Mining Electronic Health Records in the Genomics Era

Adam Roth

Pace of genome-wide association study publications since 2005

Hindorff LA, MacArthur J (European Bioinformatics Institute), Merletti F (European Bioinformatics Institute), Vujkovic N, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies
Phenotypic traits for GWAS

1. Candidate driven studies
2. Non-candidate driven studies
3. EHR
Electronic Health Record

1. Billing data
2. Lab results
3. Vital signs
4. Provider documentation
5. Documentation from reports and tests
6. Medication records

EMR vs. EHR?
Billing data

ICD codes
- International classification of diseases
- Diagnosis

Bone
170.x Malignant neoplasm of bone and articular cartilage
  170.0 Bones of skull and face, except mandible
  170.1 Mandible
  170.2 Vertebrae, excluding sacrum and coccyx
  170.3 Ribs, sternum, and clavicle
  170.4 Scapula and long bones of upper limb
  170.5 Short bones of upper limb
  170.6 Pelvic bones, sacrum, and coccyx
  170.7 Long bones of lower limb
  170.8 Short bones of lower limb
  170.9 Bone and articular cartilage site unspecified

CPT codes
- Current procedural terminology
- Services provided
- Billing
- Coronary artery bypass surgery

<table>
<thead>
<tr>
<th>Radiation Treatment</th>
<th>Radiation Treatment Management</th>
<th>77727, 77433, 77432, 77485, 77469</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reason for Refusal</td>
<td>Reason for Refusal/Fee Pt. Folio</td>
<td>Special Radiation Treatment</td>
</tr>
</tbody>
</table>

Laboratory and Vital sign codes

- LOINC
- Logical observation identifiers names and codes

<table>
<thead>
<tr>
<th>Test ID</th>
<th>Test Name</th>
<th>Result ID</th>
<th>Result Name</th>
<th>LOINC Code</th>
<th>LOINC Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>Blood Test</td>
<td>123456</td>
<td>Blood Value</td>
<td>234567</td>
<td>Blood Value</td>
</tr>
<tr>
<td>5678</td>
<td>Urine Test</td>
<td>876543</td>
<td>Urine Value</td>
<td>345678</td>
<td>Urine Value</td>
</tr>
<tr>
<td>9012</td>
<td>Hematocrit</td>
<td>123456</td>
<td>Hematocrit Value</td>
<td>654321</td>
<td>Hematocrit Value</td>
</tr>
</tbody>
</table>

...
Medication Records

• CPOE – Computerized Physician (Provider) Order Entry
  • Pharmacy, laboratory, radiology
• Barcode administration, smart pill box

Provider (Clinical) Documentation

• Dictated notes
• Most diverse source of information
• Difficult to extract
Provider (Clinical) Documentation

- Speech recognition

Structured vs. Unstructured data

- Data storage methods
- Unstructured data
  - Does not have a predefined data model or is not organized in a predefined manner.
- Structured data
  - Organized data
  - Data that resides within a fixed field within a record or file
Structured Data

Unstructured Data


<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable / Field Name</td>
<td>Section</td>
<td>Field Type</td>
<td>Field label</td>
<td>Choices, Calculations, Or Slider Labels</td>
</tr>
<tr>
<td>participant_id</td>
<td>demographics</td>
<td>text</td>
<td>Participant ID</td>
<td></td>
</tr>
<tr>
<td>enroll</td>
<td>demographics</td>
<td>text</td>
<td>Date subject enrolled</td>
<td></td>
</tr>
<tr>
<td>fname</td>
<td>demographics</td>
<td>text</td>
<td>First Name</td>
<td></td>
</tr>
<tr>
<td>lname</td>
<td>demographics</td>
<td>text</td>
<td>Last Name</td>
<td></td>
</tr>
<tr>
<td>city</td>
<td>demographics</td>
<td>text</td>
<td>City</td>
<td></td>
</tr>
<tr>
<td>state</td>
<td>demographics</td>
<td>text</td>
<td>State</td>
<td></td>
</tr>
<tr>
<td>zip</td>
<td>demographics</td>
<td>text</td>
<td>Zipcode</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>demographics</td>
<td>dropdown</td>
<td>Gender</td>
<td>0. Female</td>
</tr>
<tr>
<td>given_birth</td>
<td>demographics</td>
<td>radio</td>
<td>Has the subject given birth before?</td>
<td>0. No</td>
</tr>
<tr>
<td>num_children</td>
<td>demographics</td>
<td>text</td>
<td>How many times has the subject given birth?</td>
<td></td>
</tr>
<tr>
<td>race_other</td>
<td>demographics</td>
<td>text</td>
<td>Please describe:</td>
<td></td>
</tr>
<tr>
<td>dob</td>
<td>demographics</td>
<td>text</td>
<td>Date of birth</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>demographics</td>
<td>date</td>
<td>Age</td>
<td>round(datediff('19000101',<code>dob</code>,'y'))+1</td>
</tr>
<tr>
<td>height</td>
<td>demographics</td>
<td>text</td>
<td>Height (cm)</td>
<td></td>
</tr>
<tr>
<td>weight</td>
<td>demographics</td>
<td>text</td>
<td>Weight (kilograms)</td>
<td></td>
</tr>
<tr>
<td>bmi</td>
<td>demographics</td>
<td>float</td>
<td>BMI</td>
<td>round([weight] * 10000 / [height]^2)</td>
</tr>
<tr>
<td>jcp</td>
<td>demographics</td>
<td>dropdown</td>
<td>Does patient have a primary care physician?</td>
<td>1. Yes</td>
</tr>
<tr>
<td>upload</td>
<td>demographics</td>
<td>file</td>
<td>Upload record documents</td>
<td></td>
</tr>
</tbody>
</table>
EHR Data overview
Intelligent Character Recognition

- Recognize handwriting
- More information
- Neural Networks
  - Speech recognition
- Unstructured to unstructured
- How to get structured?

Dealing with Unstructured Text

- Keyword matching
  - ICD
  - UMLS, SNOMED-CT
  - RxNORM
- Regular expression pattern matching

```
\b[\w.\%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
```

Parse: username@domain.TLD (top level domain)
Dealing with Unstructured Text

- Natural Language Processing
  - Artificial intelligence + Linguistics
  - Understanding of text
  - Rule-based
  - Statistical based

- Parse tree
  - structures elements of a sentence

- Will go into detail tomorrow
- Context, NegEx

---

Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports

Henk Harkema¹, John N. Dowling¹, Tyler Thorblade², and Wendy W. Chapman¹

¹Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, USA
²Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, USA

Abstract

In this paper we describe an algorithm called ConText for determining whether clinical conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. The algorithm infers the status of a condition with regard to these properties from simple lexical clues occurring in the context of the condition. The discussion and evaluation of the algorithm presented in this paper address the questions of whether a simple surface-based approach which has been shown to work well for negation can be successfully transferred to other contextual properties of clinical conditions, and to what extent this approach is portable among different clinical report types. In our study we find that ConText obtains reasonable to good performance for negated, historical, and hypothetical conditions across all report types that contain such conditions. Conditions experienced by someone other than the patient are very rarely found in our report set. A
Connecting phenotype to genotype

Biobank

• Stores human biological samples
• Reuse genetic data
• eMERGE – electronic medical records & genomics network
  • BioVU
  • Opt-out approach
  • Large scale collection
  • >90% participation
  • De-identification with DE-ID
  • Health Insurance Portability and Accountability Act
  • 150,000 subjects
Phenotype selection algorithm

- 4 ‘buckets’
  - Definite cases
  - Controls
  - Probable case
  - excluded
- 4 types of data
  - Billing data
  - Other structured data (lab values)
  - Medication information
  - NLP

EDGR- EHR driven genomic research

- Non-disease based phenotypes
  - Hair, eye color
  - handedness
- cTAKES
  - Peripheral arterial disease
  - Keywords in radiology reports
  - AND- OR- NOT Boolean logic
- Rzhetsky
  - 1.5 million patients
  - Disease co-occurrence
EHR for Genetic Association Analysis

• BioVU
• 10,000 samples genotyped at 21 SNPs
• Atrial fibrillation, crohns disease, multiple sclerosis, rheumatoid arthritis, T2D
• Phenotype identification algorithm
  • Billing code queries, medication names
  • NLP- key findings, family history
• Result: 97% PPV
• ICD9 alone: 56-89%
  • Coding errors, misdiagnosis from non-specialist, indeterminate diagnoses

IBM Watson

https://www.youtube.com/watch?v=8lGI0h_jAp8
Works Cited

• PLOS Translational bioinformatics textbook