Electronic Phenotyping for Genomic Research

George Hripcsak, Columbia University
On behalf of Phenotyping WG
October 30, 2017
1. How can eMERGE improve upon the current labor-intensive phenotyping toward fully-automated phenotyping methods to increase phenotyping efficiency and validity using EMRs?
Phenotype sharing

• One part of the labor is sharing
  – eMERGE adopting OHDSI OMOP Common Data Model
  – Convert current eMERGE data warehouses to same schema and vocabulary
  – But preserve source information
eMERGE phenotype generation

- eMERGE phenotyping lessons
  - [Kho AN, Sci Trans Med 2011]
- Complexity of eMERGE phenotypes
  - [Conway M, AMIA 2011]
- Multi-modal approaches
  - [Peissig PL, JAMIA 2012]
- Use of NQF Quality Data Model
  - [Thompson WK, AMIA 2012]
- Improving validation
  - [Newton KM, JAMIA 2013]
- Design patterns
  - [Rasmussen LV, JBI 2014]
- PhEMA: Phenotype Execution and Modeling Architecture
  - [Pathak et al.]
Phenotype generation lessons

• Challenge of billing codes
• Importance of NLP
  – And multimodal in general
• Complexity of effective phenotype definitions
• Possible improvement from tools and reuse, but mostly just slogging it out
• Differing goals:
  – Knowledge discovery via GWAS needs high PPV
  – Knowledge deployment for decision support also needs sensitivity
Phenotyping for the future

• **High-fidelity phenotypes** [Hripcsak G, JAMIA 2017]
  – Encode degree, severity of condition
    • Redo for past phenotypes?
  – Exploit time to create more accurate phenotypes
  – Encode time of condition
    • Disease course, response to treatment
  – Continuous states (topology, where not dichotomous)
  – Hidden physiologic phenotypes (data assimilation)
  – Latent abstract states (deep learning)
  – Accommodate health care process bias
High-fidelity phenotypes

• Encode degree, severity of condition
High-fidelity phenotypes

- Exploit time to create more accurate phenotypes
- Encode time of condition
High-fidelity phenotypes

• Continuous states (topology, where not dichotomous)

Nicolau, PNAS 2011
High-fidelity phenotypes

- Hidden physiologic phenotypes (data assimilation)

Albers, PLOS Comp Bio 2017
High-fidelity phenotypes

• Latent abstract states (deep learning)
High-fidelity phenotypes

- Accommodate health care process bias

<table>
<thead>
<tr>
<th>concept</th>
<th>hc</th>
<th>sparkline</th>
</tr>
</thead>
<tbody>
<tr>
<td>inr</td>
<td>admit</td>
<td></td>
</tr>
<tr>
<td>ptt</td>
<td>admit</td>
<td></td>
</tr>
<tr>
<td>inr</td>
<td>ambsurg</td>
<td></td>
</tr>
<tr>
<td>ptt</td>
<td>ambsurg</td>
<td></td>
</tr>
<tr>
<td>inr</td>
<td>discharge</td>
<td></td>
</tr>
<tr>
<td>ptt</td>
<td>discharge</td>
<td></td>
</tr>
<tr>
<td>inr</td>
<td>ed</td>
<td></td>
</tr>
<tr>
<td>ptt</td>
<td>ed</td>
<td></td>
</tr>
<tr>
<td>inr</td>
<td>outpatient</td>
<td></td>
</tr>
<tr>
<td>ptt</td>
<td>outpatient</td>
<td></td>
</tr>
</tbody>
</table>

Hripcsak, JAMIA 2013
2. How might machine-learning and other advanced computational tools be used to improve electronic phenotyping in the eMERGE network?
Advanced computational tools

• Natural language processing
  – Large proportion of phenotypes employ it
  – Disparate systems across the network
  – Most get by with relatively simple processing
  – Working on sharing NLP!
Advanced computational tools

• Machine learning research
  – eMERGE research: see following slides
  – Anchors, noisy sets to learn from imperfect training data (MIT, Stanford, Columbia)
  – Active learning to reduce training set labor (Marshfield, ...)
  – Deep learning to characterize patients (Mt. Sinai, ...)
  – Physiologic phenotypes via data assimilation (Columbia)
    • E.g., kidney & liver function, body space, insulin excretion
  – Topology for continuous phenotypes (Stanford, Columbia)
Harvard eMERGE – Rheumatoid Arthritis Machine Learning Phenotype Algorithm

- Machine learning algorithms can be effectively and efficiently applied to a large population to accurately phenotype patients
- Algorithms provide flexibility to adjust sensitivity and specificity to varied use cases compared to pre-defined rules-based algorithms

Rheumatoid Arthritis Algorithm Development Workflow

- Create a training set using clinician chart review (N=200)
- Train a machine learning algorithm
- Use algorithm to identify cases (and controls)
- Validation based on additional 100 chart review (PPV = 0.92)

Rheumatoid Arthritis Algorithm Final Feature Betas

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Beta (weight)</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.017</td>
<td>Model intercept (beta 0)</td>
</tr>
<tr>
<td>patient_dxencnt</td>
<td>-0.954</td>
<td>Number of encounters with an ICD-9 code</td>
</tr>
<tr>
<td>RA_COD_DX_RheumatoidArthritis_v2</td>
<td>1.937</td>
<td>Number of coded Rheumatoid arthritis diagnoses</td>
</tr>
<tr>
<td>RA_COD_DX_Psoriaticarthritis_v2</td>
<td>-0.122</td>
<td>Number of coded Psoriatic arthritis diagnoses</td>
</tr>
<tr>
<td>RA_COD_DX_Lupus</td>
<td>-0.529</td>
<td>Number of coded Lupus diagnoses</td>
</tr>
<tr>
<td>RA_COD_LAB_RFpos1</td>
<td>1.639</td>
<td>Binary indicator where 1=any positive Rheumatoid Factor (RF) lab, else = 0</td>
</tr>
</tbody>
</table>
On Mapping Textual Queries to a Common Data Model

Sijia Liu*†, Yanshan Wang*, Na Hong*, Feichen Shen*, Stephen Wu†, William Hersh†, Hongfang Liu*
*Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA
†Department of Computer Science and Engineering, University at Buffalo, Buffalo, New York, USA
‡School of Medicine, Oregon Health & Science University, Portland, Oregon, USA

• Challenges faced in using NLP for computational phenotyping
  – Poor portability caused by syntactic, semantic, and process variations
  – Semantic gaps among users, experts, and data
  – It is not “one size fits all” solutions for computational phenotyping

• Solutions proposed
  – Improve **syntactic interoperability** by adopting common data models
  – Mitigate the semantic gaps through a combination of deep learning representation, information retrieval, informatics extraction, and late binding NLP and data normalization
  – Develop a platform for sharing NLP knowledge artifacts and mapping between data semantics and expert semantics
PhEMA

- **PhEMA: Phenotype Execution and Modeling Architecture** [Pathak et al.]
  - Standards-based representation of phenotypes
  - Visual tool for authoring phenotypes (PhAT)
  - Execution against OMOP or i2b2 (PheX)
  - Developing NLP & ML extensions
  - Integrates with PheKB
NLP – ML Approach

- Apply exclusion and inclusion criteria based on ICD9 code filtering
- Acquire EMR data for the filtered patients
- Process clinical notes to discover SNOMED-CT and RxNORM concepts with their attributes (Apache cTAKES) and generate feature vectors
- Apply machine learning prediction on feature vectors based on training from expert-provided labels
- Communicate ML model to other sites to run on their data
Relational machine learning for electronic health record-driven phenotyping

Peggy L. Peissig a,*, Vitor Santos Costa b, Michael D. Caldwell c, Carla Rottsheit a, Richard L. Berg a, Eneida A. Mendonca d,e, David Page d,f

Table 5. Comparison of eMERGE phenotyping model precision to ILP+BP

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>eMERGE 1</th>
<th>eMERGE at Marshfield</th>
<th>ILP+BP 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataract</td>
<td>0.960 - 0.977</td>
<td>0.956²</td>
<td>0.877</td>
</tr>
<tr>
<td>Dementia</td>
<td>0.730 – 0.897</td>
<td>0.897³</td>
<td>0.936</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>0.982 – 1.000</td>
<td>0.990³</td>
<td>0.926</td>
</tr>
<tr>
<td>Diabetic Retinopathy</td>
<td>0.676 – 0.800</td>
<td>0.800³</td>
<td>0.976</td>
</tr>
</tbody>
</table>

1 eMERGE precision range taken from Table 3 in Newton et al [6]. The range represents multiple eMERGE institution precision estimates.
2 Precision for Marshfield eMERGE cohort indicating the combined cohort precision definition in Peissig et al [28].
3 eMERGE precision for Marshfield taken from Table 3 in Newton et al [6].
4 ILP+BP: Inductive Logic Programming + Borderline Positives taken from Table 3.
Active Machine Learning

Start with a few labeled cases

raw unlabeled data → human labeling → machine learning → learned model

machine decides which data the human expert should label next

active learning finds optimal classifier with much less human assistance!
3. How can eMERGE assess phenotype comparability across diverse patient populations and diverse healthcare settings (e.g. academic and county hospitals, community clinics and other national healthcare systems)?
Diverse populations and settings

• Design specific eMERGE experiments
  – Busy now with existing phenotypes

• Collaborate with All of Us Research Program
  – Getting up to speed; uses same data model

• Collaborate with OHDSI
  – Large, international set for phenotype part