Terminology information loss and gain: mapping ICD9CM to OMOP with eMERGE case study
Matthew Levine and George Hripcsak
Department of Biomedical Informatics, Columbia University; Observational Health Data Sciences and Informatics

The problem: Characterizing information loss and gain when mapping into standard OMOP terminologies.

- The benefits of the CDM are abundantly clear—just look around!!
- In addition, OMOP standard terminologies, like SNOMED, nicely facilitate concept set definitions.
- BUT, terminology mapping involves some information loss.
- Possible information loss may be a perceived barrier to potential new members of the OHDSI community.
- By more deeply studying multiple and missed mappings between standard and non-standard terminologies in the CDM, we can:
  - Further improve the CDM
  - Identify pitfalls and trustworthy uses of terminology mapping

**Experimental Overview**

1. We examine eMERGE phenotype condition concept sets (ICD9 only)
2. Identify ICD9 codes with null/invalid/multiple standard mappings
3. Identify patients with condition_source_concept_id in each set of eMERGE ICD9s.
4. Map ICD9 codes to standard SNOMED concepts, and take all the standard descendants of the mapping.
5. Identify patients with condition_concept_id in mapped descendants.
6. Count how many patients are returned ONLY after mapping vs returned ONLY via source codes vs returned from either mapped OR source codes.

**ICD9 to SNOMED mappings**

<table>
<thead>
<tr>
<th>N</th>
<th>eMERGE ICD codes with N mappings</th>
<th>ICD9 codes with N mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6%</td>
<td>1.0%</td>
</tr>
<tr>
<td>1</td>
<td>97.3%</td>
<td>97.8%</td>
</tr>
<tr>
<td>2</td>
<td>1.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>3</td>
<td>0.1%</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

**The data**

NewYork-Presbyterian Hospital clinical data warehouse
- OMOP CDMv5
- Over 3 million patients
- 30 years old

We acknowledge NLM R01 LM06910 for financial support.

**Patient gain/loss when mapping eMERGE concept sets**

Patients are dropped and added from eMERGE cohorts when mapping ICD9 to SNOMED.

**Distributions of patient gain/loss when mapping eMERGE concept sets**

- 70% of cohorts lost 0 patients
- 17% of cohorts lost >1000 patients
- 93% of cohorts gained patients
- Concept sets that contained ICD9 codes that had INVALID or MULTIPLE mappings both GAINED and LOST more patients.

**Distributions of patient gain/loss when mapping ICD9 concepts to SNOMED**

- 88% of eMERGE ICD9 codes mappings lose 0 patients.
- 2% of eMERGE ICD9 codes mappings lose >1000 patients
- 65% of eMERGE ICD9 code mappings gain patients
- eMERGE ICD9 codes with only INVALID mappings both GAINED and LOST more patients.
- eMERGE ICD9 codes with MULTIPLE mappings GAINED more patients, but rarely LOST patients.

**Net change in cohort size due to ICD9 to SNOMED mappings**

Mapping ICD9 to SNOMED usually induces a net increase in cohort size.

**Conclusions and Future Directions**

- We observe changes in cohort size, but we do not yet know whether these are information LOSSES or GAINS
- The next step is to do manual clinical review to determine whether patients should be dropped or added.