THE EHR LANGUAGE GARDEN

Leveraging Variability in Health Documentation

Denis Newman-Griffis

NIH Clinical Center / University of Pittsburgh

September 1, 2020
Research ➔ Practice
Sublanguage: the secret sauce

Clinical decision support
Benefits administration
Biosurveillance

NLP

Nursing  Radiology  Pharmacy

GARDEN DESIGN

INSPERATION AND IDEAS
ERH data variability at scale

Content – what do the records say?

Form – how do they say it?

Structure – what are the pieces?
Context: SSA disability programs

- National data
- All providers/EHRs
- Unreliable metadata
"Defining" disability

Medical conditions

- High mortality conditions
- Medical listings (business rules)

Functional limitations

- Ability to perform work-related activities
- Substantial Gainful Employment

Need NLP that can handle both!
Planting the garden: findings

Content – Rehabilitation medicine as a sublanguage

Form

Structure
Health strategies – Rehabilitation

Goal: Restore/optimize function

- Adapt to health condition (e.g., chronic or incurable)
- Interactions with world

Under-studied domain
Health strategies – Curative

Goal: Cure health conditions

- Diagnosis
- Treatment
- Physiological/internal

Most of clinical NLP!
Multi-institution data

BTRIS
- 155K records
- Research patients
- 130 doctypes

OSUMC
- 418K records
- Chronic diseases
- 43 doctypes

MIMIC-III
- 2M records
- ICU admissions
- 25 doctypes*
Data classifications

- **Domain**
  - **Diagnostic** – most concerned with diagnosis and treatment
  - **Functioning** – most concerned with functioning status

- **Discipline**
- **Functional Area**
Data classifications

- **Medical** – curative documents from "standard" specialties
- **Therapy** – therapeutic specialties (PT, OT, RT, etc.)
- **Ancillary** – non-physiological (mental health, social work)
- **Other** – administrative documents
Data classifications

Domain

Discipline

Functional Area

PT – physical therapy
OT – occupational therapy
RT – recreational therapy
SLP – speech/language pathology
Psych – psychological/iatric
Neuro – neurological
SW – social work
General – catchall bucket
Rehab medicine vocabulary is distinct

BL = Variance within document types
D = Variance between doctype types in Diagnostic
D/F = Variance between Diagnostic doctype types and Functioning doctype types
Rehab medicine vocabulary is distinct

Document classification accuracy (unigram features)
Significant differences across institutions
Significant differences across institutions
Ejection fraction: 90%
Lab creatinine: 3 mg/dL

There has been removal of [a swan-ganz catheter]_{Treatment} and placement of [a right internal jugular vascular catheter]_{Treatment}.

Pt 45 yr old tech worker, sedentary activity but hikes on weekends.

[Ambulation: 4]_{Mobility}
Observations:
Pt is weight bearing: [she ambulates independently w/o use of assistive device]_{Mobility}. Limited to very brief examination.
Planting the garden: findings

**Content**

**Form** – Differences in clinical concept usage

**Structure**

Characterizing document types

Document/section structural patterns inform meaning

- Field names vs observations
- Temporality (future/past/recurrent)
- Perceived importance (e.g. Chief Complaint)

Document types change priors for disambiguation

- “Depression” in Psychiatric Consult vs GE Exam
Conceptual vs lexical analysis

Prior work used lexical content to describe clinical sublanguages
- Feldman et al, 2016
- Grön et al, 2019

Concepts (symptoms, diseases, procedures, etc) are stock in trade of clinical language
- Multiple surface forms
- Ambiguity ("Cold")
Learning concept embeddings: JET

- Train word/term/concept embeddings jointly
- Distant supervision using known terminology
- Noisy, but good quality

Terminologies: SNOMED CT, LOINC

Data: MIMIC-III

Measuring concept usage similarity

- Measured by overlap of nearest neighbor sets
- Similarity metric in \([0,1]\)
- Compare inter-type overlaps to intra-type overlaps

<table>
<thead>
<tr>
<th>Neighbors of Onion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set A</strong></td>
</tr>
<tr>
<td>Cucumber</td>
</tr>
<tr>
<td>Squash</td>
</tr>
<tr>
<td>Beans</td>
</tr>
<tr>
<td>Green</td>
</tr>
<tr>
<td>Pasta</td>
</tr>
</tbody>
</table>
Inter-type similarity is significantly lower than intra-type

<table>
<thead>
<tr>
<th></th>
<th>Case Management</th>
<th>Discharge Summary</th>
<th>Echo</th>
<th>Nursing/Other</th>
<th>Nutrition</th>
<th>Physician</th>
<th>Radiology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case Management</strong></td>
<td>0.75</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td><strong>Discharge Summary</strong></td>
<td>0.01</td>
<td>0.67</td>
<td>0.24</td>
<td>0.32</td>
<td>0.00</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Echo</strong></td>
<td>0.01</td>
<td>0.24</td>
<td>0.65</td>
<td>0.13</td>
<td>0.00</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td><strong>Nursing/Other</strong></td>
<td>0.00</td>
<td>0.32</td>
<td>0.13</td>
<td>0.60</td>
<td>0.00</td>
<td>0.27</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Nutrition</strong></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.73</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Physician</strong></td>
<td>0.00</td>
<td>0.34</td>
<td>0.36</td>
<td>0.27</td>
<td>0.01</td>
<td>0.57</td>
<td>0.26</td>
</tr>
<tr>
<td><strong>Radiology</strong></td>
<td>0.01</td>
<td>0.33</td>
<td>0.40</td>
<td>0.31</td>
<td>0.00</td>
<td>0.26</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Nearest neighbors:
Diabetes Mellitus (C0011849)

<table>
<thead>
<tr>
<th>Discharge Summary</th>
<th>Nursing/Other</th>
<th>Radiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes (C0011847)</td>
<td>Gestational Diabetes (C0085207)</td>
<td>Poorly controlled (C3853134)</td>
</tr>
<tr>
<td>Type 2 (C0441730)</td>
<td>A2 immunologic symbol (C1443036)</td>
<td>Insulin (C0021641)</td>
</tr>
<tr>
<td>Type 1 (C0441729)</td>
<td>Diabetes Mellitus, Insulin-Dependent (C0011854)</td>
<td>Diabetes Mellitus, Insulin-Dependent (C0011854)</td>
</tr>
<tr>
<td>Gestational Diabetes (C0085207)</td>
<td>Factor V (C0015498)</td>
<td>Diabetes Mellitus, Non-Insulin-Dependent (C0011860)</td>
</tr>
<tr>
<td>Diabetes Mellitus, Insulin-Dependent (C0011854)</td>
<td>A1 immunologic symbol (C1443035)</td>
<td>Stage level 5 (C0441777)</td>
</tr>
</tbody>
</table>

Strings: “diabetes mellitus”, “diabetes mellitus dm”
### Nearest neighbors:
**Mental state (C0278060)**

<table>
<thead>
<tr>
<th>Discharge Summary</th>
<th>Echo</th>
<th>Radiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherent (C4068804)</td>
<td>Donor [LOINC] (C3263710)</td>
<td>Mental status changes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(C0856054)</td>
</tr>
<tr>
<td>Confusion (C0009676)</td>
<td>Donor person (C0013018)</td>
<td>Abnormal mental state</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(C0278061)</td>
</tr>
<tr>
<td>Respiratory status [LOINC]</td>
<td>Respiratory arrest (C0162297)</td>
<td>Level of consciousness</td>
</tr>
<tr>
<td>(C2598168)</td>
<td></td>
<td>(C0234425)</td>
</tr>
<tr>
<td>Respiratory status</td>
<td>Organ donor [LOINC] (C1716004)</td>
<td>Level of consciousness</td>
</tr>
<tr>
<td>(C1998827)</td>
<td></td>
<td>[LOINC] (C4050479)</td>
</tr>
<tr>
<td>Abnormal mental state</td>
<td>Swallowing (C4281783)</td>
<td>Mississippi (C0026221)</td>
</tr>
<tr>
<td>(C0278061)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Strings: “mental status”, “mental state”
Embeddings pick up template patterns

*Mental status* in Echo notes

**PATIENT/TEST INFORMATION**
Indication: Pt presents with reduced *mental status*

**PATIENT/TEST INFORMATION**
Indication: Pt presents in vegetative state, consider for *organ donation*
Planting the garden: findings

Content

Form

Structure — Structural text features capture format and content variation

Work by Bart Desmet, Guy Divita, and Aya Zirikly
Sources of variability in SSA data

Document Source
- SSA Consultative Exams, CCDA documents from EHR, VA data, scanned notes

Content Types
- SOAP notes, radiology reports, labs, surveys

Formatted Structure
- Headers/footers, columns, section names, checkboxes
Classify early, process better

- 70K documents
- Disability claimants from 5 states
- Unreliable doctypes
Page-level Features

- Number of Characters, Words, Lines, Sentences
- Number of Punctuation, Delimiters
- Number of Section Names, Section Zones, Nested Sections
- Number of Slot Values, Slot Names, Slot Value Values
- Number of Check Boxes (this wasn’t actually working as it turns out)
- Number of Tables
- Number of Lists, List Elements
- Number of Questions
- “Text Tiling” Vector fingerprint (2 numbers)
Related Work: Text Tiling

Marti Hearst (1994): Using word sequences to build a signal to indicate topic/paragraph shifts.

Figure 6
Results of the block similarity algorithm on the Stargazer text with $k$ set to 10 and the loose boundary cutoff limit. Both the smoothed and unsmoothed plot are shown. Internal numbers indicate paragraph numbers, $x$-axis indicates token-sequence gap number, $y$-axis indicates similarity between blocks centered at the corresponding token-sequence gap. Vertical lines indicate boundaries chosen by the algorithm; for example, the leftmost vertical line represents a boundary after paragraph 3. Note how these align with the boundary gaps of Figure 5 above.

Page-level PCA

**Input**
- Structural features for each page

**Output**
- Orthogonal transform into a set of principal components
- Dimensionality reduction and variance identification
PCA “Angelfish” Plot
PCA “Angelfish” Plot
Density estimation (dot=100pgs)
Density estimation (dot=500pgs)
Density estimation (dot=1000 pgs)
Density estimation (dot=3000pgs)
Density estimation (dot=5000pgs)
Observations from the Density Plot
# Topic analysis ➔ semantic correlations

<table>
<thead>
<tr>
<th>Topic</th>
<th># non-relevant pages</th>
<th>#relevant pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social/family history for mental disorder</td>
<td>11358</td>
<td>975</td>
</tr>
<tr>
<td>mental status evaluation -risk of suicide</td>
<td>14239</td>
<td>122</td>
</tr>
<tr>
<td>Mental disorder symptoms and treatment history</td>
<td>9900</td>
<td>3613</td>
</tr>
<tr>
<td>Impression of mental disorder</td>
<td>15500</td>
<td>100</td>
</tr>
<tr>
<td>Lab test results [Topic 6]</td>
<td>11959</td>
<td>2406</td>
</tr>
</tbody>
</table>
Planting the garden: findings

Content
Form
Structure

Technology development
Whither sublanguage analysis?
EpiBio: Heterogeneous SSA data

- Geographic variation in mental health-related documentation
  - Stigma
  - Lack of details / re-coding

- Format → structure
  - Sectionizing
  - Semi-structured forms
Pitt: EHR language and health equity

- Documentation differences for patients of different races
  - What is recorded?
  - How is it recorded?

- Integrating patient-generated language with clinical observations
  - Self-reported functional status

- Ambiguity in health language
VA: Knowledge exchange

- Challenges shared by national health systems
  - Geographic and institutional variation
  - Large portion of SSA medical evidence comes from VA

- Cerner transition
  - Changes in documentation practice
  - Effect on NLP pipelines
Acknowledgments

- Aya Zirikly (NIH)
- Guy Divita (NIH)
- Bart Desmet (NIH)
- Jona Camacho Maldonado (NIH)
- Pei-Shu Ho (NIH)
- Beth Rasch (NIH)
- Eric Fosler-Lussier (Ohio State)
- Albert Lai (WashU)

Funding support from NIH Intramural Research Program and the US Social Security Administration.
Thank you!

denis.griffis@nih.gov
dnewmangriffis@pitt.edu