Summer Progress Report @Hyperfine, Inc.

Raziuddin Mahmood

UC Berkeley

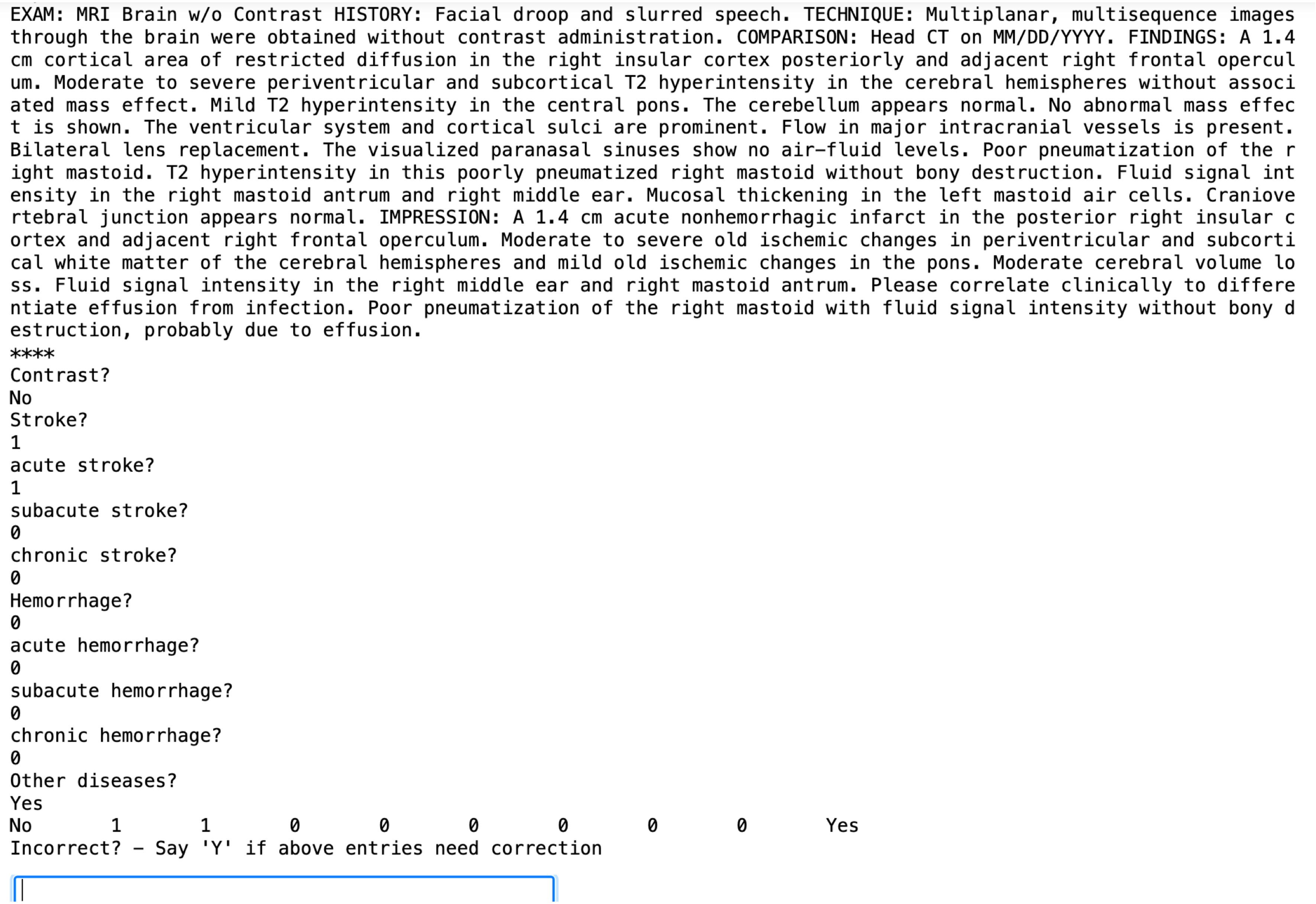
Tasks Attempted this Summer

- Task1:
  - Collating ground truth segmentation masks produced by radiologists with their original data files
- Task2:
  - Manual ground truth labeling of brain MRI reports
- Task3 (self-assigned):
  - Developing a new algorithm for automatic ground truth labeling and compared its performance to manual ground truth generated
- Task4 (self-assigned):
  - Assisting manual ground truth labeling with automated labeling for double-check of manual ground truth labeling

Task 1 : Collating Segmentations with Original MRI studies

- Overall goal:
    - The overall goal was to automate and expedite the recording process in the MRI segmentation ground truth generation workflow
- Problem:
    - Original MRI study data was organized into files in subdirectories
    - These files were assigned to clinicians to segment
    - The result of segmentation was deposited in a different subdirectory hierarchy
    - It was difficult to know which annotations were completed since the correspondence of segmentation files to original MRI files was not known unless it was recorded in an excel file.
    - Recording in the excel file was a tedious process of reading the file names in subdirectories for original and ground truthed data and manually making the correspondence between the two and noting it in the excel file
- Solution:
    - Wrote python code to navigate the directory hierarchy looking for potential corresponding pairs of segmentation masks done by clinicians and the raw MRI data and record the pairing automatically in an excel spreadsheet
    - Prepared a python script that can be run on any pair of original and segmentation root directories to produce the status file.
    - Used the status file to form the payment calculations for clinician annotation.
- Code:
    - Task1.ipynb already checked into Hyperfine GitHub
- Terminology learning
    - Neurologists were consulted for understanding the relevance of terms.
        - Example: Remote infarct == old infarct

Example – Manual labeling



```
EXAM: MRI Brain w/o Contrast HISTORY: Facial droop and slurred speech. TECHNIQUE: Multiplanar, multisequence images
through the brain were obtained without contrast administration. COMPARISON: Head CT on MM/DD/YYYY. FINDINGS: A 1.4
cm cortical area of restricted diffusion in the right insular cortex posteriorly and adjacent right frontal opercul
um. Moderate to severe periventricular and subcortical T2 hyperintensity in the cerebral hemispheres without associ
ated mass effect. Mild T2 hyperintensity in the central pons. The cerebellum appears normal. No abnormal mass effec
t is shown. The ventricular system and cortical sulci are prominent. Flow in major intracranial vessels is present.
Bilateral lens replacement. The visualized paranasal sinuses show no air-fluid levels. Poor pneumatization of the r
ight mastoid. T2 hyperintensity in this poorly pneumatized right mastoid without bony destruction. Fluid signal int
ensity in the right mastoid antrum and right middle ear. Mucosal thickening in the left mastoid air cells. Craniove
rtebral junction appears normal. IMPRESSION: A 1.4 cm acute nonhemorrhagic infarct in the posterior right insular c
ortex and adjacent right frontal operculum. Moderate to severe old ischemic changes in periventricular and subcorti
cal white matter of the cerebral hemispheres and mild old ischemic changes in the pons. Moderate cerebral volume lo
ss. Fluid signal intensity in the right middle ear and right mastoid antrum. Please correlate clinically to differe
ntiate effusion from infection. Poor pneumatization of the right mastoid with fluid signal intensity without bony d
estruction, probably due to effusion.
****
Contrast?
No
Stroke?
1
acute stroke?
1
subacute stroke?
0
chronic stroke?
0
Hemorrhage?
0
acute hemorrhage?
0
subacute hemorrhage?
0
chronic hemorrhage?
0
Other diseases?
Yes
No      1       1       0       0       0       0       0       0       Yes
Incorrect? - Say 'Y' if above entries need correction
```

Sample run in ManualGT.ipynb

Task2: Manual ground truth labeling of MRI reports

- Goal: Obtain a clean ground truth labels for images from their corresponding MRI radiology reports.
    - 12 possible labels:
        - Without contrast, with contrast, with and without contrast
        - Stroke, acute, subacute, chronic
        - Hemorrhage, acute, subacute, chronic
        - Other findings
- Problem:
    - Training machine learning models for recognition of findings in images requires training data that pairs images with finding labels seen in the image
    - Assigning labels to images manually by viewing them is a tedious process
    - Labeling can be expedited by using their corresponding reports
- Challenge:
    - The original report set came as a single excel file with the report buried in a column.
    - Annotating the report and hence the corresponding data was tedious for the selected attributes
        - 1 report was being annotated in 10 minutes. Most of the time was spent in extracting the text from excel column, placing it in a notepad, reading it, and then making the entries in the excel file. Data entry errors with misalignment of rows could cause errors in labeling
- Solution:
    - Developed Python code to show the report, manually record the required ground truth, and automatically save in excel file.
    - Significantly sped up the workflow
        - 1 report in 30 second-1min
    - **Produced manual ground truth labels for 12 labels for 600 reports = 7200 annotations**
- Code:
    - Project: MRIReportGT
    - Main notebook:ManualGT.ipynb
    - Documentation provided in code

Task 3: Developing a new method for automatic ground truth labeling from reports

- Even though this task was not given by HyperFine staff, I felt I could attempt this problem as large scale manual image labeling by visual examination or image or textual report data is not scalable because :
    - Requires clinical expertise
    - Tedious
    - Time consuming
    - Costly
    - May not be exhaustive, i.e. capture all findings present
    - Needs knowledge of all possible labels first
- Automatic labeling is an active area of research in clinical informatics with potential IP and publication implications.

Existing approaches to automatic labeling: Supervised machine learning

- Several researchers have tried using a supervised learning approach and built machine learning models
    - E.g. cheXpert from Stanford
    - Amazon comprehend medical
- Pros:
    - Once a machine learning model is trained, the extraction of labels is straightforward inference per report
    - Complex relationships between diseases mentioned in different sentences and orders can be learned through a machine learning approach.
- Cons:
    - The labels themselves are not complete, so only a handful of labels are recognized
        - CheXpert used originally 14 labels for chest X-rays while there are hundreds of possible findings in chest X-rays
    - Obtaining training data requires manual supervision to select relevant sentences from textual reports corresponding to the labels to train the machine learning model
    - The accuracy of the models is not very high, so the labels are at best suspect
    - Need to resort to full-scale manual verification even after automatic labeling, which only adds to the workload
    - Hence companies are preferring to go back to pure manual labeling which still requires some clinical expertise or familiarity with the textual terminology making the process of deriving labels slow.

Automatic image labeling from textual reports: Unsupervised detection of labels by NLP

- This approach promoted by IBM Research (https://patents.google.com/patent/US20160232658A1/en)
- Pros:
  - Combines natural language analysis of sentences with lexical pattern matching to group relevant phrases
  - Can detect both positive and negative findings (i.e lack of X, or normal)
  - Uses a large vocabulary lexicon to drive the detection.
  - State of the art performance with over 98% accuracy
  - Won the Homer Warner Award at AMIA 2020
- Cons:
  - Vocabulary-generation phase required clinician involvement, consensus process, and active learning using a domain learning assistant. This was the long tail process which took several months to year per disease and body region (e.g. chest X-ray lexicon has 250 findings and 16000 spoken variants of these findings. Discovering all of them was a semi-automatic process)
  - Negation detection errors still persist although small (2-3%)

Automatic labeling from textual reports: My Approach

- Uses a combination of NLP and lexical processing
- Main innovation is in rapid vocabulary/lexicon generation
- Improved negation pattern detection since sentence fragments are already captured. Pre and post negation patterns.
- Key ideas:
  - Vocabulary generation - Find patterns describing the target findings
  - Vocabulary detection – Detect presence of vocabulary terms in reports using the generated vocabulary
  - Negation detection – Rule out negative occurrences of findings from the labels
  - Sanity check – Analyze the statistics of generated labels and do some consistency checking.

Automated labeling algorithm- Rapid Vocabulary generation

- Semi-automatic process:
  - Record statistics of frequently occurring terms in reports
  - Analyze and curate the terms to classify the terms into the target finding types
- Statistical analysis of reports
  - N-gram analysis
    - Certain findings are spoken in a very consistent manner – these can be discovered through n-gram analysis
    - N-grams from 1 to 15 are usually sufficient to discover all such patterns
  - Phrasal grouping:
    - Use dependency parsing of sentences in reports to identify phrasal groups
- Record frequently occurring terms
  - present in at least 1% of the reports can be deemed important.
  - Reduces the manual verification load and avoids long tail
- Final produced list is through manual verification of a starting list of about 2000 or so phrases per modality/body part (e.g. brain MRI).
  - The final list will identify:
    - The pattern found in text
    - Its frequency
    - Its finding type
    - Its resolved finding name across all variants or as per machine learning model requirement

Example – n-gram pattern generation

- Only some patterns are useful, so frequency analysis will weed out the long tail:
- Example:
- *The pituitary gland and midline structures are unremarkable*
    - 4-gram :
        - The pituitary gland and
        - pituitary gland and midline
        - gland and midline structures
        - and midline structures are
        - **midline structures are unremarkable**
    - 3-gram
        - **The pituitary gland**
        - pituitary gland and
        - gland and midline
        - and midline structures
        - midline structures are
        - **structures are unremarkable**
    - 1-13 grams will contain some useful patterns

Example – Phrasal grouping

- Using Spacy for dependency parsing:
    - *A few small acute nonhemorrhagic infarcts in the left frontal and parietal lobes in the distribution of the left anterior cerebral artery with the largest one measuring about 1 cm.*
- Phrasal groups extracted from dependency parsing:
    - '*A few small acute nonhemorrhagic infarcts’*
    - *'the frontal lobes’*
    - *'the distribution of with’*
    - *'few small acute nonhemorrhagic infarcts in in .’,*
    - *'the left frontal and parietal lobes’,*
    - *'the left anterior cerebral artery’,*
    - *'the distribution’,*
    - *'the largest one measuring’,*
    - *'left frontal and parietal’,*
    - *'about 1 cm’*
- *Phrasal groups are better at grouping modifiers with findings. For example acuteness of infarct can be inferred from one of the phrasal groups above.*
    - *They are a way of associating context*

Automatic labeling – Detection in reports

- Vocabulary spotting for core findings (contrast, stroke, hemorrhage, others):
  - Once the vocabulary is obtained, the matching is using the same process used in vocabulary generation, i.e. n-gram generation or phrasal grouping to identify the phrases and match them as is.
    - Since the statistics were generated from the same collection, exact lookup is sufficient by repeating the same processing as in vocabulary generation
- Negation spotting:
  - Check for presence of pre and post negation terms within the sentence that contains the finding
- Fine-grained finding association:
  - Look for terms indicating the severity of the diseases
    - Acute, subacute or chronic (using their term variants from the vocabulary)
- Code developed:
  - AutoGT.ipynb in the project MRIReportGT

Example -1

- 1087 :EXAM: MRI BRAIN W/O CONTRAST HISTORY: R25.1, E11.9 Z79.4, G47.33 E11.42, M54.2, M54.81. Dizziness. COMPARISON: MM/DD/YYYY. TECHNIQUE: Multiplanar, multisequence imaging of the brain without contrast. FINDINGS: No diffusion or gradient signal abnormality. Midline structures are satisfactory. The craniocervical junction is within limits of normal. There is mild atrophy. Mastoid air cells and paranasal sinuses are grossly clear. There is underlying age-related change/small vessel ischemic disease. IMPRESSION: No acute abnormality. Underlying atrophy and age-related change.
- Detected findings:
  - Contrast? – No
  - Stroke? – No
  - Hemorrhage? – No
  - Other findings -> Yes

Example 2:

- 342 :MRI BRAIN WITHOUT CONTRAST, MM/DD/YYYY. HISTORY: Left shoulder numbness. COMPARISON: CT head, MM/DD/YYYY. TECHNIQUE: Sagittal and axial T1, axial T2, FLAIR, diffusion, and T2 gradient images were obtained. FINDINGS: The paranasal sinuses are clear. Mastoid air cells are clear. Nasopharynx is normal. Masticator spaces are normal. Orbital contents are normal. Extracranial soft tissue structures are unremarkable. Ventricles are normal in size. No hydrocephalus. No mass effect. No shift of midline. There is diffusion restriction in the right thalamus consistent with acute lacunar infarction. This measures 1.5 cm. No significant mass effect. No hemorrhage. No masses. IMPRESSION: 1. Acute lacunar infarction in the right thalamus measuring 1.5 cm. No mass effect or hemorrhage. 2. No masses.
- Contrast? –No
- Stroke? - Yes
- Acute stroke? – Yes
- Subacute stroke? – No
- Chronic stroke? – No
- Hemorrhage? –No
- Other findings – No (other than from stroke)

Results

- Testing automated ground truth labeling versus manual labeling on a labeled set of 601 reports

Finding	Number of Reports	P	TP	FP	Precision=TP/(TP+FP)	Recall=TP/P
Contrast match	601	601	565	36	0.94	0.94
Other findings	601	601	563	38	0.93	0.93
Stroke	601	593	543	6	0.98	0.91
Hemorrhage	601	60	42	14	0.75	0.7
Acute Stroke	601	585	420	4	0.99	0.72
Subacute Stroke	601	137	98	3	0.97	0.72
Chronic Stroke	601	213	141	11	0.93	0.67
Acute hemorrhage	601	28	4	3	0.57	0.14
Subacute hemorrhage	601	7	0	3	0	0
Chronic hemorrhage	601	35	14	1	0.93	0.4

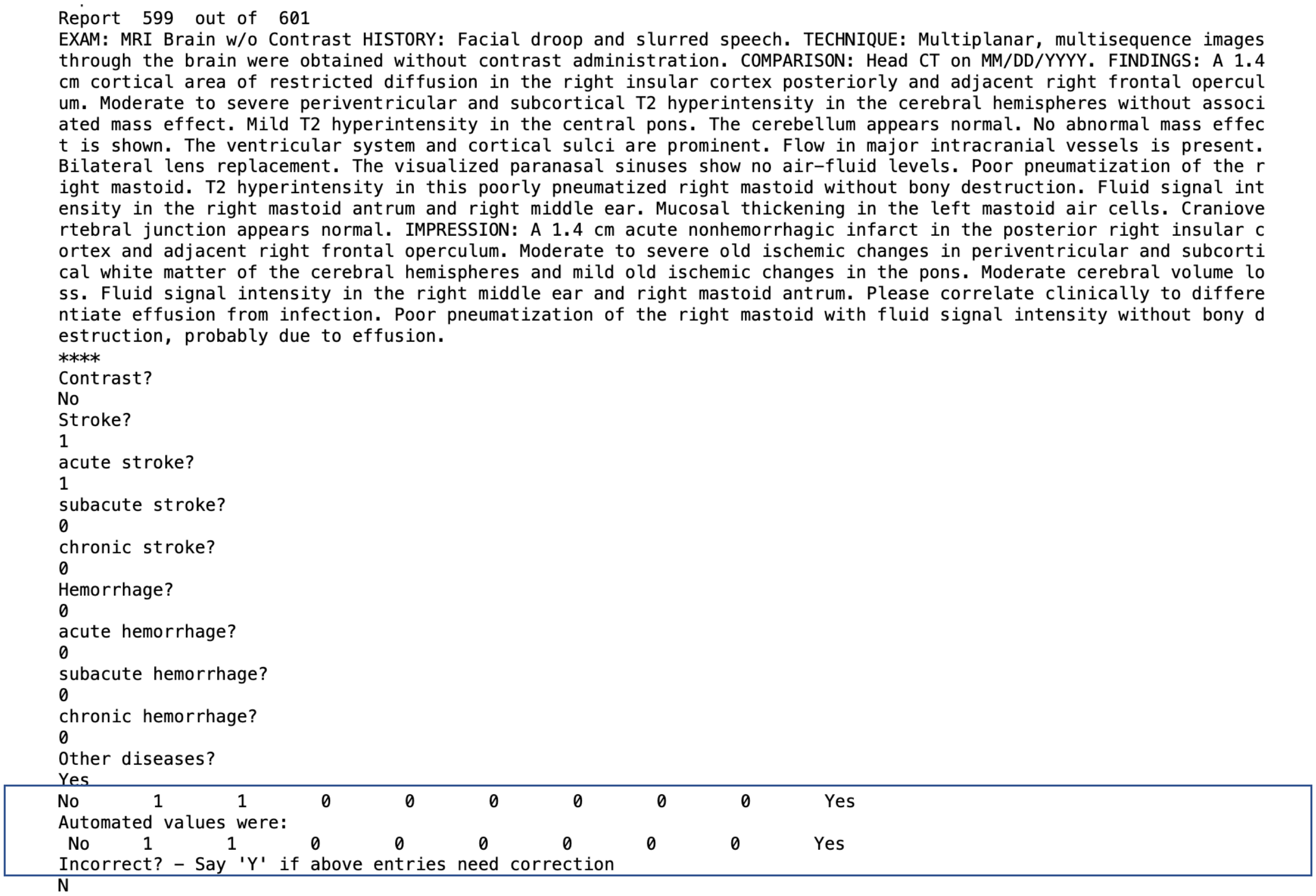
Task 4: Reliable manual labeling

- Since automated algorithm was able to run through all the reports, it could be used to validate the manual labeling or double checking during manual ground truth labeling
- In order to avoid the initial bias, the manual ground truth labeling should be done first.
- Then the labeler is shown the automated label
- The labeler can double check by re-reading the report (hint is not given by automated deliberately)
  - The labeler can choose to either agree with automated detection and correct it or ignore the automated detection

Task4: Augmenting manual ground truth labeling

- Developed Python code to present manual followed by automated detection and allowing for corrections of ground truth
  - Code available in ManualGT.ipynb
  - Documentation provided in the helper files in utils directory
    - Manualgt.py – has the code to show the reports and collect the labels
    - Sentenceprocess.py – has all NLP processing of reports to collect the vocabulary
    - Matchterms.py – implements most of the functions of automated vocabulary detection/labeling and performance evaluation

Augmentation example



```
Report  599  out of  601
EXAM: MRI Brain w/o Contrast HISTORY: Facial droop and slurred speech. TECHNIQUE: Multiplanar, multisequence images
through the brain were obtained without contrast administration. COMPARISON: Head CT on MM/DD/YYYY. FINDINGS: A 1.4
cm cortical area of restricted diffusion in the right insular cortex posteriorly and adjacent right frontal opercul
um. Moderate to severe periventricular and subcortical T2 hyperintensity in the cerebral hemispheres without associ
ated mass effect. Mild T2 hyperintensity in the central pons. The cerebellum appears normal. No abnormal mass effec
t is shown. The ventricular system and cortical sulci are prominent. Flow in major intracranial vessels is present.
Bilateral lens replacement. The visualized paranasal sinuses show no air-fluid levels. Poor pneumatization of the r
ight mastoid. T2 hyperintensity in this poorly pneumatized right mastoid without bony destruction. Fluid signal int
ensity in the right mastoid antrum and right middle ear. Mucosal thickening in the left mastoid air cells. Craniove
rtebral junction appears normal. IMPRESSION: A 1.4 cm acute nonhemorrhagic infarct in the posterior right insular c
ortex and adjacent right frontal operculum. Moderate to severe old ischemic changes in periventricular and subcorti
cal white matter of the cerebral hemispheres and mild old ischemic changes in the pons. Moderate cerebral volume lo
ss. Fluid signal intensity in the right middle ear and right mastoid antrum. Please correlate clinically to differe
ntiate effusion from infection. Poor pneumatization of the right mastoid with fluid signal intensity without bony d
estruction, probably due to effusion.
****
Contrast?
No
Stroke?
1
acute stroke?
1
subacute stroke?
0
chronic stroke?
0
Hemorrhage?
0
acute hemorrhage?
0
subacute hemorrhage?
0
chronic hemorrhage?
0
Other diseases?
Yes
No      1       1       0       0       0       0       0       0       Yes
Automated values were:
 No     1       1       0       0       0       0       0       0       Yes
Incorrect? - Say 'Y' if above entries need correction
N
```

Conclusions

- Label generation for deep learning – an interesting research problem
- Rich vocabulary generated for findings- Knowledge alone may be useful for future work
  - 322 terms
    - 39 No contrast patterns
    - 36 Contrast patterns
    - 12 With and without contrast patterns
    - 45 stroke patterns
    - 16 hemorrhage patterns
    - 174 Other finding patterns
- Automated labeling giving good results for stroke patterns
- Some deeper semantics modeling needed for improvement in accuracy for cases, particularly for acute hemorrhage
- Phrasal grouping errors with modifiers account for most of the recall issues
- Further improvement is possible with a few more pieces of information added.