>
<td>analytics</td>
<td>Discovery</td>
</tr>
<tr>
<td></td>
<td>of potential EHR</td>
<td>framework</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>data partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Horizon Scan: DI1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Onboarding EHR data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Onboarding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EHR data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adding unstructured</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and necessary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data elements (DI2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source data mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DI3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotyping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>framework (FE1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLP tools for</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cohort identification,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exposure assessment,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ascertainment (Scalable NLP: FE2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improving</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>probabilistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotyping of</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incident outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(FE3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLP-assisted chart</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abstraction tool (FE4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating targeted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learning in EHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data (Enhancing CI:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Causal inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>framework (CI2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CI4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approaches for</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing data (CI3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distributed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implementation (CI6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identification and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation of</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EHR detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>approaches (DA1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of EHR-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>detection approaches</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development of EHR-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>based detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developing and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>advancing EHR-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>detection methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods for signal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>detection for</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pregnancy/birth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>outcomes (DA4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pregnancy and birth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>outcomes signal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>detection tool</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods for cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>signal detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tool development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV signal detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DA6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance metrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CI5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approaches for</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing data (CI3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distributed regression implementation (CI6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DI2:** Representation of unstructured data across Common Data Models

**DI3:** Identification and mitigation of structured EHR source data mapping issues
Challenges and Opportunities in Integrating Electronic Health Record (EHR) Data in Sentinel

Keith Marsolo, PhD
Associate Professor
Department of Population Health Sciences
Duke Clinical Research Institute
Duke University School of Medicine
Purpose
IC Projects -- Highlight Challenges and Opportunities

As the Sentinel Innovation Center works to establish an infrastructure of administrative claims linked with electronic health record (EHR) data on 10 million+ lives:

- Focus = two projects that develop aspects of the infrastructure needed to bring EHR data into the Sentinel framework

Each highlights potential challenges and opportunities presented by EHR

**DI2: Representation of unstructured data across Common Data Models**

**DI3: Identification and mitigation of structured EHR source data mapping issues**
DI2: Representation of Unstructured Data Across Common Data Models
Incorporating Unstructured Data into a Common Data Model

Goal: To guide the Sentinel Network on **how best to incorporate information derived from unstructured data into a Common Data Model (CDM) framework.**

Objectives:

1) **What information is important?** – Identify the priority elements that should be derived from unstructured data

2) **What NLP tools are in use & how are they used?; What information is available within a note?** – Assess the overall availability of the priority elements within the Sentinel ecosystem

3) **How to best represent information derived from unstructured text?** – Recommend how those priority elements should be represented in the Sentinel Common Data Model

Project completion date: **May 31, 2022** (to be extended)
Project team

HPHCI
- Judy Maro, Co-Investigator
- Kathleen Shattuck, Project Manager
- Tyler Erikson, Statistician
- Ziyang He, Student
- Jaiden Dumas, Clinical Research Coordinator

Duke
- Keith Marsolo, PI
- Lesley Curtis, Co-Investigator
- Chuan Hong, Co-Investigator
- Sarah Palmer, Project Leader/Analyst

Vanderbilt
- Ruth Reeves, Co-Investigator
- Dax Westerman, Principal Application Developer
- Juan Zhao, Co-Investigator
- Jill Whitaker, Data Analyst
- Tina French, Data Analyst
- Liz Hanchrow, Senior Application Analyst

BWH
- Li Zhou, Co-Investigator
- Suzanne Blackley, Applications Analyst
- John Laurentiev, Applications Analyst

FDA
- Sarah Dutcher
- Efe Eworuke
- Aida Kuzucan

Subject Matter Experts
- Joseph Plasek (BWH); others TBD
Objective 1 – What information is important?

Process:
Generated list of concepts from commonly-used NLP pipelines (commercial & open-source)

- Focused mainly on broad categories, not specific items, unless called out in documentation (e.g., medications, not aspirin)
- Looked at the basic functionality provided by each tool, not every research project
- Generated “good enough” list – stopped when we reached saturation

FDA reviewed list, identified any missing elements & assigned priority rankings (high / medium / low) - highest priority given to those concepts not easily obtained from claims that are also important for drug safety studies

End Product:
Set of priority elements to be derived from unstructured text.

Image source: https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-for-health
## Example priority rankings (subset)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Concept(s)</th>
<th>Priority</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>Site</td>
<td>High</td>
<td>Several ARIA insufficiency rankings due to lack of data on cancer (e.g., staging)</td>
</tr>
<tr>
<td></td>
<td>Histology</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>Diagnoses</td>
<td>Medium</td>
<td>Often captured in claims</td>
</tr>
<tr>
<td></td>
<td>Signs / Symptoms</td>
<td>High</td>
<td>Less available in claims, useful in different aspects of studies</td>
</tr>
<tr>
<td></td>
<td>Family History (Type)</td>
<td>Medium</td>
<td>Useful in some studies, but not all</td>
</tr>
<tr>
<td></td>
<td>Medical History (Type)</td>
<td>High</td>
<td>Often gaps in EHR data, medical history important to capture</td>
</tr>
<tr>
<td>Medication</td>
<td>Class</td>
<td>Low</td>
<td>Can be inferred from drug name</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept(s)</th>
<th>Priority</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timing &amp; duration of medication</td>
<td>High</td>
<td>Particularly important for inpatient medications</td>
</tr>
<tr>
<td>Physical findings (e.g., vital signs)</td>
<td>High</td>
<td>Key covariate for FDA studies, under-captured in claims</td>
</tr>
<tr>
<td>Indication for a drug</td>
<td>High</td>
<td>Rationale for why a drug is given</td>
</tr>
<tr>
<td>Oxygen support</td>
<td>High</td>
<td>Relevant for many COVID-19 studies</td>
</tr>
<tr>
<td>Death (date) &amp; cause</td>
<td>Low*</td>
<td>Capture of death data varies by Sentinel Data Partner</td>
</tr>
</tbody>
</table>
Objective 2 – What NLP tools are in use and how are they used? What information is available within a note?

Process:
Distributed survey to partners within the Sentinel ecosystem to assess their NLP capabilities (e.g., tool(s) used, notes processed, concepts extracted, etc.) – understand how well the current state of NLP use aligns with the priority concepts identified by FDA

Perform chart annotations at 2 sites (Vanderbilt, Brigham & Women’s Hospital) to assess availability of priority elements within 2 different use cases (in progress)

End Product:
Survey responses from Partners on their ability to extract priority data elements from unstructured text, and statistics on the overall availability of priority data elements within the unstructured data as determined by chart annotation.
NLP capabilities survey (initial results)

Distributed to 14 Sentinel Data Partners & 8 partners affiliated with the Innovation Center

A total of 17 responses received (13 from Sentinel Data Partners)

- 12 use NLP in some capacity
- 50% for project-specific research; 50% for research & “operational” purposes

Wide variety of tools used / notes processed (type, number of years)

Scope of concepts extracted also varies widely

- 9 of 12 report being able to extract Diagnoses (highest percentage)
- Handful of other concepts extracted by >50% of respondents (e.g., cancer site & histology, smoking status, signs & symptoms)
Chart annotation - Motivation

Vision
A future state where Sentinel partners with access to EHR data have processed all / some of their clinical notes through an NLP pipeline (or pipelines).

- Some projects may require the development of new pipelines/classifiers,
- Others will rely on the “stock” NLP outputs.

We want to use those derived data elements in a Sentinel analysis.

Issues to consider:
- What note type(s) need to have been processed?
- What time frame had to have been covered?

Example
Looking for history of MI:
- patient had MI 10 years ago

Can we assume it is mentioned in the note at every visit, or just a subset (i.e., first visit with a new provider; every visit for the 2 years after the event, etc.)?
Chart annotation (in progress)

Focus on two use cases

• Hospitalized patients with COVID-19
• Cancer

For both, we propose to look at a subset of notes, since we will not necessarily be able to assume that (future) partners will have run NLP on everything (e.g., all hospital discharge summaries are included, but not respiratory therapist notes)

Purpose is not to develop a classifier or a pipeline, but to describe the information contained in the notes of the patients in each cohort
Hospitalized patients with COVID-19

Population:
- Index event - inpatient encounter with an admitting diagnosis of COVID-19 between April 1, 2020 and December 31, 2021
- Limit to patients who are age >= 18 at the time of admission.

Sampling strategy:
- Cohort 1 – patients without a billing code for supplemental oxygen. Select 35 patients at random.
- Cohort 2 – patients with a billing code for supplemental oxygen. Select 35 patients at random.

Analysis:
- Primary – Pull the discharge summary associated with the hospitalization and annotate priority concepts (e.g., oxygen use, conditions, medication exposure & metadata, smoking status)
- Secondary – For a subset of patients in each cohort (5-10, randomly selected), run a query to identify all notes that include keywords related to oxygen use. Review note / paragraph / sentences around the keyword and determine whether it indicates oxygen use.

Rationale for design choices:
- The secondary analysis will allow us to characterize the degree of “missingness” related to oxygen use, as discharge summaries are not expected to contain the full detail related to oxygen use
- Discharge summaries were chosen because if we are planning to use pre-computed NLP concepts in an analysis, discharge summaries are more likely to be processed across a network than specialty notes (e.g., respiratory therapy)
- Stratifying by billing codes for supplemental oxygen should ensure there is a mix of patients who did and did not receive oxygen compared with a purely random sample of hospitalized patients
Cancer

Population:
Index event
- Patients with a prescription/order for darzalex (daratumumab) and with no prescription/order for darzalex in the prior 3 years.
- Index event should be between January 1, 2016 and November 30, 2021.

Sampling strategy:
- Select 30 patients at random from the cohort.
- Annotate the physician note(s) associated with the visit where the patient was prescribed the medication (assume new prescription occurs in the outpatient setting).

Analysis:
- Annotate selected concepts (e.g., conditions, medications, smoking status, those specific to label);
- Determine if available concepts are sufficient to determine indication behind prescription.

DARZALEX example:

Medication-related concepts
- Diagnosis-related concepts

Concepts that are expected to be primarily NLP-based

1. In combination with bortezomib, melphalan and prednisone for the treatment of patients with newly diagnosed multiple myeloma who are ineligible for autologous stem cell transplant.

2. In combination with lenalidomide and dexamethasone, or bortezomib and dexamethasone, for the treatment of patients with multiple myeloma who have received at least one prior therapy (list of candidate therapies required to define this part).

3. In combination with pomalidomide and dexamethasone for the treatment of patients with multiple myeloma who have received at least two prior therapies including lenalidomide and a proteasome inhibitor.

4. As monotherapy, exclude patients with concurrent candidate therapies for the treatment of patients with multiple myeloma who have received at least three prior lines of therapy including a proteasome inhibitor (PI) and an immunomodulatory agent, or who are double-refractory to a PI and an immunomodulatory agent.
Objective 3 – How to best represent information derived from unstructured text? (in progress)

Process:
Assess current approaches for representing data derived from unstructured text (from other Common Data Models, NLP tools, etc.)

Describe tradeoffs between approaches (e.g., ease of querying, burden on partners, strengthens and weaknesses of different terminologies)

End Product:
Develop set of recommendations for the Sentinel Operations Center as they make decisions on extending the Sentinel Common Data Model
DI3: Identification and mitigation of structured EHR source data mapping issues
Mapping of EHR Data and developing quality metrics

Goal:
To assess the mapping of structured electronic health record (EHR) data to reference terminologies and to develop quality metrics to allow for comparisons across domains within a data source to further identify issues.

Objectives:
1) Develop procedures to assess the mapping of structured EHR data to reference terminologies for laboratory results, medication orders and administrations (inpatient and outpatient) & characterize the severity of issues that are uncovered.

2) Develop standardized metrics related to medications & laboratory results that allow for comparisons across domains within a data source using profiles of records across time, care setting, population, etc. This work will supplement the Sentinel Operations Center’s Data Quality Measures (DQM) in EHRs project by defining new metrics for assessments that are not routinely conducted in EHR datasets.

Project completion date: September 30, 2022
Project team

HPHCI
- Judy Maro, Co-Investigator
- Kevin Coughlin, Project Manager
- Dan Kiernan, Data Analytics

Duke
- Keith Marsolo, PI
- Lesley Curtis, Co-Investigator
- Gretchen Sanders, Project Lead
- Laura Qualls
  - Jennifer Xu, Informaticists
- Larry Hill, Statistician
- Yinhong Zhang
  - Tom Phillips, Programmers

PCORnet Sites
- Alanna Chamberlain, Co-Investigator (Mayo Clinic)
- Jiang Bian, Co-Investigator (U of Florida)

FDA
- Sarah Dutcher
- Monique Falconer
- Jose Hernandez
Motivation – Harmonization of EHR data sources

Many EHR data domains (e.g., medication orders, laboratory results) are not captured in standard formats.

To use these data for research or for data exchange, must harmonize to a reference standard.

Examples shown within these slides are taken from the National Patient-Centered Clinical Research Network (PCORnet®), but the same challenges exist regardless of the source.

For analyses that leverage linked claims-EHR data, findings from this project can provide guidance on the types of EHR data to be included in a CDM and how to ensure and verify accurate transformation.
## Representing a medication in RxNorm

<table>
<thead>
<tr>
<th>RxNorm Term Type</th>
<th>Information encoded</th>
<th>Example medication representation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Most Granular</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Branded Drug</td>
<td>X</td>
<td>Augmentin XR 12 HR 1000 MG Extended Release Oral Tablet</td>
</tr>
<tr>
<td>Semantic Clinical Drug</td>
<td>X</td>
<td>12 HR Amoxicillin 1000 MG / Clavulanate 62.5 MG Extended Release Oral Tablet</td>
</tr>
<tr>
<td>Brand Name Pack</td>
<td>X</td>
<td>N/A</td>
</tr>
<tr>
<td>Generic Pack</td>
<td>X</td>
<td>N/A</td>
</tr>
<tr>
<td>Semantic Branded Drug Form</td>
<td>X</td>
<td>Amoxicillin / Clavulanate Extended Release Oral Tablet</td>
</tr>
<tr>
<td>Semantic Clinical Drug Form</td>
<td>X</td>
<td>Amoxicillin / Clavulanate Extended Release Oral Tablet</td>
</tr>
<tr>
<td>Semantic Branded Dose Form</td>
<td>X</td>
<td>Augmentin Oral Product</td>
</tr>
<tr>
<td>Group*</td>
<td></td>
<td>Augmentin Pill (Requires two records)</td>
</tr>
<tr>
<td>Semantic Clinical Dose Form</td>
<td>X</td>
<td>Amoxicillin / Clavulanate Oral Product; Amoxicillin / Clavulanate Pill (Requires two records)</td>
</tr>
<tr>
<td>Group*</td>
<td></td>
<td>Amoxicillin 1000 MG / Clavulanate 62.5 MG</td>
</tr>
<tr>
<td>Semantic Branded Drug Component</td>
<td>X X X X</td>
<td>Amoxicillin 1000 MG / Clavulanate 62.5 MG [Augmentin]</td>
</tr>
<tr>
<td>Brand Name</td>
<td>X</td>
<td>Augmentin</td>
</tr>
<tr>
<td>Multiple Ingredients</td>
<td>X</td>
<td>Amoxicillin / Clavulanate</td>
</tr>
<tr>
<td>Semantic Clinical Drug Component*</td>
<td>X X</td>
<td>Amoxicillin 1000 MG; Clavulanate 62.5 MG (Requires two records)</td>
</tr>
<tr>
<td>Precise Ingredient</td>
<td>X</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Least Granular</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingredient*</td>
<td>X</td>
<td>Amoxicillin; Clavulanate</td>
</tr>
<tr>
<td>Non-specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dose Form</td>
<td>X</td>
<td>Extended Release Oral Tablet</td>
</tr>
<tr>
<td>Dose Form Group*</td>
<td>X</td>
<td>Oral Product; Pill (Requires two records)</td>
</tr>
<tr>
<td>Prescribable Name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonym</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tall Man Lettering Synonym</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Denotes term types that require multiple records to represent multi-ingredient medications

Within the PCORnet Common Data Model, medication orders and administrations (at most sites) are coded using RxNorm.

RxNorm is an interoperability standard maintained by the National Library of Medicine that represents medication orders and administrations at various levels of granularity.

Even if Sentinel leverages a different standard to represent EHR-based medications, data partners may still need to transform data to/from RxNorm.
PCORnet has defined a set of preferred “tiers” for the different RxNorm Term Types

<table>
<thead>
<tr>
<th>Term Type</th>
<th>Description</th>
<th>Ingredient(s)</th>
<th>Strength</th>
<th>Dose Form</th>
<th>Brand Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SBD</strong></td>
<td>Semantic Branded Drug</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>SCD</strong></td>
<td>Semantic Clinical Drug</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BPCK</strong></td>
<td>Brand Name Pack</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>GPCK</strong></td>
<td>Generic Pack</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SBDF</strong></td>
<td>Semantic Branded Drug Form</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td><strong>SCDF</strong></td>
<td>Semantic Clinical Drug Form</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SBDG</strong></td>
<td>Semantic Branded Dose Form Group*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCDG</strong></td>
<td>Semantic Clinical Dose Form Group*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SBDC</strong></td>
<td>Semantic Branded Drug Component</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BN</strong></td>
<td>Brand Name</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>MIN</strong></td>
<td>Multiple Ingredients</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCDC</strong></td>
<td>Semantic Clinical Drug Component*</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PIN</strong></td>
<td>Precise Ingredient</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td>Ingredient*</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DF</strong></td>
<td>Dose Form</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td><strong>DFG</strong></td>
<td>Dose Form Group*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>PSN</strong></td>
<td>Prescribable Name</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>SY</strong></td>
<td>Synonym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TMSY</strong></td>
<td>Tall Man Lettering Synonym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Denotes term types that require multiple records to represent multi-ingredient medications
### Example quality issue – medication mapping

<table>
<thead>
<tr>
<th>Rank based on Code</th>
<th>RxNorm Code</th>
<th>Medication name (derived from RxNorm code)</th>
<th>Record Count by Code</th>
<th>Rank based on Name</th>
<th>Medication name (from EHR)</th>
<th>Record Count by Name</th>
<th>Percent Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Null or missing</td>
<td>1257171</td>
<td>1 Null or missing</td>
<td>1257171</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>313002 Sodium Chloride 9 MG/ML Injectable Solution</td>
<td>801348</td>
<td>2 Sodium Chloride</td>
<td>1007029</td>
<td>79.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>307668 Acetaminophen 32 MG/ML Oral Suspension</td>
<td>321510</td>
<td>3 Acetaminophen 300MG / Codeine Phosphate 15 MG Oral Tablet</td>
<td>511779</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>197803 Ibuprofen 20 MG/ML Oral Suspension</td>
<td>293209</td>
<td>4 Ibuprofen 20 MG/ML / Pseudoephedrine Hydrochloride 3 MG/ML Oral Suspension</td>
<td>293218</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>540930 Water 1000 MG/ML Injectable Solution</td>
<td>286133</td>
<td>5 Water 1000 MG/ML Injectable Solution</td>
<td>287011</td>
<td>99.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>309778 Glucose 50 MG/ML Injectable Solution</td>
<td>285557</td>
<td>6 Glucose 50 MG/ML / Potassium Chloride 0.01 MEQ/ML / Sodium Chloride 0.0342 MEQ/ML Injectable Solution</td>
<td>286108</td>
<td>99.8%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shading indicates a discordance in medications (e.g., RxNorm code represents a single ingredient in RxNorm vs. multi-ingredient order within the EHR)
Objective 1: Methods to assess mapping of structured EHR data to reference terminologies

General approach:

- Develop queries to assess mapping of medication orders, medication administrations and laboratory tests – limit analysis to the top 200 by volume

- For each medication / lab, generate statistics on all the different combinations within the structured fields and “raw” source fields

- For example, for a given medication name, summarize the number of records/patients for associated RxNorm codes, dose units, dose forms, as well as the corresponding “raw” fields

<table>
<thead>
<tr>
<th>RAW Medication Name</th>
<th>RxNorm Code</th>
<th>CDM Dose Unit</th>
<th>RAW Dose Unit</th>
<th>Number of Records</th>
<th>Number of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALCICUM CARBONATE 300 MG (750 MG) CHEWABLE TABLET</td>
<td>1044532</td>
<td>Other</td>
<td>mg of elemental</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1044532</td>
<td>Other</td>
<td>mg of salt</td>
<td>50564</td>
<td>14817</td>
</tr>
<tr>
<td></td>
<td>1044532</td>
<td>Other</td>
<td>tablet</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1484737</td>
<td>Other</td>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1484737</td>
<td>Other</td>
<td>mg of elemental</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1484737</td>
<td>Other</td>
<td>mg of salt</td>
<td>51092</td>
<td>14887</td>
</tr>
<tr>
<td></td>
<td>1484737</td>
<td>Other</td>
<td>tablet</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Objective 1: Evaluation

- Generate statistics on number of medication codes/laboratory tests associated with more than one name within the EHR and vice versa
- Concordance between lab name / medication name (brand and/or ingredient) within the EHR and that derived from the associated code
- Concordance between discrete fields (e.g., lab result unit, medication dose, etc.) and those associated with the associated LOINC / RxNorm code
- Generate characterization of issues by severity (e.g., LOINC code mis-match, combination medication represented by single-ingredient RxNorm code, generic medication represented by brand name, etc.)

End product:
Procedures that can be used to assess mapping of structured EHR domains and a set of statistics on the severity of issues at 2 pilot sites (PCORnet).

<table>
<thead>
<tr>
<th>Severity</th>
<th>Example issue</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>(1) Lab test mismatch (incorrect LOINC code) (2) Multi-ingredient drug uses single ingredient RxNorm code (3) Single ingredient drug uses multi-ingredient RxNorm code</td>
<td>(1-3) The LOINC/RxNorm codes that are assigned to these records are incorrect and would not actually represent the test result or exposure to the specified medication.</td>
</tr>
<tr>
<td>Major</td>
<td>(1) Ingredient-level RxNorm code utilized when more granular available (single-ingredient drugs only) (2) More granular RxNorm code used than supported by the data</td>
<td>(1) The ingredient is correct, but the other metadata is missing, meaning those records may be excluded if the drug has forms that are not part of an analysis (i.e., topical creams). (2) This example is the inverse — records that should have been excluded were included.</td>
</tr>
<tr>
<td>Moderate</td>
<td>(1) Generic medication uses brand name RxNorm code (2) Brand name medication uses a generic-level RxNorm code</td>
<td>(1) Any study that looking for the use of a specific brand of medication will include extra records. (2) Studies that are looking at the use of a specific branded medication will miss records.</td>
</tr>
<tr>
<td>Minor</td>
<td>(1) Distribution of lab results is an outlier for a given LOINC.</td>
<td>(1) The test may be only used on specific populations (e.g., inpatients), which may bias results.</td>
</tr>
</tbody>
</table>
Example quality issue – differences based on provenance (orders vs. medication administrations)
Objective 2: Standardized metrics to generate comparisons based on provenance

General approach:

Develop queries that will support the comparison of records based on provenance – medication orders vs. administrations; billed diagnoses vs. clinician-entered – to identify potential data issues.

Define specific conditions & associated concepts to investigate (e.g., diagnoses, procedures, medications, labs). Look at values within each cohort as well as the population as a whole.

Distribute query package to partner sites to generate summary statistics. Focus of analysis will be within-DataMart comparisons, though cross-DataMart comparisons are also possible.

End product:
Set queries to support cross-domain comparisons within a dataset, at both condition and population-level, along with statistics describing the performance of each at partners sites.
Questions?
**Sentinel Initiative**

**Priorities**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data infrastructure</strong></td>
<td><strong>Feature engineering</strong></td>
<td><strong>Causal inference</strong></td>
<td><strong>Detection analytics</strong></td>
<td><strong>Innovation incubator</strong></td>
</tr>
<tr>
<td>Identification and queries of potential EHR data partners (Horizon Scan: DI1)</td>
<td>Adding unstructured data and necessary data elements (DI2)</td>
<td>Source data mapping (DI3)</td>
<td>Identifying and evaluation of EHR detection approaches (DA1)</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td>Updating CDM to include EHR data</td>
<td>Data quality metrics and quality assurance strategy</td>
<td>Empirical evaluation of EHR-based detection approaches (DA2)</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td>Source data mapping (DI3)</td>
<td>Harmonizing EHRs (DI4)</td>
<td>Development of EHR-based detection tools</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data harmonization strategy</td>
<td>Methods framework for EHR-based signal detection</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmonizing EHRs (DI4)</td>
<td>Methods for signal detection for pregnancy/birth outcomes (DA4)</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pregnancy and birth outcomes signal detection tool development</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Methods for cancer signal detection (DA5)</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cancer signal detection tool development</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
<td><strong>Master plan refinement</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FE1**: Computable phenotyping framework

**FE2**: NLP tools for cohort identification, exposure assessment, covariate ascertainment

**FE3**: Improving probabilistic phenotyping of incident outcomes
Health Outcomes and Covariates for Computable Phenotyping Using EHR Data

Lessons Learned from: Advancing scalable natural language processing approaches for unstructured electronic health record data

Workgroup Leads: David S. Carrell, PhD
Outline

• **Motivation**
  • Role of computable algorithms in Sentinel
  • Limitations of claims data
  • The promise of using EHR data and machine learning (ML) methods
• **Scalable algorithm development**
• **Filters in outcome identification**
  • Role in outcome identification
  • Data-driven, high-sensitivity filtering (HSF)
Motivation: Role of computable algorithms in Sentinel

Allow safety issues to be investigated rapidly, at ~low cost
ARIA = the Active Risk Identification and Analysis system

- Analytic Tools
- Common Data Model
- ARIA

- Claims data (no manual chart review required)
- Pre-defined, parameterized, re-usable tools

*slide courtesy of Michael Nguyen*
Motivation: Limitations of structured claims data

Reliance on existing Sentinel data in ARIA analyses has revealed various insufficiencies.

Incorporating rich EHR data may overcome some of these insufficiencies.

<table>
<thead>
<tr>
<th>Study Pop</th>
<th>Exposure</th>
<th>Outcome</th>
<th>Covariate</th>
<th>Analysis Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>13</td>
<td>66</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>22</td>
<td>21</td>
<td>54</td>
<td>26</td>
</tr>
</tbody>
</table>

This slide courtesy of Michael Nguyen
**Motivation: Promise of EHR data + ML methods**

Accurate identification of some outcomes/covariates requires information only available in EHR data and clinical notes

- Ex. 1: Identification of acute pancreatitis requires labs data (lipase)
- Ex. 2: Key facts for identifying anaphylaxis are absent in claims data but can be extracted from EHRs via natural language processing (NLP)

Relationships between rich features/predictors and outcomes are often nonlinear, making data-driven ML modeling advantageous

- Ex.: Computable algorithms for identifying anaphylaxis based on ML methods consistently outperformed simpler linear models
Scalable algorithm development

Efficiency: At reasonable cost in a ~short time frame
  • Cost/time drivers are personnel salaries, gold standard creation

Portability: Easily implemented in diverse real-world settings
  • Sharable tools/packages
  • Minimal/no local tailoring needed
  • Anticipates & accommodates local systems & data

Replicability
  • Comparable results across settings
  • Comparable results across time

Efficiency + Portability + Replicability = Scalable algorithm development

Scalable algorithm development is needed to:
  • Keep pace with demand for safety analyses
  • Produce results at reasonable cost
Filters: Their role in outcome identification

Filters are:

- Expert-specified sets of healthcare data, e.g., diagnosis, procedure, or medication codes
- That *presumptively* identify patients w/ the outcome
- For which true case status will be *determined by a computable algorithm*

Useful filters have:

- Strong face validity
- Simple and generalizable definitions
- High sensitivity (to minimize selection bias)
- Reasonable specificity
  (to limit data collection burden)

- Traditional example: COVID-19-specific ICD-10 dx codes
Filters: Data-driven, high-sensitivity filtering (HSF)

Objective:
Improve sensitivity of a “traditional” filter

HSFs use data-driven analytics to identify additional filtering codes:
• To identify patients/events overlooked by simple/traditional filters,
• With modest increase in overall sample size, and
• With reasonable effort (i.e., reusable tool applied to Sentinel data)

How do HSFs work?
1. Divide patients into two groups:
   • Ever qualified by the traditional filter
   • Never qualified by the traditional filter
2. Identify codes that are ≥10x more common in “Ever” than “Never” patients
3. Manually review and retain identified codes with face validity
4. Add patients/events w/any HSF code to the presumptive patient/event set

Filters: Data-driven, high-sensitivity filtering (HSF)

COVID-19-specific dxs \(^1\) (traditional filter)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19
Filters: Data-driven, high-sensitivity filtering (HSF)

**COVID-19-specific dxs**¹ (*traditional filter*)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19
Filters: Data-driven, high-sensitivity filtering (HSF)

COVID-19-specific dxs\(^1\) (traditional filter)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19

Algorithm to distinguish non-cases/cases
 Filters: Data-driven, high-sensitivity filtering (HSF)

COVID-19-specific dxs\(^1\) (traditional filter)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19

Algorithm to distinguish non-cases/cases

Can HSFs capture overlooked patients?
- Other diagnoses?
- Procedures?
- Medications?
- Labs? ...
Filters: Data-driven, high-sensitivity filtering (HSF)

COVID-19-specific dxs \(^1\) (traditional filter)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19

Can HSFs capture overlooked patients?
- Other diagnoses?
- Procedures?
- Medications?
- Labs? ...
Filters: Data-driven, high-sensitivity filtering (HSF)

COVID-19-specific dxs \(^1\) (*traditional filter*)
- B9729, COVID-19, pre 4/1/2020
- U07.1, COVID-19, post 4/1/2020
- Z8616, Hx of COVID-19

Can HSFs capture overlooked patients?
- Other diagnoses?
- Procedures?
- Medications?
- Labs? ...

If so ...
How many (sensitivity)?
At what cost (data burden)?
Results: COVID-19 high-sensitivity filtering (HSF)

<table>
<thead>
<tr>
<th></th>
<th>VU: +13% true cases, +22% pts</th>
<th>KP: +10% true cases, +22% pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>20,951 patients (~90% true case rate)</td>
<td>6,847 patients (~71% true case rate)</td>
</tr>
<tr>
<td>True Cases</td>
<td>4,566 patients (+22%) (~55% true case rate)</td>
<td>1,482 patients (+22%) (~38% true case rate)</td>
</tr>
</tbody>
</table>
Thank You!

David S. Carrell, PhD
Kaiser Permanente Washington Health Research Institute
Seattle, WA
david.s.carrell@kp.org
Extras
COVID-19 as a covariate in safety studies?

- **Nature Medicine**
  “... beyond the first 30 d after infection, individuals with COVID-19 are at increased risk of incident cardiovascular disease spanning several categories, including cerebrovascular disorders, dysrhythmias, ischemic and non-ischemic heart disease, pericarditis, myocarditis, heart failure and thromboembolic disease.”

- **JAMA**
  “Physicians should consider a history of COVID-19 as a cardiovascular disease risk.”

---


Abbasi J. The COVID Heart—One Year After SARS-CoV-2 Infection, Patients Have an Array of Increased Cardiovascular Risks. JAMA. Published online March 02, 2022. doi:10.1001/jama.2022.2411
**FE2: NLP tools for cohort identification, exposure assessment, covariate ascertainment ("Scalable NLP")**

**Goal:** In two heterogeneous settings develop and validate scalable and reusable NLP tools for leveraging EHR data to address known insufficiencies in existing data and methods to support FDA safety surveillance studies.

**Progress:**

- **Objective 1: Cohort identification**
  - Developed and evaluated scalable, replicable approaches to cohort identification in Sentinel safety studies
  - Products: ICPE 2022 and AMIA 2022 abstracts (at right)

- **Objective 2: Scalable NLP measures**
  - Develop, apply and evaluate scalable, replicable methods for NLP-based measurement of exposures, symptoms, and outcomes

- **Objective 3: Evaluation**
  - Compare structured data versus NLP for capturing exposures, health outcomes of interest, and covariates

**Deliverable for the IC:** Manuscript describing key products of this work and a GitHub repository of reusable tools and methods for incorporating scalable NLP into Sentinel safety studies

**WG Leads:** David Carrell, KPWA, Joshua Smith, VUMC; Timeline: 2/1/2021-1/31/2023
### VUMC patients identified by COVID-19 "base" and "high-sensitivity" (HSF) filters during study period

<table>
<thead>
<tr>
<th>Filter rank</th>
<th>COVID-19 filter category</th>
<th>N patients with this filter</th>
<th>N patients with this filter and no higher rank filters</th>
<th>Percent of all patients identified by this filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Diagnosis of U07.1 &quot;COVID-19&quot; (base #1)</td>
<td>20,840</td>
<td>20,840</td>
<td>80%</td>
</tr>
<tr>
<td>2nd</td>
<td>Any of 5 other COVID-19 diagnoses (base #2)</td>
<td>1,898</td>
<td>111</td>
<td>0.43%</td>
</tr>
<tr>
<td>3rd</td>
<td>HSF diagnoses (any of 24)</td>
<td>7,264</td>
<td>3,976</td>
<td>15%</td>
</tr>
<tr>
<td>4th</td>
<td>HSF procedures (any of 10)</td>
<td>1198</td>
<td>37</td>
<td>0.14%</td>
</tr>
<tr>
<td>5th</td>
<td>HSF medications (any of 4)</td>
<td>473</td>
<td>181</td>
<td>0.70%</td>
</tr>
<tr>
<td>6th</td>
<td>HSF problem list in EHR (any of 5)</td>
<td>9,222</td>
<td>892</td>
<td>3.4%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>26,037</td>
<td>100%</td>
</tr>
</tbody>
</table>

If we included a 7th filter, PCR+ COVID-19 test (only), **8,825 (+34%)** new patients would be added.
# High-sensitivity COVID-19 filter results -- KPWA

KPWA patients identified by COVID-19 "base" and "high-sensitivity" (HSF) filters during study period

<table>
<thead>
<tr>
<th>Filter rank</th>
<th>COVID-19 filter category</th>
<th>N patients with this filter</th>
<th>N patients with this filter and no higher rank filters</th>
<th>Percent of all patients identified by this filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Diagnosis of U07.1 &quot;COVID-19&quot; (base #1)</td>
<td>15,678</td>
<td>15,678</td>
<td>81%</td>
</tr>
<tr>
<td>2nd</td>
<td>Any of 5 other COVID-19 diagnoses (base #2)</td>
<td>1,498</td>
<td>166</td>
<td>1%</td>
</tr>
<tr>
<td>3rd</td>
<td>HSF diagnoses (any of 24)</td>
<td>5,041</td>
<td>2,789</td>
<td>14%</td>
</tr>
<tr>
<td>4th</td>
<td>HSF procedures (any of 10)</td>
<td>550</td>
<td>8</td>
<td>0.04%</td>
</tr>
<tr>
<td>5th</td>
<td>HSF medications (any of 4)</td>
<td>91</td>
<td>84</td>
<td>0.4%</td>
</tr>
<tr>
<td>6th</td>
<td>HSF problem list in EHR (any of 5)</td>
<td>4,845</td>
<td>607</td>
<td>3%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>19,332</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

If we included a 7th filter, PCR+ COVID-19 test (only), 4,737 (+25%) new patients would be added.
Automated Methods for Developing Computable Phenotypes

Lessons Learned from: Advancing scalable natural language processing approaches for unstructured electronic health record data

Workgroup Leads: Joshua C. Smith & David S. Carrell
Phenotyping

Computable phenotype algorithms typically:
- Require time-intensive expert curation and feature engineering
- Require manually-annotated gold-standard training sets
- Result in high cost and limited scalability.

PheNorm, and similar automated approaches:
- Based on natural language processing (NLP), machine learning, and (low-cost) silver-standard training labels
- Have been demonstrated to perform well for various chronic health conditions.

We evaluated PheNorm for use with acute conditions (COVID-19)
- PheNorm currently being applied to acute pancreatitis in another IC project
Rationale for exploring automating phenotyping methods

Scalability
• Manual approach is burdensome/slow, requires substantial expertise

Replicability
• Reduced operator-dependence

Hybrid solutions?
• PheNorm → PheCAP → blended methods?
Rationale for exploring automating phenotyping methods

Continuum of development approaches

Manual development
- Expert-driven
- *Manual* engineering
- Heavy reliance on *gold standard labels*
- Substantial operator dependence
- Slow

Automated development
- Data-driven
- Automated engineering
- Heavy reliance on *silver standard labels*
- Reduced operator dependence
- Fast

- Automated feature engineering (AFEP)$^1$
- Surrogate-assisted feature extraction (SAFE)$^2$
- Phenotype algorithm normalization (PheNorm)$^3$
- Phenotyping common approach (PheCAP)$^4$

---

1. Yu et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. JAMIA 2015
2. Yu et al. Surrogate-assisted feature extraction for high-throughput phenotyping. JAMIA 2017
3. Yu et al. Enabling phenotypic big data with PheNorm. JAMIA 2018
Automated modeling: PheNorm

Overview of PheNorm/PheCap


ICD-10 diagnosis codes for COVID-19 and other diagnosis codes (HSF*)
- Cases & non-cases

- Index date
- Catchment period
- Define silver label & counts of DX codes

- Feature selection
- Transformation s

- Predicted probability of being a case
- Phenotype classification (Yes/No)

- Evaluate using gold-labeled data (not included)
Automating NLP dictionary creation (AFEP)

5 clinical knowledge base articles on a topic

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

**AFEP Dictionary**
- 159 CUIs extracted from 6 articles on COVID-19

**Data/text catchment Period**
- Index date +/-30 days

**Input Data**
- KPWA: 143,584 notes from 8,329 patients
- VUMC: Approximately 1.1 million notes from 24,355 patients

**Process notes using MetaMapLite**
- Transform counts of each NLP-extracted concept from the AFEP dictionary into input vectors for PheNorm
Running PheNorm

Silver Standard Labels

1. **Structured Label** – count of days with U07.1 diagnosis code (COVID-19)
2. **Structured Label** – counts of six COVID-related CUIs
3. **NLP Label** – Cumulative count of “COVID-19” mentions in patients’ charts
4. **NLP Label** – number of days (KPWA) or notes (VUMC) in which a COVID-19 concepts was mentioned in patients charts

• Apply PheNorm, **evaluate**
COVID-19 Phenotype

Evidence of COVID-19 infection

Definite or highly probable infection
- Lab data or clinical note indicates patient was PCR-positive or
- Assertion the patient has COVID-19 in a free text statement or
- Strong evidence of proximal exposure and serologic evidence of prior infection

Probable or possible infection
- Patient symptoms are consistent with a diagnosis of COVID-19
- Absence of an explicit alternative diagnosis and/or absence of a statement that a non-COVID-19 cause is more likely
- Strong evidence of proximal exposure

Unlikely infection
- Explicit alternative diagnosis or statement that a non-COVID-19 cause is more likely
- Absence of symptoms consistent with a diagnosis of COVID-19 and absence of lab data or clinical note indicating a positive PCR test

Not infected
- No indication in the EHR of infection [i.e., symptoms, exposure, and/or labs/serology] during the relevant time window) EHR appears to thoroughly document the patient’s care during the relevant time window

Insufficient Information
- EHR appears not to be a reasonably complete source of documentation about the patient’s care during the relevant time window

Severity of illness scale (NIH)

<table>
<thead>
<tr>
<th>SEVERITY LEVEL</th>
<th>SIGN/SYMPOTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptomatic</td>
<td>No symptoms</td>
</tr>
<tr>
<td>Mild</td>
<td>Fever (&gt;=100.4F)</td>
</tr>
<tr>
<td></td>
<td>Cough</td>
</tr>
<tr>
<td></td>
<td>Sore throat</td>
</tr>
<tr>
<td></td>
<td>Malaise/fatigue</td>
</tr>
<tr>
<td></td>
<td>Headache</td>
</tr>
<tr>
<td></td>
<td>Muscle pain</td>
</tr>
<tr>
<td></td>
<td>Nausea</td>
</tr>
<tr>
<td></td>
<td>Vomiting</td>
</tr>
<tr>
<td></td>
<td>Diarrhea</td>
</tr>
<tr>
<td></td>
<td>Loss of sense of taste or smell</td>
</tr>
<tr>
<td>Moderate</td>
<td>Shortness of breath (SpO2 &gt;=94%)</td>
</tr>
<tr>
<td></td>
<td>Dyspnea (SpO2 &gt;=94%)</td>
</tr>
<tr>
<td></td>
<td>Abnormal chest imaging (SpO2 &gt;=94%)</td>
</tr>
<tr>
<td>Severe</td>
<td>SpO2 &lt;94%</td>
</tr>
<tr>
<td></td>
<td>PaO2/FiO2* &lt;300 mm Hg</td>
</tr>
<tr>
<td></td>
<td>Respiratory freq &gt;30 breaths/min</td>
</tr>
<tr>
<td></td>
<td>Lung infiltrates &gt;50%</td>
</tr>
<tr>
<td>Critical</td>
<td>Respiratory failure</td>
</tr>
<tr>
<td></td>
<td>Septic shock</td>
</tr>
<tr>
<td></td>
<td>Multiple organ dysfunction</td>
</tr>
</tbody>
</table>
# COVID-19 phenotype chart review results

<table>
<thead>
<tr>
<th>Study site</th>
<th>COVID-19 phenotype definition</th>
<th>Chart review result</th>
<th>Number of charts</th>
<th>Percent of charts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VUMC</strong> (N=483)</td>
<td>Moderate+ severity</td>
<td>Non-case</td>
<td>334</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case</td>
<td>149</td>
<td>31%</td>
</tr>
<tr>
<td></td>
<td>Mild+ severity</td>
<td>Non-case</td>
<td>188</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case</td>
<td>295</td>
<td>61%</td>
</tr>
<tr>
<td><strong>KPWA</strong> (N=437)</td>
<td>Moderate+ severity</td>
<td>Non-case</td>
<td>315</td>
<td>72%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case</td>
<td>122</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>Mild+ severity</td>
<td>Non-case</td>
<td>168</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case</td>
<td>269</td>
<td>62%</td>
</tr>
</tbody>
</table>

Chart samples were stratified to represent all filter types (not a random sample of all eligible charts)
# PheNorm Results – Moderate+ Phenotype

<table>
<thead>
<tr>
<th>Site</th>
<th>Silver Standard</th>
<th>Phenotype</th>
<th>AUC</th>
<th>Sensitivity at PPV=0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPWA</td>
<td>1 - U07.1 Days</td>
<td>Moderate+</td>
<td>0.700</td>
<td>0.07</td>
</tr>
<tr>
<td>VUMC</td>
<td>1 - U07.1 Days</td>
<td>Moderate+</td>
<td>0.814</td>
<td>0.29</td>
</tr>
<tr>
<td>KPWA</td>
<td>2 - Six-CUI Days</td>
<td>Moderate+</td>
<td>0.695</td>
<td>0.05</td>
</tr>
<tr>
<td>VUMC</td>
<td>2 - Six-CUI Days</td>
<td>Moderate+</td>
<td>0.841</td>
<td>0.47</td>
</tr>
<tr>
<td>KPWA</td>
<td>3 - COVID Mentions</td>
<td>Moderate+</td>
<td>0.674</td>
<td>0.00</td>
</tr>
<tr>
<td>VUMC</td>
<td>3 - COVID Mentions</td>
<td>Moderate+</td>
<td>0.775</td>
<td>0.29</td>
</tr>
<tr>
<td>KPWA</td>
<td>4A - CUI Days</td>
<td>Moderate+</td>
<td>0.695</td>
<td>0.00</td>
</tr>
<tr>
<td>VUMC</td>
<td>4B - CUI Notes</td>
<td>Moderate+</td>
<td>0.768</td>
<td>0.27</td>
</tr>
</tbody>
</table>
## Sentinel Initiative

### PheNorm Results – Symptomatic COVID-19

<table>
<thead>
<tr>
<th>Site</th>
<th>Silver Standard</th>
<th>Phenotype</th>
<th>AUC</th>
<th>Sensitivity at PPV=0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPWA</td>
<td>1 - U07.1 Days</td>
<td>Symptomatic</td>
<td>0.773</td>
<td>0.89</td>
</tr>
<tr>
<td>VUMC</td>
<td>1 - U07.1 Days</td>
<td>Symptomatic</td>
<td>0.901</td>
<td>0.99</td>
</tr>
<tr>
<td>KPWA</td>
<td>2 - Six-CUI Days</td>
<td>Symptomatic</td>
<td>0.766</td>
<td>0.88</td>
</tr>
<tr>
<td>VUMC</td>
<td>2 - Six-CUI Days</td>
<td>Symptomatic</td>
<td>0.899</td>
<td>0.95</td>
</tr>
<tr>
<td>KPWA</td>
<td>3 - COVID Mentions</td>
<td>Symptomatic</td>
<td>0.864</td>
<td>0.98</td>
</tr>
<tr>
<td>VUMC</td>
<td>3 - COVID Mentions</td>
<td>Symptomatic</td>
<td>0.887</td>
<td>0.94</td>
</tr>
<tr>
<td>KPWA</td>
<td>4A - CUI Days</td>
<td>Symptomatic</td>
<td>0.892</td>
<td>0.98</td>
</tr>
<tr>
<td>VUMC</td>
<td>4B - CUI Notes</td>
<td>Symptomatic</td>
<td>0.875</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Prediction Performance

Moderate+ phenotype, Silver #1 – U07.1 Days

![Prediction Performance Diagram](image1)

![Prediction Performance Diagram](image2)
Prediction Performance

Moderate+ phenotype, **Silver #2 – “Six-CUI” Days**

![Prediction Performance vs. Cutoff Graph](image1)

- **PPV = 0.80**
- Sensitivity = 0.47

![Prediction Performance vs. Cutoff Graph](image2)

- **PPV = 0.80**
- Sensitivity = 0.05
Prediction Performance

Mild+ phenotype, Silver #1 – U07.1 Days

PPV = 0.80
Sens = 0.99

PPV = 0.80
Sens = 0.89
Prediction Performance

Mild+ phenotype, Silver #3 – COVID-19 Mentions

PPV = 0.80
Sens = 0.94

PPV = 0.80
Sens = 0.98
Prediction Performance

Mild+ phenotype, **Silver #4** – COVID Notes / COVID Days

**Prediction Performance vs. Cutoff**

- PPV = 0.80
- Sens = 0.95

**Prediction Performance vs. Cutoff**

- PPV = 0.80
- Sens = 0.98
Take-home messages

- **Relevance to Sentinel safety surveillance**
  - *Relatively modest effort* was needed to implement this approach
  - *Replication* in (two) heterogeneous settings was straightforward
  - May be relevant for both chronic and acute health conditions

- **Performance of automated models**
  - “Fit” between *silver label* and phenotype definition appears important
  - “Fit” between *source data* and *phenotype definition* appears important (e.g., inpatient data needed for moderate+ severity)
  - When performance is less than desirable, automated approaches may still be a useful *starting point* for model development

- **Hybrid approaches – automated and manually-curated features**
  - PheCap and Multimodal Automated Phenotyping (MAP)
More information

Data-driven automated classification algorithms for acute health conditions: Applying PheNorm to COVID-19 disease

• Abstract submitted for AMIA 2022 Annual Symposium

Joshua Smith, PhD (VUMC)  David Carrell, PhD (KPWA)
joshua.smith@vumc.org  david.s.carrell@kp.org
Large-scale Phenotyping With Natural Language Processing

Cosmin Adrian Bejan, PhD

Department of Biomedical Informatics
Desiderata for NLP-based phenotyping

- Improve phenotype identification based on structured data
- Analyze large volumes of clinical notes
- Data-driven generation of phenotype profiles
- Minimize the amount of chart review
- Generalize across phenotypes
- Replicate across EHR repositories
Proposed NLP system architecture

1. Phenotype query
   - Term seeds
   - Ranked terms related to seeds
   - Query expansion
     - Lexical association
     - word2vec

2. Phenotype retrieval
   - TF-IDF weighting
   - Negation filtering

3. Relevance assessment

4. Vanderbilt EHR
   - > 300 million clinical notes

Query reformulation based on relevance feedback

Ranked patients
Proposed NLP system architecture

- Phenotype query
  - Query expansion
    - lexical association
    - word2vec
  - term seeds
  - ranked terms related to seeds

- Phenotype retrieval
  - TF-IDF weighting
  - negation filtering

- Vanderbilt EHR
  - > 300 million clinical notes

- Relevance assessment
- Ranked patients

Query reformulation based on relevance feedback
Proposed NLP system architecture

1. Phenotype query
   - term seeds
   - ranked terms related to seeds

2. Query expansion
   - lexical association
   - word2vec

3. Phenotype retrieval
   - TF-IDF weighting
   - negation filtering

4. Vanderbilt EHR
   - > 300 million clinical notes

5. Relevance assessment

query reformulation based on relevance feedback
Proposed NLP system architecture

3

Phenotype query

Phenotype retrieval
- TF-IDF weighting
- negation filtering

Relevance assessment

Query expansion
- lexical association
- word2vec

ranked terms related to seeds
term seeds

Vanderbilt EHR
> 300 million clinical notes

Ranked patients

query reformulation based on relevance feedback
Proposed NLP system architecture

- **Phenotype query**
  - term seeds
  - ranked terms related to seeds

- **Query expansion**
  - lexical association
  - word2vec

- **Phenotype retrieval**
  - TF-IDF weighting
  - negation filtering

- **Vanderbilt EHR**
  - > 300 million clinical notes

- **Relevance assessment**

- **Ranked patients**

query reformulation based on relevance feedback
Applications

Social determinants of health
- Homelessness (VUMC)
- Adverse Childhood Experiences (VUMC)
- Homelessness (OHSU)
- Social Isolation (OHSU)
- Financial Insecurity (OHSU)
- Chronic Stress (OHSU)

Suicide phenotypes
- Suicidal Ideation (VUMC)
- Suicide Attempt (VUMC)
- Suicide Attempt - incidence

(Bejan et al., *JAMIA* 2018)

(Dorr, Bejan et al., *MedInfo* 2019)

(Bejan et al., *medRxiv* 2022)

(Walsh et al., *submitted*)
Data-driven methods for extracting phenotype profiles

**Homelessness**

<table>
<thead>
<tr>
<th>rank[cosine(homelessness+homelessness, w)]</th>
<th>rank[cosine(homeless, w)]</th>
<th>rank[cosine(homelessness, w)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>context size=5</td>
<td>context size=15</td>
<td>context size=5</td>
</tr>
<tr>
<td>1 homeless</td>
<td>homeless</td>
<td>prison</td>
</tr>
<tr>
<td>2 homelessness</td>
<td>homelessness</td>
<td>jail</td>
</tr>
<tr>
<td>3 prison</td>
<td>methamphetamine</td>
<td>ex-wife</td>
</tr>
<tr>
<td>4 sober</td>
<td>polysubstance</td>
<td>girlfriend</td>
</tr>
<tr>
<td>5 jail</td>
<td>schizophrenia</td>
<td>sober</td>
</tr>
<tr>
<td>6 abuser</td>
<td>sober</td>
<td>abuser</td>
</tr>
<tr>
<td>7 prostitution</td>
<td>prison</td>
<td>live-in</td>
</tr>
<tr>
<td>8 polysubstance</td>
<td>prostitute</td>
<td>sober</td>
</tr>
<tr>
<td>9 ex-wife</td>
<td>ecstasy</td>
<td>ex-husband</td>
</tr>
<tr>
<td>10 ex-husband</td>
<td>mthmhi</td>
<td>fiancee</td>
</tr>
</tbody>
</table>

**Suicide phenotypes**

<table>
<thead>
<tr>
<th>suicide+ suicidal</th>
<th>suicide</th>
<th>suicidal</th>
</tr>
</thead>
<tbody>
<tr>
<td>context size = 5</td>
<td>context size = 15</td>
<td>context size = 5</td>
</tr>
<tr>
<td>1 suicide</td>
<td>suicide</td>
<td>self-harm</td>
</tr>
<tr>
<td>2 suicidal</td>
<td>suicidal</td>
<td>ideation</td>
</tr>
<tr>
<td>3 ideation</td>
<td>paranormal</td>
<td>suicidal</td>
</tr>
<tr>
<td>4 homicidal</td>
<td>homicidal</td>
<td>suicide/ideation</td>
</tr>
<tr>
<td>5 self-harm</td>
<td>ideation</td>
<td>suicidal</td>
</tr>
<tr>
<td>6 ideations</td>
<td>manic</td>
<td>mania</td>
</tr>
<tr>
<td>7 paranoia</td>
<td>self-harm</td>
<td>paranoia</td>
</tr>
<tr>
<td>8 paranoid</td>
<td>mania</td>
<td>s/h</td>
</tr>
<tr>
<td>9 suicidal</td>
<td>suicidality</td>
<td>self-mutilization</td>
</tr>
<tr>
<td>10 suicidality</td>
<td>paranoia</td>
<td>manic</td>
</tr>
</tbody>
</table>
Building phenotype queries (I)

**ACE**
- child abuse
- sexual abuse
- child neglect
- childhood trauma
- child protective service
- physical abuse
- psychological abuse
- verbal abuse
- poverty
- food insecurity
- cps supervisor
- cps report
- cps worker
- cps investigation

**Homelessness**
- homeless
- homelessness
- shelter
- unemployed
- jobless
- incarceration
Building phenotype queries (II)

- **Suicidal Ideation**
  
  suicid(al|e) idea(tion|s)*
  suicid(al|e) thought(s)*
  thought(s)* of suicide
  
  (wish|wishes|intent|intend|intends|plans) to commit suicide
  (want|wish) (s|ing|es)* to die
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (take|end) (ing)* (my|his|her|their) (own)* life
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about)
  (kill|shot|shoot|hang|poison|asphyxiate|asphyxiat|mutilate|mutilat|harm|overdose|overdos|cut|cutt|gas|gass|slash) (ing)* (myself|himself|herself|themself)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (slit|slitt|cut|cutt|slash) (ing)* (my|his|her|their|the)* (wrist|arm|throat)
  
  feel(s|ing) (very)* suicidal
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) off (a|the|interstate|my|his|her|their)*
  (bridge|building|balcony|window|roof)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) out of (a|the)* moving (vehicle|car)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) from a moving (vehicle|car)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) out of (his|her|the|a)* (\(d+\)) (nd|rd|th)
  (floor|story|balcony|window)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) in front of a (car|truck|train|vehicle)
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) into interstate
  (thoughts|think|want|wish) (s|ing|es)* (of|to|about) (jump|jumping) out of (a|the|his|her)* (window|balcony)
Building phenotype queries (III)

- **Suicide Attempt**
  - suicid(al|e) attempt
  - suicid(al|e) ideation and attempt
  - (attempted|committed) suicide
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (take|end) (ing)* (my|his|her|their) (own)* life
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (kill|shot|shoot|hang|poison|asphyxiate|asphyxiat|mutilate|mutilat|harm|overdose|overdos|cut|cutt|gas|gass|slash) (ing)* (myself|himself|herself|themselves)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (slit|slitt|cut|cutt|slash) (ing)* (my|his|her|their|the)* (wrist|arm|throat)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) off (a|the) interstate (my|his|her|their)* (bridge|building|balcony|window|roof)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) out of (a|the)* moving (vehicle|car)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) from a moving (vehicle|car)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) out of (his|her|the)a* (d+) (nd|rd|th) (floor|story|balcony|window)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) in front of a (car|truck|train|vehicle)
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) into interstate
  - (try|tried|tries|trying|attempted|attempts|attempting) (of|to) (jump|jumping) out of (a|the|his|her)* (window|balcony)
Patient retrieval evaluation (Top K)

Homelessness
- P@185 = 93%
- N=35,220

ACE
- P@185 = 76%
- N=27,861

Suicidal Ideation
- P@200 = 98.5%
- N=187,047

Suicide Attempt
- P@200 = 96.5%
- N=52,738
ICD-based identification of suicide phenotypes

**Suicidal Ideation**
- \( P(\text{ICD10CM}) = 96\% \)

**Suicide Attempt**
- \( P(\text{ICD10CM}) = 85\% \)
From ranking to classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Patient ID</th>
<th>NLP score</th>
<th>Case label</th>
<th>K</th>
<th>P@K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>4,717</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>0</td>
<td>4</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>?</td>
<td>5</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>?</td>
<td>6</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Nrank</td>
<td>.</td>
<td>?</td>
<td>N</td>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>
## From ranking to classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Patient ID</th>
<th>NLP score</th>
<th>Case label</th>
<th>K</th>
<th>P@K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>4,717</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>4</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>5</td>
<td>?</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
<td>?</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>?</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
<td>?</td>
</tr>
<tr>
<td>Nrank</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>N</td>
<td>?</td>
</tr>
</tbody>
</table>

K = ? & P@K=70

**cases**

**non-cases**
From ranking to classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Patient ID</th>
<th>NLP score</th>
<th>Case label</th>
<th>K</th>
<th>P@K</th>
<th>P(NLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,717</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>75</td>
<td>.99</td>
</tr>
<tr>
<td>.</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
</tr>
<tr>
<td>Nrank</td>
<td>?</td>
<td>?</td>
<td>N</td>
<td>N</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>
From ranking to classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Patient ID</th>
<th>NLP score</th>
<th>Case label</th>
<th>K</th>
<th>P@K</th>
<th>P(NLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>4,717</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>75</td>
<td>.99</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>?</td>
<td>6</td>
<td>6</td>
<td>?</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>?</td>
<td>.</td>
</tr>
<tr>
<td>Nrank</td>
<td>.</td>
<td>?</td>
<td>N</td>
<td>N</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

u ~ Uniform(0, 1)
From ranking to classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Patient ID</th>
<th>NLP score</th>
<th>Case label</th>
<th>K</th>
<th>P@K</th>
<th>P(NLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,717</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>75</td>
<td>.99</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>?</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>Nrank</td>
<td>.</td>
<td>?</td>
<td>N</td>
<td>?</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic labeling of cases

$P(NLP)$
Probabilistic labeling of cases

P(NLP) + gold labels
Probabilistic labeling of cases

P(NLP)

P(NLP) + gold labels

P(NLP+ICD) + gold labels
Classification of suicide phenotypes

AUPRC improvement based on negation detection:

- Suicidal ideation: 2.3% (NLP), 3.7% (NLP+ICD)
- Suicide attempt: 0.7% (NLP), 1.2% (NLP+ICD)

NLP vs. NLP+ICD
Classification of suicide phenotypes

AUPRC improvement based on negation detection:
- Suicidal ideation: 2.3% (NLP), 3.7% (NLP+ICD)
- Suicide attempt: 0.7% (NLP), 1.2% (NLP+ICD)

NLP vs. NLP+ICD
From prevalence to incidence

Phenotype: suicide attempt
Retrieval: “day of notes”
Output: <patient, day>
Weighted sampling of charts
Double chart review

Results:
• 263,403 <patient, day> retrieved
• 3,566 reviewed charts
• AUPRC range: 0.88-0.92
• Good inter-rater agreement (K=.89)
Conclusions

- **Scalable** NLP system for extracting low-prevalence (under-coded and under-reported) phenotypes from EHR
- Proved the **generalizability** of the method over multiple phenotypes
- Showed **replication** of results across two EHR repositories
- **Data-driven generation** of phenotype profiles leveraging unsupervised learning
- Extraction of phenotype cases with **high precision**
- **Diagnostic coding and NLP** yield optimal ascertainment
- Demonstrated the feasibility of the method for identifying **incidents of suicide attempt**
Acknowledgements

VUMC:
- Katelyn Robinson
- Michael Ripperger
- Drew Wilimitis
- Ryan Ahmed
- JooEun Kang
- Theodore J. Morley
- Jhansi Kolli
- John (Jack) Angiolillo
- Douglas Conway
- Robertson Nash
- Jana K. Shirey-Rice
- Loren Lipworth
- Gabriella Papa
- Robert M. Cronin
- Jill Pulley
- Sunil Kripalani
- William Stead
- Shari Barkin
- Kevin B. Johnson
- Joshua C. Denny

VUMC (cont.):
- Michael Matheny
- Aileen Wright
- Qingxia Chen
- Daniel Fabbri
- Douglas M. Ruderfer
- Colin G. Walsh

OHSU:
- David Dorr
- Ana Quinones
- Christie Pizzimenti
- Sumeet Singh
- Matt Storer

Harvard:
- Sruthi Adimadhyam
- Shamika More
- Adee Kennedy

FDA:
- Andy D. Mosholder
- Sai H. Dharmarajan
- Danijela Stojanovic

KPWA:
- David Carrell

Funding:
- U54 MD010722
- R01 LM010685
- R01 GM103859
- UL1 TR000445
- CDRN-1306-04869
- R01 MH121455
- R01 MH116269
- R01 MH118233
- FDA
Questions?
<table>
<thead>
<tr>
<th>Priorities</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data infrastructure</strong></td>
<td>Identification and queries of potential EHR data partners (Horizon Scan: DI1)</td>
<td>Adding unstructured data and necessary data elements (DI2)</td>
<td>Source data mapping (DI3)</td>
<td>Onboarding EHR data partners</td>
<td>Updating CDM to include EHR data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data quality metrics and quality assurance strategy</td>
<td>Harmonizing EHRs (DI4)</td>
<td></td>
<td>Data governance process</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FHIR preparedness (DI7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Feature engineering</strong></td>
<td></td>
<td>Computable phenotyping framework (FE1)</td>
<td>NLP tools for cohort identification, exposure assessment, covariate ascertainment (Scalable NLP: FE2)</td>
<td>Increasing automation in computable phenotyping</td>
<td>Enhancing transportability of phenotypes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Improving probabilistic phenotyping of incident outcomes (FE3)</td>
<td></td>
<td>NLP tool prototyping and expansion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Expanding phenotyping for incident outcomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Developing NLP-assisted chart abstraction tool (FE4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Causal inference</strong></td>
<td>Evaluating targeted learning in EHR data (Enhancing CI: CI1)</td>
<td>Targeted learning tool development</td>
<td>Performance metrics (CI5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Causal inference framework (CI2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Approaches for missing data (CI3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Detection analytics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Innovation incubator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Enhancing Causal Inference in the Sentinel System

Leveraging unstructured electronic health records for large-scale confounding control in real-world evidence studies

Richard Wyss, PhD, MSc
Background
• Confounding arising from non-randomized treatment choices remains a fundamental challenge for extracting valid evidence to help guide treatment and regulatory decisions.

• Standard tools for confounding adjustment have typically relied on adjusting for a limited number of investigator specified variables.
  • Adjusting for investigator-specified variables alone is often inadequate
    - Some confounders are unknown at the time of drug approval
    - Many confounders are not directly measured in routine-care databases.
Background: Proxy Confounder Adjustment

- Healthcare databases may be understood and analyzed as a high-dimensional set of “proxy” factors that indirectly describe the health status of patients (Schneeweiss 2009, 2017).

<table>
<thead>
<tr>
<th>Unobserved confounder</th>
<th>Observable proxy measurement</th>
<th>Coding examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very frail health</td>
<td>Use of oxygen canister</td>
<td>CPT-4</td>
</tr>
<tr>
<td>Sick but not critical</td>
<td>Code for hypertension during a hospital stay</td>
<td>ICD-9, ICD-10</td>
</tr>
<tr>
<td>Health-seeking behavior</td>
<td>Regular check-up visit; regular screening examinations</td>
<td>ICD-9, CPT-4, #PCP visits</td>
</tr>
</tbody>
</table>
Background: High-Dimensional Proxy Confounder Adjustment

- How to identify/generate proxy variables for adjustment?
  - High-dimensional propensity score (Schneeweiss 2009)
    - Does not require data pre-processing
  - OMOP approach:
    - Pre-process data into a common data model then use machine learning algorithms for variable selection (e.g., Lasso)
- Current approaches for generating proxy variables for confounder adjustment do not leverage information from unstructured EHR text notes.
Background: Leveraging Unstructured Electronic Health Records for Large-Scale Proxy Adjustment.

• NLP tools turn free-text notes from EHR data into structured features that can supplement confounding adjustment.
  • However, traditional applications are difficult to scale for large-scale proxy adjustment.

• **Project Objective 3 (use of NLP-generated information from unstructured data):** To explore if unsupervised NLP can be used to generate high-dimensional sets of features from free-text notes for improved large-scale proxy confounding control
  • **Aim 1:** To use scalable applications of NLP to generate structured features from high-dimensional data for large-scale proxy adjustment.
    - leverages work from RO1 (Josh Lin, PI; Richie Wyss, Co-PI; Sebastian Schneeweiss, Co-PI)
  • **Aim 2:** To better understand what machine learning tools for confounder selection perform well for large-scale proxy adjustment in ultra high-dimensional RWE studies.
Methods
Methods: Data Source for Generating Cohort Studies

- Mass General Brigham (MGB) Research Patient Data Registry (RPDR)
  - The electronic health records (EHR) of all the patients aged 65 and above identified in the Mass General Brigham (MGB) Research Patient Data Registry (RPDR) were linked to Medicare claims data.

- Linked RPDR-Medicare claims were used to generate 3 cohort studies comparing different classes of medications (details on later slide).
  - Purpose: case studies for evaluating and testing various methods for NLP feature generation for ultra high-dimensional proxy confounder adjustment.
Methods: Using NLP to Generate Structured Features.

• We used ‘bag-of-words’ to generate features for the top 20,000 most prevalent terms from free-text notes.
  • Very common, simple, and flexible NLP approach
  • Measures the frequency (occurrence) of words within a document
    - Order and structure of words in the document is discarded.
    - The model is only concerned with whether words occur in the document, not where in the document or in relation to other words

• Each word count is then a feature that can be used for modeling
## Methods: Study Cohorts

### Table 1. Study Cohorts

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Total N</th>
<th>Treatment (%)</th>
<th>Outcome (%)</th>
<th>Investigator Specified</th>
<th>Claims Codes</th>
<th>EHR features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>High vs low intensity statin with an outcome of major cardiac events</td>
<td>3,529</td>
<td>1,244 (35.3)</td>
<td>138 (3.9)</td>
<td>39</td>
<td>18,409</td>
<td>20,017</td>
</tr>
<tr>
<td>2.</td>
<td>Oral anti-coagulants vs non-use with an outcome of stroke and major bleeding</td>
<td>9,571</td>
<td>5,991 (62.6)</td>
<td>158 (1.7)</td>
<td>39</td>
<td>19,517</td>
<td>20,051</td>
</tr>
<tr>
<td>3.</td>
<td>High vs. low dose PPI with an outcome of peptic ulcer complications</td>
<td>20,862</td>
<td>7,108 (34.1)</td>
<td>234 (1.1)</td>
<td>39</td>
<td>28,041</td>
<td>20,025</td>
</tr>
</tbody>
</table>
Methods: How to best identify confounder information in ultra high-dimensional real-world data?

• Predictive performance did not improve when modeling the outcome, but does this mean that there is no additional confounder information in EHR generated variables?

• Begin by considering various methods for confounder selection
  • Focus on lasso-based approaches
    • Regular Lasso
    • Outcome adaptive lasso
    • Collaborative controlled lasso
    • Outcome highly-adaptive lasso
Methods: How to make objective decisions on which modeling approach is best?

- Cannot use actual study with estimated effects to make modeling decisions
- Recent papers have proposed using synthetic control studies to help assess validity of alternative causal inference models and tailor analyses to the given study (Alaa & Van Der Scharr 2019; Schuler et al. 2017; Athey S et al. 2019; Bahamyirou A., et al. 2018; Schuemie MJ, et al. 2018; Petersen et al. 2012)
  - Provides an objective assessment of validity and model selection.
  - A common theme is that they use a variation of ‘plasmode simulation’ (Franklin et al. 2014).

<table>
<thead>
<tr>
<th>Variation of the parametric bootstrap where we bootstrap from the original study population, but simulate some aspects of the data structure while leaving other features of the data unchanged.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typically, we set the outcome data aside (outcome blind data), then simulate the outcome while leaving baseline covariates and treatment status unchanged.</td>
</tr>
<tr>
<td>Try to generate synthetic control outcomes (and treatment) that mimic as closely as possible the observed confounding structure in the study cohort.</td>
</tr>
<tr>
<td>Will be inexact, but close approximations can be useful for testing robustness and validity of causal inference methods for the study at hand.</td>
</tr>
</tbody>
</table>
### Confounder Selection & Propensity Score Models

<table>
<thead>
<tr>
<th>Lasso PS Models</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard Lasso</strong></td>
<td>Lasso modeling treatment assignment with penalty factor (lambda) that optimizes CV treatment prediction</td>
</tr>
<tr>
<td><strong>CTMLE Lasso w/ predictions</strong></td>
<td>Collaborative controlled lasso—Lasso modeling treatment assignment but uses ctmle to choose penalty factor. We include initial predictions for the counterfactual outcomes using an outcome lasso model.</td>
</tr>
<tr>
<td><strong>CTMLE Lasso w/ no predictions</strong></td>
<td>Collaborative controlled lasso—Lasso modeling treatment assignment but uses ctmle to choose penalty factor. We did not include initial predictions for the counterfactual outcomes (only included treatment in the initial outcome model).</td>
</tr>
<tr>
<td><strong>Outcome Adaptive Lasso (OAL)</strong></td>
<td>Adaptive lasso modeling treatment assignment with a penalty factor set by user. We assigned a penalty of 0 for all variables selected by the outcome lasso and a penalty of 1 for all other variables (i.e., we forced variables selected by outcome lasso into the lasso model for treatment).</td>
</tr>
<tr>
<td><strong>CTMLE OAL w/ predictions</strong></td>
<td>Collaborative controlled outcome adaptive lasso with initial predictions for the counterfactual outcomes</td>
</tr>
<tr>
<td><strong>CTMLE OAL w/ no predictions</strong></td>
<td>Collaborative controlled outcome adaptive lasso with no initial predictions for the counterfactual outcomes (initial outcome model includes only treatment)</td>
</tr>
</tbody>
</table>

- For each PS model, we estimated the treatment effect using Targeted Maximum Likelihood Estimation (TMLE) that included initial predictions from an outcome lasso model and PS weighting.
Simulation Results
Selected Simulation Results for Prediction
Same Sample

Cross Validated

- Method 1: Uses PS selected by Lasso with optimizing CV prediction
- Method 2: Uses PS selected by CTMLE Lasso with initial outcome predictions
- Method 3: Uses PS selected by CTMLE Lasso with no initial outcome predictions
- Method 4: Uses PS selected by Adaptive Lasso optimizing CV prediction
- Method 5: Uses PS selected by CTMLE Adaptive Lasso with initial outcome predictions
- Method 6: Uses PS selected by CTMLE Adaptive Lasso without initial outcome predictions
Selected Simulation Results for Bias
Lambda Selection for Lasso PS Model

- Unadjusted
- PS Model 1: Traditional Lasso
- PS Model 2: CTMLE Lasso with predictions
- PS Model 3: CTMLE Lasso no predictions
- PS Model 4: Outcome Adaptive Lasso (OAL)
- PS Model 5: CTMLE OAL with predictions
- PS Model 6: CTMLE OAL no predictions
- Oracle: includes all confounders
General points for discussion

• Selecting models based on collaborative learning improved bias reduction even though predictive performance declined.
  - Outcome adaptive lasso with collaborative selection generally performed best.
  - Some degree of overfitting is beneficial for confounding control when using Machine Learning to data-adaptively select (model) high-dimensional sets of variables

• Bias increased as the number of spurious variables available for selection increased.
• Bias can result from two sources
  1. Lasso model not selecting confounding variables
  2. Even when lasso selects confounders there can still be regularization bias (Chernozhukov 2018).

• Use relaxed lasso to reduce regularization bias in sparse high-dimensional data (Meinshausen 2007).
Relaxed lasso

Use relaxed lasso to reduce regularization bias (Meinshausen 2007).

- Runs regularized regression twice:
  1. First runs lasso to select lambdas to control variable selection (which variables are selected for each lambda);
  2. Second step runs regularized regression again for each set of variables selected by each lambda with less penalization to control shrinkage level of coefficients. The shrinkage penalization in the second step can be selected using Cross Validation.

- ‘Idea of the relaxed lasso is to take the lasso fitted object and then for each lambda, refit the variables in the active set with either no penalization or less penalization. This gives the “relaxed” fit’. (Hastie & Tibshirani 2021)

- Relaxed lasso can often improve predictive performance by fitting more parsimonious models with less penalization in sparse high-dimensional data (Meinshausen 2007).
Selected Simulation Results for Variable Selection and Prediction with Relaxed Lasso
Selected Simulation Results for Bias with Relaxed Lasso
Discussion
General Points for Discussion after running ‘relaxed’ lasso

• Relaxed lasso reduced bias in effect estimate compared with standard lasso

• Selecting models based on collaborative learning still improved bias reduction at the expense of predictive performance.
  • Outcome adaptive lasso with collaborative selection generally performed best.
  • Some degree of overfitting is beneficial for confounding control when using Machine Learning to data-adaptively select (model) high-dimensional sets of variables

• Still some bias with large numbers of variables
  • May need large samples to use ML to identify confounders in sparse high-dimensional data.
Future work/next step is to apply top performing models from simulations to empirical studies
Research team

**Food and Drug Administration**
- Hana Lee, PhD, MS
- Sarah Dutcher, PhD, MS

**Brigham and Women’s Hospital**
- Richard Wyss, PhD, MS
- Rishi Desai, PharmD, PhD
- Josh Lin, MD, PhD
- Shirley Wang, PhD, MS
- Yinzhu Jin, MS, MPH
- Shamika More, MS
- Luke Zabotka, BA

**DPM/HPHCI/Operations Center**
- Darren Toh, ScD

**Kaiser Washington/University of Washington**
- Jennifer Nelson, PhD

**University of California, Berkeley**
- Mark van der Laan, PhD
- Lars van der Laan

**University of Michigan**
- Xu Shi, PhD

**Putnam Data Sciences**
- Susan Gruber, PhD, MS, MPH
Questions?
<table>
<thead>
<tr>
<th>Priorities</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data infrastructure</strong></td>
<td>Identification and queries of potential EHR data partners (Horizon Scan: DI1)</td>
<td>Adding unstructured data and necessary data elements (DI2)</td>
<td>Source data mapping (DI3)</td>
<td>Data quality metrics and quality assurance strategy</td>
<td>Onboarding EHR data partners</td>
</tr>
<tr>
<td></td>
<td>Master plan</td>
<td>Master plan refinement</td>
<td>Data harmonization strategy</td>
<td>FHIR preparedness (DI7)</td>
<td>Harmonizing EHRs (DI4)</td>
</tr>
<tr>
<td><strong>Feature engineering</strong></td>
<td>Calculable phenotyping framework (FE1)</td>
<td>Improving probabilistic phenotyping of incident outcomes (FE3)</td>
<td>Developing NLP-assisted chart abstraction tool (FE4)</td>
<td>Expanding phenotyping for incident outcomes</td>
<td>Calibration methods (CI4)</td>
</tr>
<tr>
<td></td>
<td>Causal inference framework (CI2)</td>
<td>Increasing automation in computable phenotyping</td>
<td>Implementing NLP-assisted chart abstraction tool</td>
<td>Performance metrics (CI5)</td>
<td>Approaches for missing data (CI3)</td>
</tr>
<tr>
<td><strong>Causal inference</strong></td>
<td>Performance metrics (CI5)</td>
<td>Increasing automation in computable phenotyping</td>
<td>Implementing NLP-assisted chart abstraction tool</td>
<td>Performance metrics (CI5)</td>
<td>Approaches for missing data (CI3)</td>
</tr>
<tr>
<td><strong>Detection analytics</strong></td>
<td>Evaluating targeted learning in EHR data (Enhancing CI: CI1)</td>
<td>Targeted learning tool development</td>
<td>Performance metrics (CI5)</td>
<td>Approaches for missing data (CI3)</td>
<td>Distributed regression implementation (CI6)</td>
</tr>
<tr>
<td></td>
<td>Identification and evaluation of EHR detection approaches (DA1)</td>
<td>Empirical evaluation of EHR-based detection approaches (DA2)</td>
<td>Development of EHR-based detection tools</td>
<td>Development of EHR-based signal detection</td>
<td>Methods framework for EHR-based signal detection</td>
</tr>
<tr>
<td><strong>Innovation incubator</strong></td>
<td>Data Sandbox Discovery Phase</td>
<td>Data Sandbox Implementation Phase</td>
<td>Data Sandbox Discovery Phase</td>
<td>Data Sandbox Implementation Phase</td>
<td>Data Sandbox Discovery Phase</td>
</tr>
</tbody>
</table>

**CI2**: A causal inference framework for Sentinel
A Causal Inference Framework for Sentinel

Rishi J Desai, PhD,
Assistant Professor, Division of Pharmacoepidemiology and Pharmacoeconomics,
Brigham and Women’s Hospital, Harvard Medical School,
Boston, MA
Background and motivation
Why do we need another framework?

**Quality assessment tools**

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

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

**Reporting tools**

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

Sinead M. Langan, 1 Spinus AJ Schmidt, 1 Kevin Wong, 1 Vera Ehrenstein, 1 Stuart G Nicholl, 1, 3 Kristian B Filbin, 1, 3 Olaf Klingel, 1 Irene Petersen, 1 Henrik T Sorensen, 1, 3 William G Davis, 1, 3 Arndt Gutmann, 1, 3 Karl Hamer, 1, 3 Lars G Hemkens, 1, 3 David Mohr, 1, 3 Sebastian Schneeweiss, 1, 3 Liam Smith, 1, 3 Miriam Sterkenboom, 1, 3 Erik von Elm, 1, 3 Shirley V Wong, 1, 3 Eric I Benchimol, 1, 3

**Best practices**

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

Guidance for Industry

**Misc:** Highly specific or focusing on parts of the process

The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology (Revision 9)
Why do we need another framework?

What do we have?

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

What don’t we have?

- None of these tools offer a general framework to guide decision making at various steps along the way

Vision for a framework to guide principled investigations using non-randomized, secondary data

- The Sentinel Innovation Center is developing a causal inference framework proposing a stepwise process that systematically considers key choices with respect to design and analysis that influence the validity of studies conducted with non-randomized, secondary data
- A standardized “industrial” process that will be outlined in this framework will serve as a guide to inform the conduct of non-randomized secondary database studies of drug-outcome evaluation
- Key considerations to meet the FDA need of informing regulatory decision making based on such investigations
  - Limit variations across investigators by outlining a general process
  - Focus on repeatability of the process
  - Written and endorsed by independent experts
A draft of the proposed framework
Step 1: Well defined research question in the target trial framework specifying PICOTS

- First and non-negotiable step in any framework that intends to generate causal inference from observed data
- Target trial framework, which is conceptualized as envisioning a hypothetical prospective randomized controlled trial, provides a useful and practical device to sharply define a causal question of interest
- Explicit identification of the following key study parameters
  - patient population (P)
  - the intervention (I) specifying the medical product under investigation,
  - a comparator group (C)
  - the outcome (O) along with an appropriate time horizon (T)
  - setting (S) where the study is implemented
2a. Is the exposure of interest captured?
- Yes (e.g. prescription medications)
- No (e.g. blood transfusion products)

2b. Is the outcome of interest measured with sufficient validity*?
- Yes (e.g. serious events like AMI)
- No (e.g. pancreatitis)

2c. Is the population identifier measured with sufficient validity*?
- Yes (e.g. diagnosis of indications, important comorbid illnesses)
- No (e.g. HF with EF class)

2d. Are key confounders identifiers measured?
- Yes (e.g. diagnosis of indications, important comorbid illnesses)
- No (e.g. renal function based on laboratory test results)

* Validity as demonstrated by parameters including PPV, sensitivity, specificity for binary outcomes; proportion missing for continuous outcomes; and accurate onset for time to event outcomes and availability of long-term follow-up data for latent outcomes.
**Step 3: Refining target trial parameters\(^1\) and translate to RWE study parameters**

(Using a hypothetical example case study of SGLT2 inhibitors and the risk of genital infection in a claims-EHR linked data source)

<table>
<thead>
<tr>
<th>Element</th>
<th>Ideal trial</th>
<th>RWE translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exposure (&quot;treatment strategies&quot;)</strong></td>
<td>Randomly assigned initiation of SGLT2i (canagliflozin, dapagliflozin, empagliflozin) versus a DPP4 inhibitors</td>
<td>First prescription dispensing of SGLT2i (canagliflozin, dapagliflozin, empagliflozin) or DPP4 inhibitors identified based on pharmacy claims</td>
</tr>
<tr>
<td><strong>Eligibility</strong> (assessed at baseline, prior to time 0)</td>
<td>Patients aged 18 years or older, with type 2 diabetes mellitus, and no use of study medications before randomization</td>
<td>Observability related: continuous enrollment for 12 months and &gt;80% mean capture proportion(^2) in EHRs before study medication initiation. Treatment related: No prior use of study medications. Indication related: Diagnosis of type 2 diabetes based on diagnosis codes or HbA1c results. Other: Age 18 or older</td>
</tr>
<tr>
<td><strong>Follow-up start</strong> (Time 0)</td>
<td>At randomization</td>
<td>At prescription dispensing</td>
</tr>
<tr>
<td><strong>Follow-up end</strong></td>
<td>1-year post-randomization unless patients are lost to follow-up or die or have the outcome</td>
<td>Earliest of the outcome, death, insurance disenrollment, or 1-year post initiation</td>
</tr>
<tr>
<td><strong>Primary outcome</strong></td>
<td>Hospitalization for genital infections</td>
<td>Hospitalization for genital infections assessed based on primary discharge diagnosis codes</td>
</tr>
<tr>
<td><strong>Baseline covariates</strong></td>
<td>-</td>
<td>Demographics, diabetes severity related variables including micro and macrovascular complications and laboratory test results such as HbA1c and serum creatinine, comorbid conditions, comediations, markers for healthy behavior and healthcare utilization</td>
</tr>
<tr>
<td><strong>Causal estimand</strong></td>
<td>Intent-to-treat (ITT)</td>
<td>Observational analogue of ITT</td>
</tr>
<tr>
<td><strong>Statistical analysis</strong></td>
<td>A Cox proportional hazards model</td>
<td>Adjustment of baseline confounding with propensity score matching followed by an outcome analysis using a Cox proportional hazards model</td>
</tr>
<tr>
<td><strong>Subgroup analyses</strong></td>
<td>Stratified by gender</td>
<td>Same as ideal trial</td>
</tr>
</tbody>
</table>

---


\(^2\)Lin et al. *Epidemiology* 2018;29: 356–363
**Step 4: Feasibility analysis**

**Cohort feasibility**
- Implement inclusion/exclusion criteria
- Exposure group assignment
- Exposure patterns e.g. average time on treatment

If cohort size deemed feasible

**Outcome feasibility**
- Outcome counts and rates either in the full cohort without stratification by exposure or just in the reference exposure group

Diagnostic evaluations
1. Evaluate distribution of key patient characteristics
2. If using propensity score (PS) based confounding adjustment methods
   a. Construct propensity score model
   b. Evaluate overlap to ensure comparability
   c. Evaluate balance conditional on the PS, update modeling choices until the balance is achieved

Outcome counts
- Diagnostics passed
- Potential issues diagnosed
- Outcome counts insufficient to support reliable causal analysis
- Outcome counts sufficient to support reliable causal analysis

Go back to Step 3, consider design modifications (e.g. relaxing inclusion/exclusion criteria)

Proceed to Step 5

If cohort size deemed not feasible to support the proposed analysis, wait until accumulation of additional data
Step 5: Pre-specification of robustness evaluations

**Robustness evaluations**

- **Deterministic sensitivity analyses**
  - Varying design assumptions, variable measurement methods, or analytic choices

- **Quantitative bias analyses**
  - For exposure/outcome misclassification
    - Probabilistic sensitivity analysis¹
  - For unmeasured confounding
    - E.g. Array or rule out methods²

- **Trial calibration***
  - Duplicating inclusion/exclusion criteria and all design aspects of the trial to evaluate whether primary outcome is replicable in the data source³
  - *such trial may not always exist

- **Net bias analysis**
  - Control/tracer analysis

- **Control analysis**
  - Negative control exposure/outcome⁴

---

³ Khosrow-Khavar et al. Annals Rheum Dis. 2022
Summary and next steps

• Continuing to fine tune the framework steps
• Conducting a demonstration project to highlight how decisions are made at each step along the way and walk users through the steps based on a realistic case-example
• The goal is dissemination of this framework in peer-reviewed publication by early next year
Questions?
Closing remarks

• Through initiatives such as those discussed today, Sentinel Innovation Center is making strides in helping to achieve the FDA’s vision of a Medical Data Enterprise with a query-ready system containing >10 million EHR lives

• Key research needs have been identified and ongoing research projects are addressing some salient challenges presented by EHRs in 4 key domains
  • Data infrastructure
  • Feature engineering
  • Causal inference
  • Detection analytics

• Highly interdisciplinary research work being conducted at the Innovation Center involving experts in the fields of epidemiology, informatics, medicine, and statistics, will generate unique insights regarding meaningful use of EHRs for clinical research and provide practical solutions
Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework

Rishi J. Desai, Michael E. Matheny, Kevin Johnson, Keith Marsolo, Lesley H. Curtis, Jennifer C. Nelson, Patrick J. Heagerty, Judith Maro, Jeffery Brown, Senghee Toh, Michael Nguyen, Robert Ball, Gerald Dal Pan, Shirley V. Wang, Joshua J. Gagne and Sebastian Schneeweiss

The Sentinel System is a major component of the United States Food and Drug Administration’s (FDA) approach to active medical product safety surveillance. While Sentinel has historically relied on large quantities of health insurance claims data, leveraging longitudinal electronic health records (EHRs) that contain more detailed clinical information, as structured and unstructured features, may address some of the current gaps in capabilities. We identify key challenges when using EHR data to investigate medical product safety in a scalable and accelerated way, outline potential solutions, and describe the Sentinel Innovation Center’s initiatives to put solutions into practice by expanding and strengthening the existing system with a query-ready, large-scale data infrastructure of linked EHR and claims data. We describe our initiatives in four strategic priority areas: (1) data infrastructure, (2) feature engineering, (3) causal inference, and (4) detection analytics, with the goal of incorporating emerging data science innovations to maximize the utility of EHR data for medical product safety surveillance.

npj Digital Medicine (2021)4:170; https://doi.org/10.1038/s41746-021-00542-0
Innovation Center collaborating organizations

Lead sites:

- Kaiser Permanente
- Vanderbilt University Medical Center
- Brigham and Women's Hospital
- University of Washington
- HealthCore
- Healthagen
- Aetna
- Optum
- Veradigm
- Northeastern University
- Health Sciences South Carolina
- Aetion
- Epic
- patientslikeme
- Concerto HealthAi
- University of Florida
- Penn State University
- Rutgers University
- Stanford University
- HealthPartners Institute
- UC Berkeley
- UC Illinois
- UC Michigan
- AMIA
- DOC.ai
- HCA Healthcare
- Oak Ridge National Laboratory
- OHDSI
- Medical University of South Carolina
- Columbia University
- Putnam Data Sciences, LLC
- Department of Population Health Sciences, Duke University School of Medicine
- Microsoft
- Microsoft Research
- MDxHealth
- UAB
- Robert J. Margolis, MD Center for Health Policy
Thank you