Welcome to the Sentinel Innovation Center Webinar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today’s webinar will be available on the recording posted at the link above.
Harmonizing Electronic Health Records from Heterogeneous Systems via Automated Translation of Medical Concepts

Xu Shi
Department of Biostatistics, University of Michigan

August 19, 2020
Demand for health information exchange

• **Goodbye “meaningful use”, hello “promoting interoperability”**
  - Centers for Medicare & Medicaid Services (CMS) renamed EHR incentive program
  - To advance integration and sharing of healthcare data

---

**BREAKING: CMS Finalizes “Promoting Interoperability” Rule**

August 2, 2018 by Rajiv Leventhal

The federal agency has finalized 90-day reporting periods for 2019 and 2020, while requiring 2015 CEHRT starting in 2019

Just three months after issuing a proposal, the Centers for Medicare & Medicaid Services (CMS) has finalized a rule late this afternoon that will overhaul the meaningful use program with a core emphasis on advancing health data exchange among providers.

The final rule issued today makes updates to Medicare payment policies and rates under the Inpatient Prospective Payment System (IPPS) and the Long-Term Care Hospital (LTCH) Prospective Payment System (PPS) that will incentivize value-based, quality care at these facilities.

CMS Administrator Seema Verma said in a statement, “We’ve listened to patients and their doctors who urged us to remove the obstacles getting in the way of quality care and positive health outcomes. Today’s final rule reflects public feedback on CMS proposals issued in April, and the agency’s patient-driven priorities of improving the quality and safety of care, advancing health information exchange and usability, and removing outdated or redundant regulations on healthcare providers to make way for innovation and greater value.”

According to CMS, the rule applies to about 3,300 acute care hospitals and 420 long-term care hospitals, and will take effect Oct. 1

Semantic interoperability: EHRs do not talk to each other

- 786.2 Cough (ICD-9)
- 780.61 Fever (ICD-9)
- 71010 Chest X-ray (CPT)

<table>
<thead>
<tr>
<th>lab_type</th>
<th>result</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEMOGRAM PLATELET COUNT</td>
<td>154</td>
<td>x10 3/uL</td>
</tr>
<tr>
<td>HEMOGRAM RED BLOOD CELL COUNT</td>
<td>4.59</td>
<td>x10 6/uL</td>
</tr>
<tr>
<td>HEMOGRAM RED CELL DISTRIBUTION WIDTH</td>
<td>20.4</td>
<td>%</td>
</tr>
<tr>
<td>HEMOGRAM WHITE BLOOD CELL COUNT</td>
<td>6.6</td>
<td>x10 3/uL</td>
</tr>
</tbody>
</table>

--- PHYSICIAN NOTE ---

History of Present Illness

Presenting problem started 5 days ago. History comes from patient. Able to get a good history. Presents with symptoms suggestive of a lower GI bleed. This is a new problem, with no prior history of similar episodes. Symptoms developed over several days. Describes stool as black in color. Passing mucoid stools. Streaks of blood noted in stool. Saw gross blood in the bowel movements. Not on iron or Pepto bismol. Estimated blood loss is less than 50 cc. No history of prior GI bleeding. No history orthostatic symptoms, excessive fatigue, or syncope.

• Standardized medical code for billing 😊
  - Common language across healthcare providers and insurers

• Inconsistent coding in practice 😞
  - System A use 786.05: shortness of breath
  - System B use 786.09: other dyspnea and respiratory abnormality
What are potential challenges?

System A

System B
What are potential challenges?
What are potential challenges?

Common Data Model

System A

System A

Phenotyping Algorithm
Or Causal Inference

786.05: shortness of breath

Anaphylaxis

System B

System B
What are potential challenges?

Common Data Model

System A

System A

Phenotyping Algorithm

Or Causal Inference

786.05: shortness of breath

Transport

System B

System B

Anaphylaxis

Anaphylaxis

5
What are potential challenges?

System A

Common Data Model

System A

786.05: shortness of breath

System B

System B

786.09: dyspnea and respiratory abnormality

Phenotyping Algorithm Or Causal Inference

Anaphylaxis

Transport
What are potential challenges?

- Performance of phenotyping algorithm can dramatically drop
- Causal inference can fail due to incorrect confounding adjustment
Manual mapping is imprecise

- **General equivalence mapping (GEM): ICD-9 (10k) ⇔ ICD-10 (60k)**
  - Approximate mappings with multiple scenarios: data merged with adhoc decisions
Data Driven Mapping of Medical Codes
The Back pain Outcomes using Longitudinal Data study

- **The elderly with back pain**
  - 5000 patients ≥ 65 years old
  - Cost-effectiveness of early diagnostic imaging

- **EHR data from three sites:**
  - Henry Ford Health System in Detroit
  - Kaiser Permanente Northern California
  - Harvard Vanguard in Boston

______________________________
Data quality check before pulling EHR data from study sites

- **Compare use of CPT codes between study sites**

<table>
<thead>
<tr>
<th>Procedure</th>
<th>% Total Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRI</td>
<td>Henry Ford: 9.4</td>
</tr>
<tr>
<td></td>
<td>Kaiser Permanente: 3.6</td>
</tr>
<tr>
<td>CT scan</td>
<td>Henry Ford: 14.5</td>
</tr>
<tr>
<td></td>
<td>Kaiser Permanente: 4.2</td>
</tr>
<tr>
<td>X-ray</td>
<td>Henry Ford: 22.3</td>
</tr>
<tr>
<td></td>
<td>Kaiser Permanente: 5.3</td>
</tr>
<tr>
<td>Physical Therapy</td>
<td>Henry Ford: 28.9</td>
</tr>
</tbody>
</table>

**Question:** can we scan for variation in the endorsement of all medical codes to identify such data quality issue?
Detect and quantify coding differences under a hierarchical structure

- **Code grouping** e.g. PheWAS (phenome-wide association studies)
  CCS (Clinical Classifications Software)

**Testing:**
Group-wise association test

codes in a group ↔ genetic variants in a region

**Estimation:**
Hierarchical shrinkage
Post-regularization inference

CPT-SCAN: [https://xu-rita-shi.shinyapps.io/CPT_SCAN/](https://xu-rita-shi.shinyapps.io/CPT_SCAN/)

Shi et al. (2017)
Further investigation into observed differences in code endorsement

- Compare use of CPT codes between study sites

![Bar chart showing the use of CPT codes between Henry Ford and Kaiser Permanente.]

- Henry Ford uses a generic code “HF0PT” for physical therapy
Can data tell me “HF0PT” = “physical therapy”?

• **Co-occurrence: semantic information from the context**
  - “HF0PT” is surrounded by codes for pain-related diseases or treatments
  - “Physical therapy” often appears in such a context
Computers learn the meaning of a word from its context

- **word2vec**: represent a word as a vector
  - Learn semantic relationship from co-occurrence
  - Words with similar meanings are close

- **code2vec**: represent a code as a vector
  - Code $\leftrightarrow$ word; Healthcare system $\leftrightarrow$ language
  - Interpret meaning of codes in clinical practice setting

Mikolov et. al. (2013), Levy & Goldberg (2014), Choi et. al. (2016), Beam et. al. (2018)
Computers learn the meaning of a word from its context

- **word2vec**: represent a word as a vector
  - Learn semantic relationship from co-occurrence
  - Words with similar meanings are close

- **code2vec**: represent a code as a vector
  - Code ⇔ word; Healthcare system ⇔ language
  - Interpret meaning of codes in clinical practice setting

**Question**: can we infer a mapping between two sets of code-vectors learned from two healthcare systems, respectively?

Mikolov et. al. (2013), Levy & Goldberg (2014), Choi et. al. (2016), Beam et. al. (2018)
From language translation to code mapping

• Inconsistent objectives in language translation with word2vec

---

Xing et. al. (2015), Conneau et. al. (2018), Shi et. al. (2018)
From language translation to code mapping

- **Inconsistent objectives in language translation with word2vec**
  - Generate word vectors: \( \max \text{ inner product} \)
  - Space alignment: \( \min \ell_2 \text{ distance} \)
  - Infer a mapping: \( \max \text{ cosine} \)

- **Length normalization: semantic information is in the direction**

Xing et. al. (2015), Conneau et. al. (2018), Shi et. al. (2018)
How do statisticians think about language translation?

\[ \mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^T_{n \times p}, \quad \mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]^T_{n \times p}: n \text{ vectors, each } \mathbf{X}_i, \mathbf{Y}_i \in \mathbb{R}^p \]

- \( n \): number of codes
- \( p \): dimension of code-vectors
How do statisticians think about language translation?

- **Classical regression**

\[
Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p}
\]

\(Y_i \sim X_i\) correctly linked
How do statisticians think about language translation?

- **Classical regression**
  \[ Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ correctly linked} \]

- **Shuffled regression**
  \[ Y_{n \times p} = \Pi_{n \times n} X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ may not correspond} \]

Introduce a mapping matrix \( \Pi \) (the “dictionary”)
no mismatch if \( \Pi = I \) is an identity matrix
How do statisticians think about language translation?

- **Classical regression**
  \[
  Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p}
  \]
  \[
  Y_i \sim X_i \text{ correctly linked}
  \]

- **Shuffled regression**
  \[
  Y_{n \times p} = \Pi_{n \times n} X_{n \times p} W_{p \times p} + U_{n \times p}
  \]
  \[
  Y_i \sim X_i \text{ may not correspond}
  \]

Each row of \( \Pi \) is like a pointer:
- match: \( \Pi_i = \mathbb{I}_i \) \( \Rightarrow \) \( Y_i \sim X_i \)
- mismatch: \( \Pi_i = \mathbb{I}_j \) \( \Rightarrow \) \( Y_i \sim X_j \)
How do statisticians think about language translation?

- **Classical regression**
  \[ Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ correctly linked} \]

- **Shuffled regression**
  \[ Y_{n \times p} = \Pi_{n \times n} X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ may not correspond} \]

Each row of \( \Pi \) is like a pointer:
- **match**: \( \Pi_i. = \mathbb{I}_i. \Rightarrow Y_i \sim X_i \);
- **mismatch**: \( \Pi_i. = \mathbb{I}_j. \Rightarrow Y_i \sim X_j \)
How do statisticians think about language translation?

- **Classical regression**
  \[
  \mathbf{Y}_{n \times p} = \mathbf{X}_{n \times p} \mathbf{W}_{p \times p} + \mathbf{U}_{n \times p}
  \]
  \[\mathbf{Y}_i \sim \mathbf{X}_i\] correctly linked

- **Shuffled regression**
  \[
  \mathbf{Y}_{n \times p} = \Pi_{n \times n} \mathbf{X}_{n \times p} \mathbf{W}_{p \times p} + \mathbf{U}_{n \times p}
  \]
  \[\mathbf{Y}_i \sim \mathbf{X}_i\] may not correspond

Each row of \(\Pi\) is like a pointer:

- **match**: \(\Pi_i = \mathbb{I}_i\) \(\Rightarrow\) \(\mathbf{Y}_i \sim \mathbf{X}_i\)
- **mismatch**: \(\Pi_i = \mathbb{I}_j\) \(\Rightarrow\) \(\mathbf{Y}_i \sim \mathbf{X}_j\)
How do statisticians think about language translation?

• Classical regression
  \[ Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ correctly linked} \]

• Shuffled regression
  \[ Y_{n \times p} = \Pi_{n \times n} X_{n \times p} W_{p \times p} + U_{n \times p} \]
  \[ Y_i \sim X_i \text{ may not correspond} \]

Each row of \( \Pi \) is like a pointer:

match: \( \Pi_i \cdot = I_i \cdot \Rightarrow Y_i \sim X_i \);

mismatch: \( \Pi_i \cdot = I_j \cdot \Rightarrow Y_i \sim X_j \)
How do statisticians think about language translation?

- **Classical regression**
  \[
  Y_{n \times p} = X_{n \times p} W_{p \times p} + U_{n \times p}
  \]
  \(Y_i \sim X_i\) correctly linked

- **Shuffled regression**
  \[
  Y_{n \times p} = \Pi_{n \times n} X_{n \times p} W_{p \times p} + U_{n \times p}
  \]
  \(Y_i \sim X_i\) may not correspond

Each row of \(\Pi\) is like a pointer:
- match: \(\Pi_i \cdot = \Pi_i \cdot \Rightarrow Y_i \sim X_i\);
- mismatch: \(\Pi_i \cdot = \Pi_j \cdot \Rightarrow Y_i \sim X_j\)
Formulating the problem: mismatched spherical data

- $\Pi$ encodes 1-to-1 and 1-to-many mapping

Allow for 1-to-many mapping

weight vector: $\Pi_i = \omega$
The statistical problem: mismatched spherical data

- $\Pi$ encodes 1-to-1 and 1-to-many mapping
- Assume $\Pi$ is block diagonal

Incorporate code-group information
mismatch only occurs within group
The statistical problem: mismatched spherical data

- $\Pi$ encodes 1-to-1 and 1-to-many mapping
- Assume $\Pi$ is block diagonal
- $W$ is an orthogonal matrix s.t. $\|WX_i\| = \|Y_i\| = 1$

$W$ rotates $X$ on the sphere
Align spherical language spaces
The statistical problem: mismatched spherical data

- $\Pi$ encodes 1-to-1 and 1-to-many mapping
- Assume $\Pi$ is block diagonal
- $W$ is an orthogonal matrix s.t. $\|WX_i\| = \|Y_i\| = 1$

Goal: estimate $(\Pi, W)$ using mismatched spherical data
Initialize $\hat{\Pi}^{[1]} = I$

Align language spaces

$\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4 \\
3 & 2 & 1 & 4 \\
4 & 3 & 2 & 1
\end{array}$

$\begin{array}{cccc}
2 & 1 & 3 & 4 \\
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4 \\
3 & 2 & 1 & 4 \\
4 & 3 & 2 & 1
\end{array}$

$\hat{\Pi}^{[1]}$

Rotate $X\hat{\Pi}^{[1]}$

Estimate mismatch pattern

$\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4 \\
3 & 2 & 1 & 4 \\
4 & 3 & 2 & 1
\end{array}$

$\begin{array}{cccc}
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast
\end{array}$

$\hat{\Pi}$

Use matched rows

$\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4 \\
3 & 2 & 1 & 4 \\
4 & 3 & 2 & 1
\end{array}$

$\begin{array}{cccc}
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast
\end{array}$

$\hat{\Pi}$

Refine $\hat{W}$

$\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & 2 & 3 & 4 \\
2 & 1 & 3 & 4 \\
3 & 2 & 1 & 4 \\
4 & 3 & 2 & 1
\end{array}$

$\begin{array}{cccc}
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast
\end{array}$

$\hat{W}$

iSphereMAP: iterative Spherical regression MAPping
iSphereMAP: iterative Spherical regression MAPping

- **Find rotation via spherical regression**
  \[
  \hat{W}[1] = \text{argmin} \| Y - XW \|_F^2 = UV^T
  \]
  where \( X^TY = UDV^T \)

---

**Initialize** \( \hat{W}^{[1]} = I \)

**Align language spaces**

**Rotate** \( X\hat{W}^{[1]} \)

**Estimate mismatch pattern**

**Refine** \( \hat{W} \)

**Use matched rows**

---

Steps:
1. Initialize \( \hat{W}^{[1]} = I \)
2. Align language spaces
3. Rotate \( X\hat{W}^{[1]} \)
4. Estimate mismatch pattern
5. Refine \( \hat{W} \)
6. Use matched rows
iSphereMAP: iterative Spherical regression MAPping

- Find rotation via spherical regression
  \[
  \hat{W}^{[1]} = \arg\min_{W:WW^T=I_p} \|Y - XW\|_F^2 = UV^T
  \]
  where \(X^TY = UDV^T\)

- Match a code to its nearest neighbor(s)
  \[
  \tilde{\Pi}^k = \arg\min \|\tilde{Y}_k - \tilde{X}_k \Pi^T\|_F^2
  \]
  where \(\tilde{Y}_k = Y^k, \tilde{X}_k = (X^k \hat{W}^{[1]})^T\)
iSphereMAP: iterative Spherical regression MAPping

- **Find rotation via spherical regression**
  \[ \hat{W}^{[1]} = \arg\min_{W:WW^T = I_p} \|Y - XW\|_F^2 = U V^T \]
  where \( X^T Y = U D V^T \)

- **Match a code to its nearest neighbor(s)**
  \[ \tilde{\Pi}^k = \arg\min \|\tilde{Y}_k - \tilde{X}_k \hat{\Pi}\|_F^2 \]
  where \( \tilde{Y}_k = Y_k^T, \tilde{X}_k = (X_k \hat{W}^{[1]})^T \)

- **Refine rotation using matched data**
  \[ \hat{W} = \arg\min_{W:WW^T = I_p} \|Y_{\text{match}} - X_{\text{match}} \hat{W}\|_F^2 \]
Theoretical guarantees

Initialize $\hat{\Pi}^{[1]} = I$

Align language spaces

<table>
<thead>
<tr>
<th>Y</th>
<th>W[1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>W + U</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Rotate $X \hat{W}^{[1]}$

Estimate mismatch pattern

<table>
<thead>
<tr>
<th>Y</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>W + U</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Use matched rows

Refine $\hat{W}$

<table>
<thead>
<tr>
<th>Y</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>W + U</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Theoretical guarantees

- **Is alignment insensitive to mismatch?**

\[
\| \hat{W}^{[1]} - W \|_F = O_p(\text{inherent noise} + \text{mismatch})
\]

Consistency requires sparse mismatch.

---

**Initialize** \( \hat{\Pi}^{[1]} = \mathbb{I} \)

**Align language spaces**

**Rotate** \( \hat{X} \hat{W}^{[1]} \)

**Estimate mismatch pattern**

**Use matched rows**

**Refine** \( \hat{W} \)
Theoretical guarantees

- **Is alignment insensitive to mismatch?**
  \[ \| \hat{\mathbf{W}}^{[1]} - \mathbf{W} \|_F = O_p(\text{inherent noise} + \text{mismatch}) \]
  Consistency requires sparse mismatch

- **Is code mapping correct?**
  Correctly map \( \mathbf{Y}_i \) to \( \mathbf{X}_j \) if one-to-one;
  Consistently estimate the weight if one-to-many
Theoretical guarantees

Initialize $\hat{\Pi}^{[1]} = \mathbb{I}$
Align language spaces

<table>
<thead>
<tr>
<th>Y</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

$\hat{W}^{[1]}$

Rotate $X\hat{W}^{[1]}$
Estimate mismatch pattern

<table>
<thead>
<tr>
<th>Y</th>
<th>II</th>
<th>11</th>
<th>22</th>
<th>33</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>II</td>
<td>22</td>
<td>11</td>
<td>33</td>
<td>44</td>
</tr>
</tbody>
</table>

$\hat{\Pi}$

Use matched rows
Refine $\hat{W}$

<table>
<thead>
<tr>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

$\hat{W}$

Is alignment insensitive to mismatch?

$\|\hat{W}^{[1]} - W\|_F = O_p(\text{inherent noise} + \text{mismatch})$
Consistency requires sparse mismatch

Is code mapping correct?

Correctly map $Y_i$ to $X_j$ if one-to-one;
Consistently estimate the weight if one-to-many

Can we better estimate $\hat{W}$?

$\|\hat{W} - W\|_F = O_p(\text{inherent noise})$
As good as if no mismatch is present
Simulation: iSphereMAP vs Mikolov et. al. 2013 (Google)

**Spherical regression error**
- Google $\hat{W}^{[1]}$
- Google $\hat{W}$
- iSphereMAP $\hat{W}^{[1]}$
- iSphereMAP $\hat{W}$

**Mapping: 1-to-1 match error**
- Google $\hat{\Pi}$
- iSphereMAP $\hat{\Pi}$
Example: ICD-9 code translation between two systems

Code Translation = Space alignment + Mapping

<table>
<thead>
<tr>
<th>Partners HealthCare</th>
<th>Veterans Health Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal chest sounds 786.7</td>
<td>Abnormal chest sounds 786.7</td>
</tr>
<tr>
<td>Painful respiration 786.52</td>
<td>Painful respiration 786.52</td>
</tr>
<tr>
<td>Cough 786.2</td>
<td>Cough 786.2</td>
</tr>
<tr>
<td>Other dyspnea and respiratory abnormality 786.09</td>
<td>Other dyspnea and respiratory abnormality 786.09</td>
</tr>
<tr>
<td>Wheezing 786.07</td>
<td>Wheezing 786.07</td>
</tr>
<tr>
<td>Shortness of breath 786.05</td>
<td>Shortness of breath 786.05</td>
</tr>
<tr>
<td>Tracheostomy complications 519.0</td>
<td>Tracheostomy complications 519.0</td>
</tr>
</tbody>
</table>
**Example: ICD9-to-10 mapping for suicide and self-inflicted injuries (SSI)**

### Manual mapping (GEM)

<table>
<thead>
<tr>
<th>SSI by jumping from unspecified site</th>
<th>E957.9</th>
<th>Y929</th>
<th>Unspecified place or not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSI by jumping from natural sites</td>
<td>E957.2</td>
<td>Y92838</td>
<td>Other recreation area</td>
</tr>
<tr>
<td>SSI by jumping from other man–made structures</td>
<td>E957.1</td>
<td>Y9289</td>
<td>Other specified places</td>
</tr>
<tr>
<td>SSI by jumping from residential premises</td>
<td>E957.0</td>
<td>Y92009</td>
<td>Unspecified place in unspecified non-institutional (private) residence</td>
</tr>
</tbody>
</table>

### Data driven (iSphereMAP)

<table>
<thead>
<tr>
<th>SSI by jumping from unspecified site</th>
<th>E957.9</th>
<th>Y929</th>
<th>Unspecified place or not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSI by jumping from natural sites</td>
<td>E957.2</td>
<td>Y92838</td>
<td>Other recreation area</td>
</tr>
<tr>
<td>SSI by jumping from other man–made structures</td>
<td>E957.1</td>
<td>Y9289</td>
<td>Other specified places</td>
</tr>
<tr>
<td>SSI by jumping from residential premises</td>
<td>E957.0</td>
<td>Y92009</td>
<td>Unspecified place in unspecified non-institutional (private) residence</td>
</tr>
</tbody>
</table>

Intentional self–harm by jumping from a high place, initial encounter
Take home messages

- EHRs need to be “semantically” translated before being fed into a phenotyping algorithm or statistical model
- Manually curated mappings are imprecise and error prone
- Data driven mappings are scalable and automated
  - Based on summary of co-occurrence: does not require individual level data
  - Unsupervised: does not rely on training labels
Thank you!
Questions?
shixu@umich.edu