Quantifying the Differences Between Human and Machine Performance in Extracting Data from Clinical Notes

James Harnett, Pfizer Senior Director Real World Data and Analytics
Bertrand Lefebvre, Optum Senior Data Scientist
Outline

• Introduction
• NLP objectives
• Extracting information from medical notes
• The mCRPC pipeline
• Evaluating the pipeline, evaluating annotators
• Conclusion
Rapid EHR Adoption in the US

Office-based Physician Electronic Health Record Adoption

Non-federal Acute Care Hospital Electronic Health Record Adoption


Office of the National Coordinator for Health Information Technology. 'Non-federal Acute Care Hospital Electronic Health Record Adoption,' Health IT Quick-Stat #47. dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php. May 2016.
Significant Interest from Regulators in Use of EHR

“The most useful source of knowledge will come from randomization in the context of clinical practice”

Former FDA Commissioner Dr. Robert Califf

"Let's make generating this evidence a lot easier and randomizing within the care system as much as we can."

Dr. Janet Woodcock, September 21, 2017
Life science companies will need EHRs

- **EHRs can be leveraged for:**
  - Supporting RW cost-effectiveness evidence generation
  - Enabling outcomes-based agreements between life science companies and payers
  - Filling post-marketing surveillance requirements
  - Matching patients to clinical trials
  - Simulating clinical trials
  - Conducting clinical trials in physician office using EHR for source data
  - Identifying new targets (esp. w/ clinico-genomic data)
But, much of the needed information is not readily available in EHRs

- EHR was not designed for research (clinical practice management/ billing)

- Much of the needed information is trapped in the unstructured part of the EHR

- Critical differences in patient, treatment and outcome characterization without unstructured data

- Manual review subject to errors, bias, time and costs
Growing interest in exploring natural language processing and machine learning methods

- Growing interest in application of NLP to medical/health records

Medline search on PubMed (3/10/2018) for articles on natural language processing + cancer + medical record/health record

Available at: https://blogs.wsj.com/cio/2017/01/25/a-new-tool-to-analyze-medical-records/
Significant need in Prostate Cancer

• Epidemiology (1)
  – American Cancer Society estimates 164,690 patients diagnosed with prostate cancer in 2018; about 1 in 9 men will develop prostate cancer in lifetime
  – Other than skin cancer, prostate cancer is the most common cancer among men in the United States, about 1 in 41 will die of prostate cancer

• Disease Course & Treatment (2)

1. Available at: https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html
2. Available at: http://www.cancernetwork.com/sites/default/files/1706arrar_figure1.png
3. Not yet recruiting, Recruiting, Enrolling by invitation, Active, not recruiting Prostate Cancer trials. Available at: https://www.clinicaltrials.gov/ct2/results?cond=prostate+cancer&recrs=b&recrs=a&recrs=f&recrs=d&age_v=&gndr=&type=&rslt=&Search=Apply
Project Overview

• To advance the availability of research ready data for prostate cancer covering the spectrum of care using NLP/machine learning methods

• Identify where (and where not) NLP/machine learning approaches are robust enough to support regulatory submissions and offer advantages over manual abstraction

• Share learnings to advance science of NLP/machine learning application and leveraging EHR for research and decision making
Specific objectives

• **Extract information allowing to reconstruct a prostate cancer patient’s therapeutic history**
  – Primary cancer(s)
  – Metastases
  – Cancer stage
  – Hormone therapy response (castration-resistance)
  – Prostate cancer histology
  – Performance status (ECOG)
  – Prostate biopsy results (Gleason score)
  – Reason for change in therapy
  – PSA levels (from structured)

• **Evaluate results**
What is Information Extraction?

Unstructured

CT-guided needle biopsy of the left iliac bone on June 4, 2013, revealed metastatic prostatic adenocarcinoma.

The PSA was 6.5 on 2/20/08. Biopsy revealed Gleason 4 + 3 = 7, 65% on the right and Gleason 4 + 3 = 7, 60% on the left.

Structured

- Cancer: prostate
- Histology: adenocarcinoma
- Metastasis: iliac bone, diagnostic date: 06/04/2013
- Gleason score: [7 (4+3), 7 (4+3)]

Other NLP tasks: Classification, topic modeling, summarization, QA, …
The data

Patients with at least 2 ICD codes of Prostate Cancer

- 33 million unique notes
- 291,797 unique patients
- 48 unique provider groups
What’s in a note?

A note is

– Patient-centric free-form *narrative*, rich in information but with lots of variability

A note is **not**

– *Clinical trial* – not designed to collect all relevant information for a specific study on a selected cohort of patients
– *Medical test* – not designed to be a diagnostic tool. Note ≠ X-ray image or lab result.
– *Scientific literature*
What’s in a note?

- **Uncertainty**: “This corresponds to a lytic abnormality adjacent to the end plate and in fact could represent degenerative arthritic changes versus possible bony metastatic disease.”

- **Ambiguity**: “I explained to the patient to start the treatment sooner rather than later does not interfere with the overall survival of patients with metastatic androgen-resistant prostate cancer”

- **Inaccuracy**: “Diagnosis: CRPC Stage IV castration sensitive prostate cancer”
What’s in a note?

- **Uncertainty**: “This corresponds to a lytic abnormality adjacent to the end plate and in fact could represent degenerative arthritic changes versus possible bony metastatic disease.”

- **Ambiguity**: “I explained to the patient to start the treatment sooner rather then later does not interfere with the overall survival of patients with metastatic androgen-resistant prostate cancer”

- **Inaccuracy**: “Diagnosis : CRPC Stage IV castration sensitive prostate cancer”

- “Caution: Voice recognition software (Dragon) was used to generate this note. Inadvertent and nonsensical statements could be made.”  
  
  \( \_-(ツ)\_/ \)
Annotating notes

CT-guided needle biopsy of the left iliac bone on June 4, 2013, revealed metastatic prostatic adenocarcinoma.

What is annotation?
– Add metadata telling what the “true” result should be for a set of tasks

What are annotations used for?
– Evaluate performance of algorithm(s) on a particular task
– For algorithms that “learn” from data, train them to perform the task

Annotation process
– Guidelines are established, defining the task and helping with consistency
– Each note from a randomly selected set is annotated by one or more people
– Annotations are “curated”, to reconcile divergences and establish final “truth”
Extraction pipeline

Pre-processing

Cancer + Metastases

Staging

Normalization

Cleanup
Sentence & word boundaries

Cancer/mets locations? Diagnosis dates?

Cancer stage?

Hormone response, Gleason score, ECOG score, …

Prostatic prostate Apr-2012 201204DD

Abdominal Pain
The primary symptoms of the illness include abdominal pain. The primary symptoms of the illness do not include fever, fatigue, shortness of breath, nausea, vomiting, diarrhea or dysuria. The current episode started more than 2 days ago (4-5 days). The onset of the illness was gradual. The problem has not changed since onset.
Evaluation metrics

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Impression</th>
<th>Stage</th>
<th>IIB</th>
<th>T2a</th>
<th>N0</th>
<th>prostate</th>
<th>adenocarcinoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>*</td>
<td>*</td>
<td>stage</td>
<td>* stage</td>
<td>stage</td>
<td>stage</td>
<td>loc_cancer</td>
</tr>
<tr>
<td>Predicted</td>
<td>*</td>
<td>*</td>
<td>stage</td>
<td>* stage</td>
<td>stage</td>
<td>stage</td>
<td>loc_cancer</td>
</tr>
</tbody>
</table>

Fraction of predictions that are correct: 
**accuracy** = \( \frac{TP+TN}{TP+TN+FP+FN} \)

= 10/12 for stage

Fraction of positives that are correct: 
**precision** = \( \frac{TP}{TP+FP} \) = \( \frac{3}{4} \) for stage

Fraction of true positives that are retrieved: 
**recall** = \( \frac{TP}{TP + FN} \) = \( \frac{3}{4} \) for stage

Precision/recall average: 
**F1** = \( \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \) = \( \frac{3}{4} \) for stage

<table>
<thead>
<tr>
<th>Tag</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>histology_pc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>stage</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>
Evaluating human annotators

Why evaluating annotators?
– Provide a “baseline” (or an upper bound)
– When crowdsourcing, select the “best” ones

How were annotators evaluated?
– Each annotator’s annotation used as a “prediction”
– Same metrics as for pipeline evaluation (other metrics exist, such as IAA)
– Same test set as for pipeline evaluation (120 notes)
– 2 annotators for each note
Human vs machine: side-by-side precision comparison

Precision

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pipeline</th>
<th>Annotator</th>
</tr>
</thead>
<tbody>
<tr>
<td>histology_pc</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>stage</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>score_ecog</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>loc_cancer_metastatic</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>hormone_response</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>loc_metastasis</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>date_cancer_dx</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>date_metastasis_dx</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Human vs machine: side-by-side recall comparison

Recall

- stage
- score_ecog
- histology_pc
- hormone_response
- loc_cancer_metastatic
- loc_cancer
- loc_metastasis
- date_cancer_dx
- date_metastasis_dx

Pipeline vs Annotator

Confidential property of Optum. Do not distribute or reproduce without express permission from Optum.
Human vs machine: takeaways

• **Pipeline tuned for Precision vs. Recall**
  – Pipeline generally has higher precision, but lower recall

• **High performance for Histology, Stage, and Performance Status (ECOG)**
  – Scores with > 0.97 precision and > 0.93 recall
  – Found primarily (if not exclusively) in unstructured data

• **Dates are a challenge for the existing pipeline**
  – in terms of both precision and recall
  – continued reliance on structured diagnosis codes for Prostate Cancer (short of additional pipeline refinement)

• **Other variables – like Hormone Response – are handled well by the pipeline**
  – opportunities to further supplement with structured data like change in PSA (lab results), combined with medication changes (written and filled medications)
Conclusion

• Built a **targeted** and **evaluated** extraction pipeline for prostate cancer studies

• Pipeline is modular, extensible and **generalizable** to other cancers

• Established an **end-to-end** infrastructure and process for IE and classification

• Possible next steps:
  – Improvements (e.g., for dates)
  – Generalize to other cancers
  – Extend scope (e.g. biomarkers)
Why is this important

- EHRs offer tremendous opportunities to match patients to RCTs or treatments at the point of care and improve efficiencies of RCTs
  - Patients in RCTs may have better survival (localized BrCa 7-yr survival: 91% vs. 82% guideline driven care and 76% for non-guideline driven care) (1)
  - Yet, only 3% of cancer patients enroll in RCTs and up to 60% sites don’t enroll a patient (2)
  - Cost of RCTs has increased by 64% between 2008 and 2011(3) and is $36,500 per patient(4)

- FDA has introduced a draft guidance for use of EHR for clinical investigation and RWE for medical devices – interest in pragmatic clinical trials in practice using EHR for potential label enhancements or expansion
  - Seeing examples of use synthetic controls based on EHR data (5,6)

- EHRs could enable alignment of outcomes and reimbursement
  - Life science companies can execute outcomes-based contracts
  - Providers can be reimbursed for outcomes vs. processes
  - Patient discounts for providing information on between visit outcomes/ adherence

2. Available at: https://www.forbes.com/sites/judystone/2015/01/06/how-can-we-encourage-participation-in-clinical-trials/#8c93f204d0c5
3. Available at: https://www.cuttingedgeinfo.com/2011/per-patient-clinical-trial-costs/ rises in clinical trial cost
5. Available e at: https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm559876.htm
6. Available at: https://www.focr.org/sites/default/files/pdf/RWE%20-%20Project%20PRE-MEETING%20DRAFT.pdf
Why is this important? (Cont’d)

• NLP/machine learning may address need to proactively, rapidly, and accurately abstract information from unstructured part of medical record
  – Current research supports potential advantages over variability in manual abstraction
  – Efforts in current project identified opportunities to scale up across 100k’s of prostate cancer patients that would be very difficult and costly to address with manual abstraction on ongoing basis
  – Also, identified areas where may need more work or may not be feasible can help avoid inefficiencies/refocus research efforts

• Future research will need to expand to other prostate cancer data needs and evaluate application to other tumors
Contributors @Optum

• *NLP team*: Kazuki Shintani, Maya Tydykov

• *Life Sciences*: Mike Crowley, Mike Sanky, Vidya Shankar

• *Privacy & Regulatory*: Cynthia Senerchia, Margaret Katana, Clara Gagnon

• *Terminology*: Jiby Joseph, Laleh Taheri, Jaqui Thomas, and more

• *Data Warehouse*: Bill Pjura, Russ Rogers
Appendix
Annotators evaluation results

- 12 annotators total
- Lots of variability from annotator to annotator
- Trends consistent with pipeline evaluation
  - Annotators agree and do well on stage, histology, castration-resistance
  - Annotators have less agreement and do less well on metastasis locations, dates
### Pipeline evaluation results

<table>
<thead>
<tr>
<th>Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>date_cancer_dx</td>
<td>0.70 (+0.02)</td>
<td>0.56 (+0.03)</td>
<td>0.63 (+0.03)</td>
<td>187</td>
</tr>
<tr>
<td>date_metastasis_dx</td>
<td>0.56 (-0.01)</td>
<td>0.34 (+0.02)</td>
<td>0.42 (+0.01)</td>
<td>121</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>0.89 (+0.14)</td>
<td>0.83 (-0.06)</td>
<td>0.86 (+0.05)</td>
<td>431</td>
</tr>
<tr>
<td>loc_cancer_metastatic</td>
<td>0.91 (-0.01)</td>
<td>0.84 (+0.03)</td>
<td>0.87 (+0.01)</td>
<td>167</td>
</tr>
<tr>
<td>loc_metastasis</td>
<td>0.75 (+0.00)</td>
<td>0.65 (+0.13)</td>
<td>0.70 (+0.09)</td>
<td>455</td>
</tr>
<tr>
<td>stage</td>
<td>0.98 (+0.00)</td>
<td>0.95 (+0.00)</td>
<td>0.96 (+0.00)</td>
<td>342</td>
</tr>
<tr>
<td>histology_pc</td>
<td>0.99 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.93 (+0.00)</td>
<td>192</td>
</tr>
<tr>
<td>date_hormone_refractory</td>
<td>0.50 (+0.25)</td>
<td>0.60 (+0.4)</td>
<td>0.55 (+0.33)</td>
<td>5</td>
</tr>
<tr>
<td>hormone_response</td>
<td>0.83 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.85 (+0.00)</td>
<td>99</td>
</tr>
<tr>
<td>score_ecog</td>
<td>0.97 (+0.01)</td>
<td>0.90 (+0.06)</td>
<td>0.93 (+0.03)</td>
<td>96</td>
</tr>
</tbody>
</table>
### Pipeline evaluation results

<table>
<thead>
<tr>
<th>Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>date_cancer_dx</td>
<td>0.70 (+0.02)</td>
<td>0.56 (+0.03)</td>
<td>0.63 (+0.03)</td>
<td>187</td>
</tr>
<tr>
<td>date_metastasis_dx</td>
<td>0.56 (-0.01)</td>
<td>0.34 (+0.02)</td>
<td>0.42 (+0.01)</td>
<td>121</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>0.89 (+0.14)</td>
<td>0.83 (-0.06)</td>
<td>0.86 (+0.05)</td>
<td>431</td>
</tr>
<tr>
<td>loc_cancer_metastatic</td>
<td>0.91 (-0.01)</td>
<td>0.84 (+0.03)</td>
<td>0.87 (+0.01)</td>
<td>167</td>
</tr>
<tr>
<td>loc_metastasis</td>
<td>0.75 (+0.00)</td>
<td>0.65 (+0.13)</td>
<td>0.70 (+0.09)</td>
<td>455</td>
</tr>
<tr>
<td>stage</td>
<td>0.98 (+0.00)</td>
<td>0.95 (+0.00)</td>
<td>0.96 (+0.00)</td>
<td>342</td>
</tr>
<tr>
<td>histology_pc</td>
<td>0.99 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.93 (+0.00)</td>
<td>192</td>
</tr>
<tr>
<td>date_hormone_refractory</td>
<td>0.50 (+0.25)</td>
<td>0.60 (+0.4)</td>
<td>0.55 (+0.33)</td>
<td>5</td>
</tr>
<tr>
<td>hormone_response</td>
<td>0.83 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.85 (+0.00)</td>
<td>99</td>
</tr>
<tr>
<td>score_ecog</td>
<td>0.97 (+0.01)</td>
<td>0.90 (+0.06)</td>
<td>0.93 (+0.03)</td>
<td>96</td>
</tr>
</tbody>
</table>
# Pipeline evaluation results

<table>
<thead>
<tr>
<th>Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>date_cancer_dx</td>
<td>0.70 (+0.02)</td>
<td>0.56 (+0.03)</td>
<td>0.63 (+0.03)</td>
<td>187</td>
</tr>
<tr>
<td>date_metastasis_dx</td>
<td>0.56 (-0.01)</td>
<td>0.34 (+0.02)</td>
<td>0.42 (+0.01)</td>
<td>121</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>0.89 (+0.14)</td>
<td>0.83 (-0.06)</td>
<td>0.86 (+0.05)</td>
<td>431</td>
</tr>
<tr>
<td>loc_cancer_metastatic</td>
<td>0.91 (-0.01)</td>
<td>0.84 (+0.03)</td>
<td>0.87 (+0.01)</td>
<td>167</td>
</tr>
<tr>
<td>loc_metastasis</td>
<td>0.75 (+0.00)</td>
<td>0.65 (+0.13)</td>
<td>0.70 (+0.09)</td>
<td>455</td>
</tr>
<tr>
<td>stage</td>
<td>0.98 (+0.00)</td>
<td>0.95 (+0.00)</td>
<td>0.96 (+0.00)</td>
<td>342</td>
</tr>
<tr>
<td>histology_pc</td>
<td>0.99 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.93 (+0.00)</td>
<td>192</td>
</tr>
<tr>
<td>date_hormone_refractory</td>
<td>0.50 (+0.25)</td>
<td>0.60 (+0.4)</td>
<td>0.55 (+0.33)</td>
<td>5</td>
</tr>
<tr>
<td>hormone_response</td>
<td>0.83 (+0.00)</td>
<td>0.87 (+0.00)</td>
<td>0.85 (+0.00)</td>
<td>99</td>
</tr>
<tr>
<td>score_ecog</td>
<td>0.97 (+0.01)</td>
<td>0.90 (+0.06)</td>
<td>0.93 (+0.03)</td>
<td>96</td>
</tr>
</tbody>
</table>
## Human vs machine: side-by-side

### Annotators

**Pipeline**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>date_cancer_dx</td>
<td>0.59</td>
<td>0.70</td>
<td>0.70</td>
<td>0.63</td>
</tr>
<tr>
<td>date_metastasis_dx</td>
<td>0.63</td>
<td>0.56</td>
<td>0.71</td>
<td>0.42</td>
</tr>
<tr>
<td>loc_cancer</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>loc_cancer_metastatic</td>
<td>0.93</td>
<td>0.91</td>
<td>0.91</td>
<td>0.87</td>
</tr>
<tr>
<td>loc_metastasis</td>
<td>0.67</td>
<td>0.75</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>stage</td>
<td>0.96</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>histology_pc</td>
<td>0.96</td>
<td>0.99</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>hormone_response</td>
<td>0.89</td>
<td>0.83</td>
<td>0.93</td>
<td>0.85</td>
</tr>
<tr>
<td>score_ecog</td>
<td>0.94</td>
<td>0.97</td>
<td>0.95</td>
<td>0.93</td>
</tr>
</tbody>
</table>

- 2 annotators
- Pipeline and annotators evaluated on the **same** test set