How to Advance Beyond ‘Regular Data’ with Text Analytics

Mike Dow
Director, Product Development, Health Catalyst

Carolyn Wong Simpkins, MD, PhD
Chief Medical Informatics Officer, Health Catalyst
Learning Objectives

- Define what natural language processing (NLP) and text analytics are.
- Demonstrate how to use NLP with clinical text.
- State the common uses of NLP at health systems today.
What Is Text Analytics?

The process of deriving high-quality information from text by applying NLP to transform text into data for analysis.
We believe that it will never be possible to anticipate every valuable piece of clinical information, so NLP should become an essential component of every healthcare analytics strategy.
80% of clinical data is locked away in unstructured notes.
Poll Question #1

Where is your organization’s clinical documentation stored? Select all that apply:

a) Paper records
b) Universal single EHR
c) Multiple EHRs for different sites/departments
d) Enterprise data warehouse (EDW)
e) Unsure and/or not applicable
A Clinician’s View of Heart Failure Information

- Current med list, current weight, current blood pressure (BP).

- Key events from most recent hospitalization (e.g., new cardiac events, discharge weight, new echo report with ejection fraction (EF) and wall motion, reason for decompensation).

- Latest EF and perhaps a graph of trend in EF over time.

- Current symptoms or complaints: weight gain, shortness of breath, peripheral edema.
Where Is This Information?

• Current med list, current weight, current blood pressure (BP).

• Key events from most recent hospitalization (e.g., new cardiac events, discharge weight, new echo report with ejection fraction (EF) and wall motion, reason for decompensation).

• Latest EF and perhaps a graph of trend in EF over time.

• Current symptoms or complaints: weight gain, shortness of breath, peripheral edema.
There’s an App for That: NLP

Past medical history, family history, social history and allergies have been reviewed with the patient and are correct per documentation.

CV history:
- CAD: 2% -LMCA, myocardial bridge LAD: LHC 10/16
- LVEF 60%
- Carotid artery disease - s/p prior CEA, followed by Dr. Oz, recent (10/16) carotid duplex 20-45% plaquing bilaterally
- HTN-controlled
- HLP: on statin therapy, lipids at goal on most recent lipid eval, tolerating without SE
- Sleep apnea
- Obesity
- Venous insufficiency, varicosities
- Tobacco use: participating in smoking cessation - current 2 cigarettes/day, goal to quit in next couple weeks all together

<table>
<thead>
<tr>
<th>PatientID</th>
<th>NoteID</th>
<th>Phrase</th>
<th>Low</th>
<th>High</th>
<th>Mid</th>
</tr>
</thead>
<tbody>
<tr>
<td>357750</td>
<td>340455</td>
<td>LVEF 60%</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>380888</td>
<td>661979</td>
<td>ejection fraction is at least 60-70%</td>
<td>60</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>363533</td>
<td>655151</td>
<td>ejection fraction of about 20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>738472</td>
<td>364993</td>
<td>EF of 45%</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>600607</td>
<td>145577</td>
<td>ejection fraction of greater than 65</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>854894</td>
<td>153107</td>
<td>LVEF of 35 - 40%</td>
<td>35</td>
<td>40</td>
<td>37.5</td>
</tr>
<tr>
<td>706071</td>
<td>564758</td>
<td>ejection fraction by visual inspection is 65</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Showing results for Collections:

ALL SOURCES

EF LVEF "ejection fraction"

e.g. defType=surround&fq(ljoin)...

<table>
<thead>
<tr>
<th>Report Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed (Completed)</td>
<td>9371</td>
</tr>
<tr>
<td>Progress Note - Generic (Progress Note - Generic)</td>
<td>5187</td>
</tr>
<tr>
<td>Cardiology Progress Note (Cardiology Progress Note)</td>
<td>4546</td>
</tr>
<tr>
<td>Cardiology Office/Clinic Note (Cardiology Office/Clinic Note)</td>
<td>3385</td>
</tr>
<tr>
<td>Office Note-Physician (Office Note-Physician)</td>
<td>2690</td>
</tr>
<tr>
<td>Discharge or Death Summary (Discharge or Death Summary)</td>
<td>2442</td>
</tr>
</tbody>
</table>
Showing results for Collections:

- ALL SOURCES

EF LVEF "ejection fraction"

fq=report_type:("Cardiology Progress Note")

Summary Reports 4,546 Patients 1,249 Report Types Cardiology Progress Note Tools

Showing a sample of 40 reports:

<table>
<thead>
<tr>
<th>#</th>
<th>Report ID</th>
<th>Patient</th>
<th>Name</th>
<th>Source</th>
<th>Date</th>
<th>Related IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>485476616E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>602070324E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>469009770E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>602668429E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>655556973E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>740628418E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>642284270E</td>
<td>Inpatient</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Stress test revealed anterior apical ischemia and poor ejection fraction of 28%. Recommend proceeding with left heart catheterization with possible PCI.

We have discussed risks, benefits, alternatives at length. She clearly understands and wishes to proceed.
Lexiscan stress test was performed which demonstrated no evidence of ischemia. There is global hypokinesis and ejection fraction 30%.

Impression: dilated cardiomyopathy

Recommendations: Beta-blockade, ACE inhibitor and diuretics
Lexiscan stress test was performed which demonstrated no evidence of ischemia. There is global hypokinesis and ejection fraction 30%.

Impression dilated cardiomyopathy

Recommendations. Beta-blockade ACE inhibitor and diuretics.
Showing results for Collections:

ALL SOURCES

EF LVEF "ejection fraction"

fq=report_type:("Cardiology Progress Note")

Returns in 17 milliseconds.

<table>
<thead>
<tr>
<th>#</th>
<th>Report ID</th>
<th>Patient</th>
<th>Name</th>
<th>Source</th>
<th>Date</th>
<th>Related IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>485476616E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>602070324E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>469090770E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>602668429E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>655556973E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>740628418E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>642284270E</td>
<td></td>
<td>Inpatient</td>
<td>A</td>
<td>[Date]</td>
<td></td>
</tr>
</tbody>
</table>
Ejection Fraction - validated  ·  Version 1

Created by mdow on 2017-05-28 08:35

```json
1 | "states": [
2 |  "name": "START",
3 |  "regex": "(?!)(?[^\n]+|\n[^\n])+(?!\n)+|(?=[\n]+),(?!([\n]+))|")",
4 |  "acceptStateName": "EF_WORD",
5 |  "rejectStateName": "END",
6 |  "comments": "Look for the ejection fracture word"
7 | },
8 |  "name": "EF_WORD",
9 |  "regex": "(([^\b])\(EF\)\L\b|EF\[\n\]I\e\[\n\]O\[\n\]jection|\n\]fraction|\n\]fraction\[\n\]fraction|\n\]fraction|\n\]fracture\)\b",
10 |  "acceptStateName": "EF_SENTENCE",
11 |  "rejectStateName": "END",
12 |  "checkWindowStateName": "START",
13 |  "checkWindowType": "INSIDE",
14 |  "expandMatchType": "CHECK_WINDOW",
15 |  "matchType": "SINGLE",
16 |  "comments": "Looks inside the chunked pieces for an ef word."
17 | },
18 |  "name": "EF_SENTENCE",
19 |  "regex": "(([^\b])\(EF\)\L\b|EF\[\n\]I\e\[\n\]O\[\n\]jection|\n\]fraction|\n\]fraction\[\n\]fraction|\n\]fraction|\n\]fracture\)\b",
20 |  "acceptStateName": "EF_WORD",
21 |  "rejectStateName": "END",
22 |  "checkWindowStateName": "START",
23 |  "checkWindowType": "INSIDE",
24 |  "expandMatchType": "CHECK_WINDOW",
25 |  "matchType": "SINGLE",
26 |  "comments": "Look for the ejection fracture word"
27 | }
```

Choose an input type:  🌐 Reports  ☐ Plain Text

Execution type: State Machine 2 (experimental)  ▼

Test Input:

Data Collection

득 COLLECTİONS

Enter up to 10 Report IDs for the reports you want to run (one Report ID per line).


Choose an input type:  
- Reports
- Plain Text

Execution type: State Machine 2 (experimental)

Test input:

Data Collection

COLLECTIONS

602070324E
Lexicon stress test was performed which demonstrated no evidence of ischemia.
There is global hypokinesis and ejection fraction 30%. Impression dilated
cardiomyopathy Recommendations. Beta-blockade ACE inhibitor and diuretics

Lexicon stress test was performed which demonstrated no evidence of ischemia.
There is global hypokinesis and ejection fraction 30%. Impression dilated
cardiomyopathy Recommendations. Beta-blockade ACE inhibitor and diuretics

Lexicon stress test was performed which demonstrated no evidence of ischemia.
There is global hypokinesis and ejection fraction 30%. Impression dilated
cardiomyopathy Recommendations. Beta-blockade ACE inhibitor and diuretics
"regex": "\(\{[\w]+\}\)((?\{[\w]+\}\)+)?\)
"acceptStateName": "EF_WORD",
"rejectStateName": "END",
"comments": "Look for the ejection fracture word"
},

"name": "EF_SENTENCE",
"regex": "(\{[\w]+\}\)((?\{[\w]+\}\)+)?\)
"acceptStateName": "EF_SENTENCE",
"rejectStateName": "END",
"checkWindowStateName": "START",
"checkWindowType": "INSIDE",
"expandMatchType": "CHECK_WINDOW",
"matchType": "SINGLE",
"comments": "Looks inside the chunked pieces for an ef word."}
Showing results for Collections:

ALL SOURCES

EF LVEF "ejection fraction"

fq=report_type:("Cardiology Progress Note")

Total Patients: 1,249

Birth Year

Gender

Female: 54%
Male: 46%
Send to Validation

You can send these query results to be validated using an existing study plan. If you want the number of results to be limited, you can enter a number to be allocated among all of your users. Selecting randomize will shuffle the current set, so that all sorting, including the default sorting or top scoring results will be randomized.

Study Plan:
Ejection Fraction Validation - demo (In Progress)

Total # of Annotations:
100

Randomize from Results

Allocation:
By Report (single report per annotation)
Send to Validation

You can send these query results to be validated using an existing study plan. If you want the number of results to be limited, you can enter a number to be allocated among all of your users. Selecting randomize will shuffle the current set, so that all sorting, including the default sorting on top scoring results will be randomized.

Study Plan:
Ejection Fraction Validation - demo (In Progress)

Total # of Annotations:
100

Allocation:
By Report (single report per annotation)
Send to Validation

You can send these query results to be validated using an existing study plan. If you want the number of results to be limited, you can enter a number to be allocated among all of your users. Selecting randomize will shuffle the current set, so that all sorting, including the default sorting on top scoring results will be randomized.

Study Plan:

Ejection Fraction Validation - demo (In Progress)

Total # of Annotations:

1

Randomize from Results

Allocation:

By Report (single report per annotation)
Send to Validation

You can send these query results to be validated using an existing study plan. If you want the number of results to be limited, you can enter a number to be allocated among all of your users. Selecting randomize will shuffle the current set, so that all sorting, including the default sorting on top scoring results will be randomized.

Study Plan:
Ejection Fraction Validation - demo (In Progress)

Total # of Annotations:
1

Randomize from Results

Allocation:
By Report (single report per annotation)
Select a study to begin annotation:

<table>
<thead>
<tr>
<th>#</th>
<th>Study Name</th>
<th>Study Manager</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Ejection fraction - May 16</td>
<td>Mike Dow</td>
<td>0%</td>
</tr>
<tr>
<td>14</td>
<td>Ejection Fraction Validation - demo</td>
<td>Mike Dow</td>
<td>40%</td>
</tr>
</tbody>
</table>
She is here for follow-up. She underwent a nuclear stress test that revealed evidence of inferior wall ischemia with an ejection fraction of 24%. The echocardiogram confirms LV dysfunction with an EF of 25%. The patient continues to complain of increasing dyspnea on exertion and occasional chest tightness.
She is here for follow-up. She underwent a nuclear stress test that revealed evidence of inferior wall ischemia with an ejection fraction of 24%. The echocardiogram confirms LV dysfunction with an EF of 25%. The patient continues to complain of increasing dyspnea on exertion and occasional chest tightness.
Validation – Precision and Recall

Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}

Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
## Validation – Precision and Recall

### Precision - are the values we have correct?

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reports</td>
<td>7,054,798</td>
</tr>
<tr>
<td>Initial EF Search</td>
<td>9,741</td>
</tr>
<tr>
<td>Reports Sent to EF State Machine</td>
<td>9,741</td>
</tr>
<tr>
<td>Number of Reports with extracted result</td>
<td>9,621</td>
</tr>
<tr>
<td>State Machine Results to be Reviewed</td>
<td>203</td>
</tr>
<tr>
<td>Review Process: subset of Notes, review. Goal is to get a 0/1 answer for the question &quot;are the extracted values correct?&quot; many reports have multiple values, for simplicity, no partial credit, any single invalid result renders the state machine's interpretation of that note incorrect.</td>
<td></td>
</tr>
<tr>
<td>Correct Results:</td>
<td>201</td>
</tr>
<tr>
<td>Calculated PPV:</td>
<td>99.0%</td>
</tr>
<tr>
<td><strong>Precision:</strong></td>
<td>201/203 = 99.0%</td>
</tr>
</tbody>
</table>

### Recall - did we get all of the right values?

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reports</td>
<td>7,054,798</td>
</tr>
<tr>
<td>Initial EF Search</td>
<td>9,741</td>
</tr>
<tr>
<td>Random Samples Size:</td>
<td>205</td>
</tr>
<tr>
<td>Review process: Review all 200 notes using validation study. Question 1: is there a valid EF value? Question 2: list the values.</td>
<td></td>
</tr>
<tr>
<td>n that had EF values</td>
<td>205</td>
</tr>
<tr>
<td>Compare the extracted values of the n Notes to the extracted values from the state machine. Score as 0/1. If if state machine extracted no values score as 0.</td>
<td></td>
</tr>
<tr>
<td>SM Results:</td>
<td>205</td>
</tr>
<tr>
<td>Calculated Sensitivity:</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Recall:</strong></td>
<td>205/205 = 100.0%</td>
</tr>
</tbody>
</table>
SELECT TOP 10
    Phrase,
    NumericLow,
    NumericHigh,
    NumericMid
FROM TextObservation.Extraction.LVEFBase

<table>
<thead>
<tr>
<th>Phrase</th>
<th>NumericLow</th>
<th>NumericHigh</th>
<th>NumericMid</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF 55</td>
<td>55.00</td>
<td>55.00</td>
<td>55.00</td>
</tr>
<tr>
<td>EF Calculated: 67</td>
<td>67.00</td>
<td>67.00</td>
<td>67.00</td>
</tr>
<tr>
<td>EF = 65</td>
<td>65.00</td>
<td>65.00</td>
<td>65.00</td>
</tr>
<tr>
<td>EF = 20-25</td>
<td>20.00</td>
<td>25.00</td>
<td>22.50</td>
</tr>
<tr>
<td>EF Calculated: 56.1</td>
<td>56.10</td>
<td>56.10</td>
<td>56.10</td>
</tr>
<tr>
<td>EF = 15-20</td>
<td>15.00</td>
<td>20.00</td>
<td>17.50</td>
</tr>
<tr>
<td>EF Calculated: 43.1</td>
<td>43.10</td>
<td>43.10</td>
<td>43.10</td>
</tr>
<tr>
<td>EF = 60</td>
<td>60.00</td>
<td>60.00</td>
<td>60.00</td>
</tr>
<tr>
<td>EF Calculated: 43.1</td>
<td>43.10</td>
<td>43.10</td>
<td>43.10</td>
</tr>
<tr>
<td>EF = 60-65</td>
<td>60.00</td>
<td>65.00</td>
<td>62.50</td>
</tr>
</tbody>
</table>
Using the Data
“No history of diabetes but family hx of heart failure.”
What You Need for NLP in Healthcare

- **Validation**
- **Extraction**
  - "Ejection fraction of 60 – 70%"
- **Context**
  - "Patient denies history of diabetes"
  - "Family history of diabetes"
- **Discovery & Synonyms**
  - NIDDM ~ diabetes
- **Search Engine**
Lessons Learned

1. Text analytics is about transforming text into data for analysis.
2. To use NLP with clinical text you need the five components in this pyramid:

   ![Diagram of pyramid with components: Search engine, Discovery & Synonyma, Context, Extraction, Validation]

3. Text analytics is commonly used at health systems:
   a) Information extraction and cohort identification.
   b) Quality reporting and specialties like oncology and obstetrics.
Poll Question #2

What pieces of information have you wanted or been asked for that can’t be found in discrete data fields? (Select all that apply.)

a) Symptoms (e.g., chest pain, shortness of breath, numbness, weakness)
b) Radiology findings (e.g., MRI, X-ray, ultrasound, CT)
c) Past medical history
d) Cardiology findings (e.g., EKG, stress test, echocardiogram)
e) Pathology findings (e.g., tumor type, stage)
f) Unsure or not applicable
Future Plans: Oncology

Extract key MRI findings from text report.

Search text notes for path diagnosis, stage.

Search for key Dx (seizures, headache) and neuro exam findings (deficits) in text notes.
Thank You