LIRE report text pre-processing of non-lumbar text

### 2020-12-23

**Goal:**

By building off of 2017-11-01 work, the goal is to isolate lumbar text in the findings and impression sections for both the index and non-index radiology reports.

Approach follows the following logic, which is in line with the annotation guidelines.

* Exclude all cervical findings.
* Exclude upper thoracic (T1-T6) findings.
* Keep lower thoracic (T7-T12) findings.
* Keep thoracic findings when levels are not explicitly stated.

We used a regular expression based programmatic approach

**Checking**:

We used a “greedy” string matching algorithm to check if cervical or thoracic mentioned. This is to ensure we catch all mentions of these terms in the reports. The regular expressions are

(cervi(c|x))|(C[1-7])

(thora(c|x))|(T[1-6])

We observed 127,086 out of 378,555 reports that had cervical and/or thoracic-related text

**Building Regex Process:**

*Examination:*

Sample 1000 reports from each site and manually extract the sentences that relate to thoracic and/or cervical in the finding and impression sections.

*Regex:*

Build regular expressions and assess if they can extract the text similar to manual approach.

Repeat *Extract* and *Regex* for another set of 1000 reports and so on.

**Regex Process:**

For a given text…

1. Remove leading and trailing white spaces
2. Add period to the end of text if not present
3. Temporarily replace decimal with "^"
4. Extract groups of cervical and upper-thoracic relate text using the following regex expression:

* **(?i)[^.]\*((cervi[cx].\*?\\.)|((C[1-7]|T[1-6])((?=\\.).\*?\\.|\\D.\*?\\.)))((?=[^.]\*(?=L[1-5]\\D|lumbar|thoraci[cx]|T[1-9]\\D|(T1[0-2])|cervi[cx]|C[1-7]\\D).\*?\\.)|(?=$))**
  + Logic:
    - Identify sentence that mentions: cervi[c|x], C[1-7], and/or T[1-6]
    - Extract that sentence and the following sentences until a sentence that mentions: cervi[c|x], lumbar, thoraci[c|x], C[1-7], L[1-5], and/or T[1-12] is identified
* Repeat Steps 1 and 2 until the end of the input text is reached

1. For these removed groups of sentences, we only consider groups of sentences if they are not edge-cases, so the first sentence cannot have any of the follow terms:

* lumbar
* L[1-5]
* T[7-12]
* T#-S#
* T[1-2](-| )weight
* T[1-2] signal
* T[1-2] (hypo|hyper|iso)intense
* T[1-2] shortening
* T[1-2] images
* (Sagittal|Axial|Coronal) T[1-2]
* NOTE: there are still other edge-cases that need to be considered. DO NOT consider this list all the edge-cases in the LIRE dataset.

1. Once we've identified groups of sentences that don't meet any of the edge-cases, we removed these sentences from the text
2. Return the decimal to the processed text
3. If the processed text is empty (i.e. “”), meaning everything was removed, then replace with a whitespace (i.e. “ )

**Results:**

* Used the Unannotated Index Reports (N = 254,546)
* Text is made up of Finding and Impression sections

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | Original Text – Thoracic Count | Original Text – Thoracic Percent | Original Text – Cervical Count | Original Text – Cervical Percent | Processed Text – Thoracic Count | Processed Text – Thoracic Percent | Processed Text – Cervical Count | Processed Text – Cervical Percent |
| All | 68956 | 27% | 3327 | 1.3% | 68356 | 27% | 1382 | 0.5% |
| 1 | 5025 | 31% | 668 | 4.2% | 5005 | 31% | 334 | 2.1% |
| 2 | 54891 | 26% | 1533 | 0.7% | 54457 | 26% | 729 | 0.4% |
| 3 | 5211 | 31% | 928 | 5.5% | 5102 | 30% | 274 | 1.6% |
| 4 | 3829 | 28% | 198 | 1.4% | 3792 | 27% | 45 | 0.3% |

**Result Examination:**

* Non-Zero Processed Text Thoracic Count and Percentage
  + “Thoracic” mentioned in sentence without indication of level.
  + Edge-Cases
  + “Thoracic” and lumbar-related information (i.e. “lumbar”, “L[1-5]”) mentioned in the same sentence.
    - Example: “thoracolumbar”
* Non-Zero Processed Text Cervical Count and Percentage
  + “Cervical” and lumbar-related information (i.e. “lumbar”, “L[1-5]”) mentioned in the same sentence

### 2017-11-01

**Motivation/Background:**

* For the LIRE cross-sectional imaging paper, we want to include primary care patients that have back pain, not primary care patients with cancer.
* Therefore, we opt to exclude patients whose radiology reports are whole spine MRIs
  + These patients are likely due to mets (unlikely for non-mets patients to receive whole spine MRI)
  + Use a CPT code approach for exclusion.
  + **Note:** This information might be useful in locating patients with mets for future research projects.
* However, we will keep patients whose radiology reports only have cervical and lumbar spine, as it is likely primary care patients would have both neck and back pain.
  + However, some lumbar reports are “merged” with non-lumbar reports and/or scout images.
* For these patients, our approach follows the following logic, which is in line with the annotation guidelines.
  + Exclude all cervical findings.
  + Exclude upper thoracic (T1-T6) findings.
  + Keep lower thoracic (T7-T12) findings.
  + Keep thoracic findings when levels are not explicitly stated.
* Goal of this note is to describe and document strategy we used.

**Pre-processing strategy/methodology:**

We used a regular expression based programmatic approach with site-specific parameters, with modules for checking, tagging, and cleaning.

**Checking**: We used a “greedy” string matching algorithm to check if cervical or thoracic mentioned. This is to ensure we catch all mentions of these terms in the reports. The regular expressions are

(cervi(c|x))|(c[1-7])

(thora(c|x))|(t[1-6])

**Tagging**: Since the checking is “greedy”, sometimes we may catch mentions of cervical/thoracic that are refers to comparison studies, imaging techniques, or a non-vertebral reference, such as:

* Cervicothoracic counter
* Cervical MRI is requested
* High T2 signal

These are phrases we want to KEEP in the report for NLP, but we don’t want them to trigger a potential manual review for cervical/thoracic.

**Cleaning**: Two types of cleaning, including segmentation and sentence removal. As described, sites 2 & 3 have very similar general approaches; however the exact regular expressions varied quite a bit by site (and not included here due to the highly iterative process).

The findings section of all sites were segmented into cervical, thoracic, and lumbar sections; the cervical sections were excluded. In addition, for sites 2 and 3, the impression section of all sites were segmented into cervical, thoracic, and lumbar sections; the cervical sections were excluded. Then, additional data cleaning by site:

**Site 1:**

* Removing sentences that contain both scout and cervic in the same sentence (reference to scout image of the cervix)
* Whenever the impression section starts off with "Spinal numbering is conventional", the report is dictated by levels, so remove segments that refer to cervix and/or upper thoracic

**Site 2:**

* Whenever the impression section is numbered (e.g. 1).. 2)...), the impression is dictated by spine levels, so remove segments that contain cervix and/or upper thoracic

**Site 3:**

* Whenever the impression section is numbered (e.g. 1).. 2)...), the impression is dictated by spine levels, so remove segments that contain cervix and/or upper thoracic

Initial number of reports that required pre-processing

|  |  |  |
| --- | --- | --- |
| **Site** | **Reports with cervic or C1-7 (%)** | **Reports with thorac or T1-6 (%)** |
| Group Health | 664 (4%) | 868 (6%) |
| Kaiser | 1415 (0.71%) | 15302 (8%) |
| Henry Ford | 911 (5%) | 1943 (12%) |

Post data cleaning:

|  |  |  |
| --- | --- | --- |
|  | # reports with cervical | # reports with thoracic |
| Site 1 | 107 | 1628 |
| Site 2 | 898 | 9674 |
| Site 3 | 395 | 1653 |

Sometimes, even though cervix/thoracic is mentioned, it is mentioned immediately in conjunction with the lumbar spine, usually in conjunction with degeneration. Examples include:

* degeneration of the cervical, thoracic, and lumbar spine
* degeneration cervical thoracic spine lumbar

These sentences should be included for NLP, but they should not trigger a manual review.

After excluding reports with patterns like that, there are still the number reports that I couldn't pre-process or tag for non-review reliably. These include

* + Reports with impression sections that are not numbered.
  + Reports with sentences in the findings that mentions cervical and/or thoracic with lumbar spine, but the patterns are more complicated
  + Reports where cervical/thoracic was mentioned, but they are normal e.g. “cervical alignment is normal”

|  |  |  |
| --- | --- | --- |
|  | # reports with cervical | # reports with thoracic |
| Site 1 | 94 | 126 |
| Site 2 | 878 | 1466 |
| Site 3 | 326 | 282 |