Adapting clinical NLP methods for multi-site medical products research

David S. Carrell, PhD
Kaiser Permanente Washington Health Research Institute
June 15, 2017, FDA White Oak Campus, Silver Spring, MD
Overview

<table>
<thead>
<tr>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary use of EHR data &amp; text</td>
</tr>
<tr>
<td>Accessing clinical text</td>
</tr>
<tr>
<td>Interpretive challenges</td>
</tr>
<tr>
<td>Multi-site clinical NLP implementation</td>
</tr>
</tbody>
</table>
Secondary use of EHR data & text

- Reusing data/text for some other purpose
- Seemingly simple tasks can be challenging
Secondary use of EHR text

Electronic exhaust of healthcare delivery

- Intended use:
  - Care of one patient
  - Through an EHR interface
  - By clinicians
  - In one healthcare system
  - In context of other encounters
  - Limited access (privacy)

- Use for purposes not originally intended may be challenging
Secondary use of EHR text

Electronic exhaust of healthcare delivery

- Information needed for patient care ≠ information needed for research

  **Example**: Does a patient have pre-cancerous colon polyps?

  - **Clinical**: Easily resolved in a patient chart.
  - **Research**: Meta data *do not link* pathology reports to their colonoscopy procedures
    - *Probabilistic* matching
Secondary use of EHR text – Pathology reports

Electronic exhaust of healthcare delivery

- Finding colonoscopy-related pathology reports requires NLP
  Rule: Report contains ≥1 of:
  anal verge, ascending colon, ascending mass, ascending polyp, Cecal, Cecum, Colon, Colonic, descending colon, ic valve, ileocecal valve, ileum, rectal, rectosigmoid, rectum, sigmoid, sigmoid polyp, splenic flexure, terminal ileum, transverse colon, transverse polyp

- Most (not all) colonoscopy and pathology reports have same date

| Pathology report counts by number of days before/after colonoscopy procedure date. |
|---------------------------------|----------------|----------------|----------------|----------------|----------------|
| Pathology date                  | Days before CSPY | Same date | Days after CSPY |
|---------------------------------|----------------|----------|----------------|----------------|----------------|
|                                 | -5-14 | -4 | -3 | -2 | -1 |                      | +1 | +2 | +3 | +4 | +5-14 |
| N Path Rpts                     | 7     | 1  | 3  | 0  | 7  | 1630                | 50 | 52 | 29 | 4  | 3  |

18 (92%) 138 (8%)
Secondary use of EHR text – “duplicate” notes

*It’s like déjà vu all over again  – Yogi Berra*

- As the legal medical record *every version* of a note is preserved
  - Every edit/save generates a new copy
  - EHR displays only the most recent version (primary use)
  - Tricky to de-duplicate in the back end database (secondary use)

<table>
<thead>
<tr>
<th>N versions</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,594,182</td>
<td>82%</td>
</tr>
<tr>
<td>2</td>
<td>107,412</td>
<td>6%</td>
</tr>
<tr>
<td>3</td>
<td>26,629</td>
<td>1%</td>
</tr>
<tr>
<td>4</td>
<td>7,151</td>
<td>0%</td>
</tr>
<tr>
<td>5</td>
<td>2,027</td>
<td>0%</td>
</tr>
<tr>
<td>6</td>
<td>618</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>224</td>
<td>0%</td>
</tr>
<tr>
<td>8</td>
<td>73</td>
<td>0%</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>0%</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>0%</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>0%</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>0%</td>
</tr>
</tbody>
</table>

1,934,518

3,025 ≥ 5 versions
## Accessing clinical text

- Need for privacy, de-identification
- Text availability
- Text degradation
Accessing clinical text

- All clinical text may contain highly sensitive information
  - Unlike structured data, can’t just suppress sensitive content
  - Risks to patient privacy
  - Institutional risk

- Clinical text seldom exits local firewall
  - Lack of freely available corpora impedes technical progress, collaboration

- Text de-identification
  - 94-98% effective
  - Requires technical sophistication, knowledge
  - Added cost
Accessing clinical text – de-identification cost

- Cost of manually annotating personally identifiable information (PII)
- 100 Family Practice notes (1,093 PII instances)
Accessing clinical text – The process

- Corpus assembly: multi-step process highly reliant on local expertise
### Accessing clinical text – What text exists?

- Availability of clinical text varies by site
- Study of prescription opioid abuse/addiction/overdose in four sites

<table>
<thead>
<tr>
<th></th>
<th>KPNW</th>
<th>Optum</th>
<th>TennCare</th>
<th>KPW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Epic EHR</td>
<td>Subset of patients w/ EHR</td>
<td>Epic EHR</td>
<td>Epic EHR</td>
</tr>
<tr>
<td></td>
<td>Outpatient</td>
<td>Outpatient</td>
<td>Outpatient</td>
<td>Outpatient clinics</td>
</tr>
<tr>
<td></td>
<td>Hospital, ER</td>
<td>NLP-extracted term lists</td>
<td>Hospital, ER</td>
<td>ER (partial)</td>
</tr>
<tr>
<td></td>
<td>In-house substance use treatment</td>
<td>NLP-extracted term lists</td>
<td>At one location (VUMC)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Claims data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- **KPNW**
  - Epic EHR
  - Outpatient
  - Hospital, ER
  - In-house substance use treatment

- **Optum**
  - Subset of patients with EHR
  - Outpatient
  - NLP-extracted term lists
  - Claims data

- **TennCare**
  - Epic EHR
  - Outpatient
  - Hospital, ER
  - At one location (VUMC)

- **KPW**
  - Epic EHR
  - Outpatient clinics
  - ER (partial)
Accessing clinical text – degraded text

**EHR version**

SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son John

Today she was in usual state of health. was going to a restaurant, and on way in "just hit her suddenly" with double vision.

```{deletions}
```

OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2 97% Estimated Body mass index is 21.92 kg/(m^2)

```{deletions}
```

LABS / STUDIES:
HEMATOCRIT (%)
Date Value Low High Status
## Accessing clinical text – degraded text

*The Worst You Ever Gave Me Was the Best I Ever Had—Frank Sinatra*

<table>
<thead>
<tr>
<th>EHR version</th>
<th>Secondary use database version</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBJECTIVE:</strong> Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplopia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in “just hit her suddenly” with double vision. <strong>[deletions]</strong></td>
<td><strong>SUBJECTIVE:</strong> Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diplopia and headache. Here with her son John Today she was in usual state of health. was going to a restaurant, and on way in “just hit her suddenly” with double vision. <strong>[deletions]</strong></td>
</tr>
<tr>
<td>OBJECTIVE: BP 165/82</td>
<td>Pulse 69</td>
</tr>
<tr>
<td><strong>LABS / STUDIES:</strong></td>
<td><strong>LABS / STUDIES:</strong></td>
</tr>
<tr>
<td>HEMATOCRIT (%)</td>
<td>Date Value Low High Status</td>
</tr>
<tr>
<td>3/21/2011 46 34 49 Final</td>
<td></td>
</tr>
<tr>
<td>3/29/2010 41 34 49 Final</td>
<td></td>
</tr>
<tr>
<td>6/18/2010 39 34 49 Final</td>
<td></td>
</tr>
<tr>
<td>10/4/2007 47 34 49 Final</td>
<td></td>
</tr>
<tr>
<td>8/20/2006 41 34 49 Final</td>
<td></td>
</tr>
<tr>
<td>8/23/2007 37 34 49 Final</td>
<td></td>
</tr>
<tr>
<td><strong>[deletions]</strong></td>
<td><strong>[deletions]</strong></td>
</tr>
<tr>
<td><strong>ASSESSMENT / PLAN:</strong></td>
<td><strong>ASSESSMENT / PLAN:</strong></td>
</tr>
<tr>
<td>1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double vision worrisome for posterior circulation. I didn’t see notes of prior anisocoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA Spoke with Barb, RN at Pleasant Valley Medical Center ER and gave heads up</td>
<td>1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double vision worrisome for posterior circulation. I didn’t see notes of prior anisocoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA Spoke with Barb, RN at Pleasant Valley Medical Center ER and gave heads up</td>
</tr>
<tr>
<td>Christine Jones, M.D.</td>
<td></td>
</tr>
<tr>
<td>(123) 456-7890</td>
<td></td>
</tr>
<tr>
<td>(123) 456-7891 FAX</td>
<td></td>
</tr>
</tbody>
</table>
### Accessing clinical text – degraded text

*The Worst You Ever Gave Me Was the Best I Ever Had* – Frank Sinatra

<table>
<thead>
<tr>
<th>EHR version</th>
<th>Secondary use database version</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJECTIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diploria and headache. Here with her son John</td>
<td></td>
</tr>
<tr>
<td>Today she was in usual state of health, was going to a restaurant, and on way in &quot;just hit her suddenly&quot; with double vision.</td>
<td></td>
</tr>
<tr>
<td><strong>{deletions}</strong></td>
<td></td>
</tr>
</tbody>
</table>
| OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2
97% Estimated Body mass index is 21.92 kg/(m\(^2\)) |
| **{deletions}** |
| LABS / STUDIES: |
| HEMATOCRIT (%) |
| Date Value Low High Status |
| 3/21/2011 46 34 49 Final |
| 3/29/2010 41 34 49 Final |
| 6/18/2010 39 34 49 Final |
| 10/4/2007 47 34 49 Final |
| 8/20/2008 41 34 49 Final |
| 8/23/2007 37 34 49 Final |
| **{deletions}** |
| ASSESSMENT / PLAN: |
| SUBJETIVE: Jane B Smith is a pleasant 77 year old female with the above past medical history / problem list who comes in with complaint(s) of diploria and headache. Here with her son John Today she was in usual state of health, was going to a restaurant, and on way in "just hit her suddenly" with double vision. **{deletions}** |
| OBJECTIVE: BP 165/82 | Pulse 69 | Temp 97.8 °F (36.6 °C) | Resp 19 | SpO2
97% Estimated Body mass index is 21.92 kg/(m\(^2\)) **{deletions}** |
| LABS / STUDIES: |
| HEMATOCRIT (%) |
| Date Value Low High Status |
| 3/21/2011 46 34 49 Final |
| 3/29/2010 41 34 49 Final |
| 6/18/2010 34 49 Final |
| 10/4/2007 34 49 Final |
| 8/20/2008 41 34 49 Final |
| 8/23/2007 37 34 49 Final |
| **{deletions}** |
| ASSESSMENT / PLAN: 1. Double vision - acute onset
Accessing clinical text – degraded text

I never said most of the things I said – Yogi Berra

- Progress note in EHR:

  COPD oxygen and steroid dependent on oxycontin and oxycodone for dyspnea and pain. Follow up in 1 week.

- Back-end database version:

  COPD oxygen and steroid dependent on oxycontin and oxycodone for dyspnea and pain. Follow up in 1 week.

- False positive NLP hit in abuse/addiction study.
Interpretive challenges in clinical text

- The synonymy problem
- The copy-and-paste problem
- The freedom problem in free-text
Interpretive challenges: *Synonymy*

*I know it when I see it* – Justice Potter Stewart

- **Example:** *Clostridium difficile infection* ("c diff")
  - 1 term covers 63%
  - 6 more terms cover 32%; 23 terms cover the remaining 5%

<table>
<thead>
<tr>
<th>Negation Status of NLP-Extracted Mammogram Findings</th>
<th>NLP-Extracted Finding</th>
<th>Negated</th>
<th>Affirmed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>mass or lump</td>
<td>8,691 (95%)</td>
<td>453 (5%)</td>
<td>9,144</td>
<td></td>
</tr>
<tr>
<td>calcification</td>
<td>6,739 (87%)</td>
<td>1,046 (13%)</td>
<td>7,785</td>
<td></td>
</tr>
<tr>
<td>architectural distortion</td>
<td>1,875 (78%)</td>
<td>542 (22%)</td>
<td>2,417</td>
<td></td>
</tr>
<tr>
<td>asymmetry</td>
<td>100 (17%)</td>
<td>486 (83%)</td>
<td>586</td>
<td></td>
</tr>
<tr>
<td>distortion</td>
<td>137 (66%)</td>
<td>71 (34%)</td>
<td>208</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17,542 (87%)</td>
<td>2,598 (13%)</td>
<td>20,140</td>
<td></td>
</tr>
</tbody>
</table>

Kaiser Permanente
Interpretive challenges: Copy-and-paste

I never said most of the things I said. —Yogi Berra

- EHR text has a copy-and-paste problem
- Most valuable text is manually entered
- <20% of a note’s text is manually typed
  - Wang, JAMA 2017
- The text we most care about is written:
  - In a hurry
  - While multi-tasking
  - Usually without editing
  - Often without punctuation
Interpretation: Medication side effect mentions

KPW medication side effects project (underway)

- 89,377 patients with 113,564 new anti-depressant medication episodes
- 21,602 patients with 25,439 new anti-psychotic medication episodes
- 2005-2016
- Goal: Capture side effects described in Family Practice and Behavioral Health notes
Interpretation: Medication side effect mentions

The importance of context

Found in the notes of patient starting a new anti-depressant:

- “Zoloft caused increased anxiety”
- “feels panicked”
- “not able to sleep”
- “better experience with newer medication (Lexapro)”

DAY 0

SUBJECTIVE: … [NN] year old [MALE/FEMALE] with recurrent depression … Early AM wakening ... no energy … Hx of depression in past with poor medication experience. Tried Imipramine with bladder sx, Zoloft caused increased anxiety (felt unable to cope and poor sleep), Trazadone didn’t help with depression. Sister coping with depression with better experience with newer medication (Lexapro) …

ASSESSMENT/PLAN: Depression, severe … start with Celexa

Day 4

SUBJECTIVE: … feels panicked by not able to sleep …

ASSESSMENT/PLAN: Acute sleep disturbance … add short term Alprazolam for acute anxiety and sleep

Day 18

SUBJECTIVE: … checking back on depression/anxiety

ASSESSMENT/PLAN: Depression/anxiety improving. Continue current dose of Celexa … Alprazolam

Day 39

SUBJECTIVE: … Feels like has returned to "normal"
Interpretation: Typical side effect mentions

- Attribution?
- Severity?
- Rare MSEs?

I am glad the nortriptyline is helping with your pain and burning. Since it also seems to be causing severe dry mouth, I recommend decreasing your dose...

Stopped the Sertraline a week ago secondary to worsening SI after increasing the dose.

He had difficulty achieving orgasm on fluoxetine.

Last visit changed from Prozac to Wellbutrin. Decreased libido is not better.

He is experiencing a diminished libido but says this is tolerable right now.

Is concerned about weight gain.

Less anxious ... Happy about weight gain.

Was on higher dose but was causing side effects.

SE: Unable to cry; early a.m. Awakening; occasionally feels nauseated.
## Interpretation: Abbreviation SE (side effect)

Word sense of "SE" in a sample of 50 clinic notes randomly selected from 18,542 notes containing "SE" 2005-2016.

<table>
<thead>
<tr>
<th>Sense of &quot;SE&quot;</th>
<th>Mentions</th>
<th>Percent</th>
<th>Estimate in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinic address (quadrant)</td>
<td>17</td>
<td>34%</td>
<td>6,304</td>
</tr>
<tr>
<td>Patient address (quadrant)</td>
<td>16</td>
<td>32%</td>
<td>5,933</td>
</tr>
<tr>
<td>Hypothetical MSE(^3)</td>
<td>10</td>
<td>20%</td>
<td>3,708</td>
</tr>
<tr>
<td>MSE(^4)</td>
<td>3</td>
<td>6%</td>
<td>1,113</td>
</tr>
<tr>
<td>MSE confirmed absent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misspelling(^6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish pronoun(^7)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) The 18,542 notes were ~1% of the 2,138,866 notes for 81,804 patients with new anti-depressant medications during the 180-day period following medication start.

"started on fluoxetine ... Mood is somewhat better, however ... notes significant spaciness and inability to concentrate at work ... symptom occurred in the first week, but has worsened on the higher dose ... ASSESSMENT: major depression on fluoxetine with significant SE unlikely to resolve PLAN: Switch to citalopram"
Interpretation: *Hypothetical* mentions

- Hypotheticals are 3-5 times more common than actual side effect mentions

| **Dry mouth is a common side effect** of many prescription and nonprescription drugs |
| **You question if metoprolol causes dry mouth**, this is possible but unlikely |
| **Potential side effects could include** but not limited to: fatigue, mild to severe somnolence, increased appetite/potential weight gain, dizziness, orthostatic hypotension, elevated liver function tests, or paradoxical reactions such as insomnia, anxiety, agitation, panic attacks, insomnia, irritability, hostility, or worsening depression or suicidal ideation. |
| **If side effects** are tolerable patient will continue Rx and they will likely abate |
| **Benefits, risks, SE, alternative of no meds yes** |
Interpretation: List-style reporting

Benefits, risks, SE, alternative of no meds yes

**ASSESSMENT:**

- Therapeutic alliance in full
- Has not had full treatment course yet distinct clinical criteria
- Course of time

**PLAN:**

- Monitor course, progress, and success criteria. Patient doing well and no side effects. Mother in support group. Success criteria and self-care.
- Stress and depression (increased progress)
- Brain imaging resources
- Nurse success in no self-harm behavior and will develop

-Symptom(s): see subjective portion of note and above
- Goal(s): reduction in symptoms and improved functioning
- Intervention: psychotropic medication, also see instructions

Benefits, risks, SE, alternative of no meds yes

Rx course & natural history w & w/o treatment yes
Review studies available yes
Refill policy, Stimulant policy no
Patient resources provided yes
Interpretation: List-style reporting

- Checklist style (is the “X” before or after the symptom?)
- Note misspelled “abominal”
  - In original (and 591 other notes since 2011)
  - Spelling distance algorithms may help (e.g., allow 1 missing letter)
  - But, allowing 1 missing letter: “addiction [to] opioids” = “addition [of] opioids”

### ROS (x indicates positive finding) (Remainder of comp review of systems is negative except as noted in HPI)

- [ ] Fatigue, weakness
- [ ] Loss of appetite
- [ ] Weight loss
- [ ] Weight gain
- [ ] Fever
- [ ] Night sweats
- [ ] Trouble swallowing
- [ ] Heartburn
- [ ] Indigestion
- [ ] Nausea, vomiting
- [ ] Abominal pain, discomfort
- [ ] Change in bowel habits
- [ ] Constipation
- [ ] Diarrhea
- [ ] Blood with bowel movement
- [ ] Fecal incontinence
- [ ] Eye problems
- [ ] Nose bleeds
- [ ] Mouth sores
- [ ] Cough
- [ ] Shortness of breath
- [ ] Excessive snoring, sleep apnea
- [ ] Chest pain
- [ ] Irregular heartbeat or palpitations
- [ ] Swelling in ankles
- [ ] Confusion
- [ ] Numbness
- [ ] Depression or anxiety
- [ ] Trouble sleeping
- [ ] Trouble with urination
- [ ] Joint problems
- [ ] Itching or rash
- [ ] Skin problems.
Multi-site clinical NLP implementation

- Sending algorithms to the text
- Bringing text to the algorithms
- Salient issues in multi-site NLP
## Multi-site NLP implementation strategies

*Send algorithms to the text or bring text to the algorithms*

<table>
<thead>
<tr>
<th>Sending algorithms to the text</th>
<th>Bringing text to the algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td><strong>Advantages</strong></td>
</tr>
<tr>
<td>– Maintains local control over text (no data sharing)</td>
<td>– Uniform processing</td>
</tr>
<tr>
<td>– More likely to discover bugs/errors</td>
<td>– Simplifies NLP system tailoring</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td><strong>Disadvantages</strong></td>
</tr>
<tr>
<td>– Additional software engineering</td>
<td>– De-identification need (local)</td>
</tr>
<tr>
<td>– Variation in sites’ technical proficiency</td>
<td>– Loss of local control over text (mitigated by DUA)</td>
</tr>
<tr>
<td>– Site differences unavoidable</td>
<td></td>
</tr>
</tbody>
</table>

- *Local assembly of corpora always necessary*
Multi-site NLP implementation strategies

Sending algorithms to the text

- Advantages
  - Maintains local control over text (no data sharing)
  - More likely to discover bugs/errors

- Disadvantages
  - Additional software engineering
  - Variation in sites’ technical proficiency
  - Site differences unavoidable
Multi-site NLP: Sending algorithms to text

Portable NLP system used in the eMERGE consortium

- Task: Process free-text imaging reports from 7 eMERGE consortium sites to support GWAS study of CAAD
- Rule-based system captured numeric ("... 50–69% stenosis ...") and qualitative ("... completely occluded ...") descriptions
- Designed for portability and ease of use
  - Self-installing JAVA application
  - Simple GUI, data output
  - Trouble-shooting features
- Columbia, Geisinger, Harvard
- Three revisions needed
  - 1) Original, 2) Check exam type, 3) Ignore normative range boilerplate
Multi-site NLP: Sending algorithms to text

Portable NLP system used in the eMERGE consortium

Issues:

• Design NLP system for portability
• Accommodate site differences in sources, language
• Tailoring needed at each site
• Tailoring across firewalls is tricky (*tele-informatics*)
• Validation needed at each site
• Simpler NLP tasks are the best candidates
Multi-site NLP: Bringing text to algorithms

- **Advantages**
  - Uniform processing
  - Simplifies NLP system tailoring

- **Disadvantages**
  - De-identification need (local)
  - Loss of local control over text (mitigated by DUA)
Multi-site NLP: Bringing text to the algorithms

Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)

- Task: Calculate colonoscopy quality metrics in diverse settings
- Centralized adaptation of an existing NLP system
  - Centralized, manually-annotated training and validation corpora
  - All software engineering at lead site
- Each site assembles, de-identifies corpora
  - 2 years of colonoscopy & associated pathology reports
  - De-identification via open-source Miter MIST or commercial De-ID®
  - DUAs & IRB reviews
- KPW, UNC, UPMC, Central Illinois Endoscopy (CIE)
- 2+ years to adapt existing NLP system!
## Multi-site NLP: Bringing text to the algorithms

*Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)*

<table>
<thead>
<tr>
<th>Site</th>
<th>Location</th>
<th>Practice type</th>
<th>EHR type</th>
<th>Compensation</th>
<th>No. of MDs</th>
<th>No. of CSPYs</th>
<th>Report characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPW</td>
<td>West</td>
<td>HMO</td>
<td>Comprehensive</td>
<td>Salary</td>
<td>18</td>
<td>12,098</td>
<td>980/716</td>
</tr>
<tr>
<td>CIE</td>
<td>Mid-W</td>
<td>Private</td>
<td>GI specialty</td>
<td>Fee-for-service</td>
<td>11</td>
<td>13,036</td>
<td>504/214</td>
</tr>
<tr>
<td>UNC</td>
<td>South</td>
<td>Univ.</td>
<td>Comp. &amp; GI</td>
<td>Salary + incentive</td>
<td>53</td>
<td>19,062</td>
<td>733/247</td>
</tr>
<tr>
<td>UPMC</td>
<td>NE</td>
<td>Univ. &amp; private</td>
<td>Comp. &amp; GI</td>
<td>FFS, salary + incentive</td>
<td>119</td>
<td>73,990</td>
<td>388/595</td>
</tr>
</tbody>
</table>
Multi-site NLP: Bringing text to the algorithms

Centralized NLP in a 4-site colonoscopy study (Ateev Mehrotra, PI)

Table 2. Performance of the NLP system for 13 colonoscopy quality metrics as measured by accuracy, average recall, precision, and F measure by study site. (1)

<table>
<thead>
<tr>
<th>Quality metric</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Sites</td>
<td>UPMC</td>
<td>UNC</td>
<td>CIE</td>
</tr>
<tr>
<td>Indications</td>
<td>0.93</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>Preservation</td>
<td>0.94</td>
<td>0.91</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>Attenuated</td>
<td>0.95</td>
<td>0.80</td>
<td>0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>Cecum</td>
<td>0.87</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Ileo</td>
<td>0.87</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Biopsy</td>
<td>0.99</td>
<td>0.97</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>Any villous/dysplastic chlns.</td>
<td>0.99</td>
<td>0.92</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>Location in colon</td>
<td>0.87</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(1) Study sites: UPMC = University of Pittsburgh Medical Center, UNC = University of North Carolina, CIE = Central Illinois Endoscopy, KPW = Kaiser Permanente Washington.

(2) These variables had >2 possible outcomes and were therefore evaluated using average recall, average precision, and average F measure, defined as the arithmetic mean of the respective performance metric calculated for each outcome category individually.
Multi-site NLP: Challenges and strategies

Many challenges in clinical NLP have little to do with NLP – Josh Denny

Issues:

- Local corpus assembly challenges
- 4 different environments (3 of 4 Epic®)
- Changes over time
- Re-assembly often needed
- De-identification & DUAs non-trivial
- Linguistics challenges multiply
- Textual diversity by site
- Changes over time
- Adapting existing NLP system non-trivial
- Can be successful
Summary

Many challenges in clinical NLP have little to do with NLP—Josh Denny

- “Secondary use” implies many challenges, often unanticipated
- Assembling the right clinical corpus takes effort, local expertise
- Clinical text availability/quality often less than ideal
- Challenges of interpretation, especially in multi-site context
- Simpler NLP tasks → higher likelihood of success
Abstract

Adapting clinical NLP methods for multi-site medical products research

David S. Carrell, PhD

Medical product clinical trials and postmarketing safety surveillance are increasingly coordinated across multiple institutional settings where secondary use of electronic health record (EHR) data makes large-scale ascertainment of outcomes more efficient. Many important outcomes are captured only in unstructured clinical narrative. Harmonizing information extracted from unstructured text in these settings entails challenges similar to those encountered when combining structured EHR data from geographically and institutionally diverse delivery systems. The adage emerging from these efforts, that “all data are local,” is at least as relevant to unstructured clinical data as it is to more widely used structured EHR data. This presentation will describe salient issues confronted when adapting and applying natural language processing (NLP) methods across multiple institutional settings. Seemingly simple tasks, such as assembling complete and representative clinical corpora, can be surprisingly challenging. Idiosyncratic characteristics of clinical documentation, including language usage, document structure, and content, makes the application of NLP methods in multi-site settings an endeavor that requires forethought and attention to detail. These and related issues will be illustrated with examples from recent NLP projects in several clinical domains, including a project now underway to extract from clinical progress notes information about patient-reported medication side effects.
Other issues …

– Character set clashes
– Good PDFs / Bad PDFs (scanned images of text)
– Dictated/transcribed notes (misunderstandings)
– Open source vs. proprietary software
– Machine learned algorithms and HIPPA PHI
– Co-reference resolution
– Negation
– Epic “CareEverywhere” (= “DataNowhere”)