

# Using Natural Language Processing (NLP) to Identify Lines and Devices in Portable Chest X-Ray (CXR) Reports

VA Health Services Research & Development Cyber Seminar  
March 29, 2012

Presenters:

**Mary K. Goldstein, MD (Principal Investigator)**

*Director, Geriatrics Research Education and Clinical Center (GRECC) at VA Palo Alto Health Care System*

*Professor of Medicine (Center for Primary Care and Outcomes Research), Stanford University*

**Dan Y. Wang, PhD**

*Research Health Science Specialist, VA Palo Alto Health Care System*

*VP of Technology, Medcivic LLC*

**Tammy S. Hwang, BA**

*Research Health Science Specialist, VA Palo Alto Health Care System*



## Acknowledgments

This study was undertaken as part of the VA Health Services Research & Development (VA HSR&D) Consortium for Healthcare Informatics Research (CHIR) Translational Use Case Project (TUCP) grant HIR 09-007 (PI: Goldstein), a component of CHIR HIR 08-374 (PI: Samore).

We also made use of report extraction facilities and a secure server workspace provided by the VA HSR&D Informatics and Computing Infrastructure (VINCI).



**VA Consortium of Healthcare  
Informatics Research**



**VA Informatics and Computing  
Infrastructure**

Views expressed are those of the presenters and not necessarily those of the Department of Veterans Affairs

## Investigator Team

Dan Y. Wang, PhD, VA Palo Alto Health Care System

Daniel L. Rubin, MD, MS, Stanford University

Tammy S. Hwang, BA, VA Palo Alto Health Care System

Dallas A. Chambers, BS, VA Palo Alto Health Care System

Justin G. Chambers, BS, VA Palo Alto Health Care System,  
Medcivive LLC

Brett R. South, MS, VA Salt Lake City Health Care System

Mary K. Goldstein, MD, MS, VA Palo Alto Health Care System,  
Stanford University

Additional subject matter expertise from Matthew Samore,  
MD, PhD of CHIR project and other experts

Statistical consultants: Shuying Shen, PhD, and Andrew Redd,  
PhD

# Poll

Please help us understand our audience.

Select all that apply:

- (A) I am primarily interested in the potential application of this tool
- (B) I am primarily interested in the underlying technology/  
technical background of this tool
- (C) I do clinical work as a licensed health professional at the VA
- (D) A substantial part of my work includes informatics
- (E) Quality assessment/measurement is a substantial part of my  
work
- (F) Research is a substantial part of my work

# Goals of the Session

At the end of this seminar, the participants should be able to...

- Explain steps involved in conducting a project extracting information from free-text of VA Electronic Health Records
- Describe how to use an annotation tool (Knowtator) to create a reference standard
- Understand a natural language processing (NLP) technique that works for information extraction from chest x-ray (CXR) reports

# Outline

- Background
- Methods
- Results
- Comments
- Audience questions and discussion

# Background

# Structured and Unstructured Data

- Electronic health records have extensive information important for patient care, quality assessment, quality improvement, and epidemiologic surveillance
  - structured data elements
  - unstructured data (free text)
- Consortium for Healthcare Informatics Research (CHIR)
  - Translational Use Case Projects



# Portable CXR in ICU

- Patients in Intensive Care Units (ICU) often have medical devices inserted
  - Such devices can lead to complications such as blood-borne infections, increased morbidity, and increased cost
  - Line/device related infections have been correlated with length of time of presence and line/device type (Goetz, 1998; Tokars, 2007)
    - Hospitals are often required to report a daily count of patients with specific lines/devices in place
- Inserted devices are radio-opaque
  - usually visible in Portable chest X-ray (CXR) images
  - usually documented in free-text CXR reports

# Our Aims in Design

- Use Natural Language Processing (NLP)
  - avoid manual chart review
  - extract detailed information about devices
  - potentially enable infection surveillance and epidemiological research
- Focus on the part of the radiology report that describes what the radiologist sees on the x-ray (not clinical history entered by the ordering provider)
- Initially focus on accuracy of information extracted from within a single report as compared to human reader of that report
  - Reserve as future work the comparison of reports for same patient through time

# Methods

# Overview of Methods

- Specify the report-types of relevance
  - We decided to focus on portable chest x-ray (CXR) reports for patient in ICU at time the study was done
    - In future work expanding to all CXR reports for patient during the same admission
- Specify what information should be identified in each report
- Identify a source of documents
  - We worked in VA VINCI secure computing environment
  - Document selection
- Develop a reference standard of annotated reports that are not shared with NLP developers
- Develop NLP code to process the text
  - Separate set of documents made available to NLP developers to train the system
- Compare the output from the NLP with the reference standard

# Staged Project

- Early stage project
  - Evaluated with 90 reports (previously reported)
- More recent broader evaluation and improvements
  - Evaluated with 500 reports
- Ongoing and future work
  - Move beyond one CXR report at a time to link reports for same patient through time during acute hospitalization

# Identifying Records

- Identify EHR free-text records to extract
  - Use structured data to identify which reports are relevant
  - Procedure names and CPT codes
    - The CPT code for “Chest X-ray” had 1,749 distinct procedure names
    - Many reports had Null procedure code, with procedure names that appeared to be CXR
    - We developed list of procedure names for SQL code search
      - At present no way to know if we captured every single portable CXR report
        - » Not essential for purpose of this project
        - » A major issue if intent were to apply to every relevant case in VA
      - Available to other VA staff who wish to use it and improve it
        - » Perhaps libraries of such lists could be made available as shared resource

# Specifying Terms to Extract

- How we identified device/line terms to extract
  - Resolved questions about categories of items to include in discussion with subject matter experts
    - For purpose of this study, included lines and tubes in the chest, but not hardware in the spine and not heart valves
  - Device terms obtained from UMLS (Bodenreider, 2004) and RadLex (Rubin, J Digital Imaging 2008)
    - Note that standardized terminologies do not capture the actual language used in many reports
  - Reviewed actual CXR reports to identify additional terms

# Developing the Reference Standard

- Reference Standard: A set of carefully annotated reports against which we can measure the accuracy of the NLP algorithm
- Annotation training
- The annotators refer to a written annotation guideline
  - Includes a lines/device appendix with list of terms
  - Developed in a process of annotating to understand what issues are confusing for annotators
- The annotation guideline is shared with the NLP Developer
  - So everyone is using the same rules
- The Reference Standard set of documents is blinded to the NLP developer
  - So that the evaluation will test the NLP on documents not used to train it
- Span of text
  - every character in the documents is numbered, starting with 1 as the first character in the report. the “span” of text is the character positions of a particular portion of the text.
  - Example
    - report text: “the tube is present.”
    - the word “tube” has span from 5 to 8



## Our Reference Standard

- 500 randomly selected portable CXR exams obtained in the ICU between 1/1/2006 to 1/1/2009 from 46 randomly selected, geographically dispersed, Veteran Affairs (VA) medical centers cross the country
  - excluded sites with small number of ICU patients
- Reports stratified by medical center and length of Stay (LOS). LOS groups included short LOS (1-3 days), medium LOS (4-7 days), and long LOS (8-10 days) in a 40:40:20 ratio

# Evaluation of NLP accuracy

- Compare NLP output with reference standard for each line/device and insertion status term
- TP (True Positive): found in both reference standard and NLP output
- FN(False Negative): found only in reference standard
- FP (False Positive): found only in NLP output
- Compute standard overall measures for information retrieval:
  - Precision =  $TP/(TP+FP)$
  - Recall =  $TP/(TP + FN)$
- Sebastiani ACM 2002

# Poll

Select all that apply:

- (A) I have experience using text annotation tools
- (B) I have some familiarity with the concept of text annotation
- (C) I am not familiar with text annotation other than what I have heard in this talk

## **Annotation of Reference Standard:**

Team of annotators used Knowtator Tool, (Ogren, 2006) a plug-in of open-source knowledge acquisition tool Protégé ([protégé.stanford.edu](http://protégé.stanford.edu))

- Knowtator uses an annotation schema
- Annotations done by point-clicking or drag-and-drop of the mouse over terms or phrases which are visually highlighted
- 4 classes: Device/Line, Device/Line Status, Laterality and Device/Line Quantity
- Each device term was associated with an insertion status term and assigned an insertion status
- For comparison with NLP output, Knowtator provides XML output text file

Knowtator Classes Instances Slots

fast annotate  Device/Line  Device/Line\_Status  Laterality  Device/Line\_Gity  PHI  Document\_Quality quit

annotation schema

- Device/Line (6)
- Device/Line\_Status (4)
- Laterality (3)
- Device/Line\_Gity
- PHI
- Document\_Quality

text source: Testpt1Series1Day7.txt filter: show all

Impression:

- Interval placement** of a **nasogastric tube** with the distal tip in the abdomen and not visualized. Stable tracheostomy tube and right-sided dialysis catheter and PICC line. Stable right upper quadrant surgical drain. Stable left axillary line.
- Low lung volumes with minimal left basilar opacity likely representing atelectasis. Stable cardiomeastinal silhouette.

Interval placement

span edit: delete

annotated class

Device/Line\_Status (4)

slots of annotated class (3 values)

A. Device/Line Status Determination

Inserted

B. Device/Line Reference

nasogastric tube

C. Annotator Confidence

Instances of 'Device/Line\_Status'

Impression:

1. Interval placement of a nasogastric tube with the distal tip in the abdomen and not visualized. Stable tracheostomy tube and right-sided dialysis catheter and PICC line. Stable right upper quadrant surgical drain. Stable left axillary line.
2. Low lung volumes with minimal left basilar opacity likely representing atelectasis. Stable cardiomeastinal silhouette.



Knowtator Classes Instances Slots

**fast annotate**  Device/Line  Device/Line\_Status  Laterality  Device/Line\_Qty  PHI  Document\_Quality quit

annotation schema

- Device/Line (6)
- Device/Line\_Status (4)
- Laterality (3)
- Device/Line\_Qty
- PHI
- Document\_Quality

text source: Testpt1Series1Day7.txt filter: show all

Impression:

- Interval placement of a nasogastric tube with the distal tip in the abdomen and not visualized. Stable tracheostomy tube and right-sided dialysis catheter and PICC line. Stable right upper quadrant surgical drain. Stable left axillary line.
- Low lung volumes with minimal left basilar opacity likely representing atelectasis. Stable cardiomeastinal silhouette.

Interval placement

span edit: delete

annotated class

Device/Line\_Status (4)

slots of annotated class (3 values)

A. Device/Line Status Determination

Inserted

B. Device/Line Reference

nasogastric tube

C. Annotator Confidence

Instances of 'Device/Line\_Status'



## **Annotation Procedure:**

- 25 batches of 20 reports each
- Each batch of reports annotated by 3 annotators: 2 who independently annotate it first, then a 3<sup>rd</sup> as an adjudicator to decide on possible differences
- Roles rotated for different batches
- If no agreement could be reached by team of 3, refer to subject matter expert
- Annotation guideline and device appendix

## Evaluation:

- 3 rounds of testing
- Round 1: NLP vs. original annotations
- Round 2: NLP vs. corrected annotations
- Round 3: Corrected NLP vs. corrected annotations

## NLP Algorithm: Base GATE pipeline

- Base toolset: GATE (General Architecture for Text Engineering)
  - Particular modules: sentence splitter, English tokenizer, Gazetteer, JAPE (Java Annotation Patterns Engine), POS (Part of Speech) Tagger.
- GATE Pipeline Functions:
  - identify sentence boundaries
  - identify custom token types via the Gazetteer: devices, device fragments, terms indicating insertion, removal, presence
- Cunningham et al, App Proc Mtg for Comp Ling, 2002; 168-175

## NLP Inference Logic:

- After GATE pipeline, custom Java modules to associate statuses with devices
  - for each line/device term found, find nearest insertion status term within the same sentence subject to refinements.
  - Example: "Since previous exam, the endotracheal tube remained in satisfactory position."
- "endotracheal tube" found as device term
- "in satisfactory position" found as an insertion status term of type "present"
- Custom module makes association and inference

## NLP Training Process:

- Incorporation of more complex sentences/text patterns:
  - single sentence with multiple insertion status terms
  - multiple devices with single insertion status terms
  - multiple devices with multiple insertion status terms
- Each iteration involves expanding the NLP rules and adding terms
- First promising results reported in AMIA 2010
  - Using 90 annotated reports, achieved over 90% precision and recall in identifying lines/devices and their associated insertion statuses
    - (Rubin et al, AMIA Annu Symp Proc 2010; 692-6)

## NLP Algorithm Refinements:

### Clause Boundaries Trump Proximity

Example: "Nasogastric tube redemonstrated, the IJ catheter in the main pulmonary artery removed."

- The status associated with "IJ catheter" is "removed", signifying removal, and not "redemonstrated" signifying presence
  - "redemonstrated" is separated from "IJ catheter" by the clause boundary ",".

## **NLP Algorithm Refinements:**

- Precedence Order of Statuses

Example: "Interval placement of the left subclavian line is seen with its tip projecting over the cavoatrial junction."

- A single device term "subclavian line"
- Two insertion status terms "Interval placement", "seen"
- Use precedence of recent insertion (and removal) status terms over presence status terms for resolution

## Recent NLP Algorithm Refinement

- Device Fragments:

Example: “Interval removal of chest and nasogastric tubes.”

- The words “chest” and “nasogastric” are identified as device fragments, but by themselves do not necessarily indicate a device.
- The presence of the plural term “tubes” and the conjunction “and” does form a distinct text pattern recognized by our NLP Algorithm.



## Evaluation of NLP accuracy

- Comparison with reference standard for each device and insertion status term
- Overall Measures as noted above:
  - Recall =  $TP / (TP + FN)$
  - Precision =  $TP / (TP + FP)$

# NLP Training Tool, Debugger Platform

Font: A+ / a-

## Chest X-Ray Reports: Automated Identification of Inserted Devices

### CXR Recall and Precision

	Total	Total(NLP)	TP	FN	FP	Recall	Precision
device	74	84	73	1	11	0.986	0.869
status-insertion	3	4	3	0	1	1.0	0.75
status-removal	5	4	4	1	0	0.8	1.0
status-presence	65	76	64	1	12	0.984	0.842
status-neg. of presence	0	0	0	0	0	NaN	NaN

INS
REM
PRE
ABS

All 
W/Err 
W/O Err

1 2 3

### Analysis Summary

File	Show Mult	XML exist	TP	FN	FP	TPPM	TPEM
Testpt1 Series1Day16.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	10	0	0	3	7
Testpt1 Series1Day17.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	10	0	2	3	7
Testpt1 Series1Day18.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	8	0	0	2	6
Testpt1 Series1Day19.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	4	0	0	0	4
Testpt1 Series1Day2.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	4	0	0	0	4
Testpt1 Series1Day3.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	8	0	4	1	7
Testpt1 Series1Day4.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	8	0	2	0	8
Testpt1 Series1Day5.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	10	2	6	1	9
Testpt1 Series1Day6.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	10	0	0	1	9
Testpt1 Series1Day7.txt	<input checked="" type="checkbox"/>	<u>Yes</u>	12	0	0	1	11

# NLP Training Tool, Debugger Platform

Font: A+ / a-

## Chest X-Ray Reports: Automated Identification of Inserted Devices

### CXR Recall and Precision

	Total	Total(NLP)	TP	FN	FP	Recall	Precision
device	74	84	73	1	11	0.986	0.869
status-insertion	3	4	3	0	1	1.0	0.75
status-removal	5	4	4	1	0	0.8	1.0
status-presence	65	76	64	1	12	0.984	0.842
status-neg. of presence	0	0	0	0	0	NaN	NaN

Sub    INS   REM   PRE   ABS

All            

W/Err           

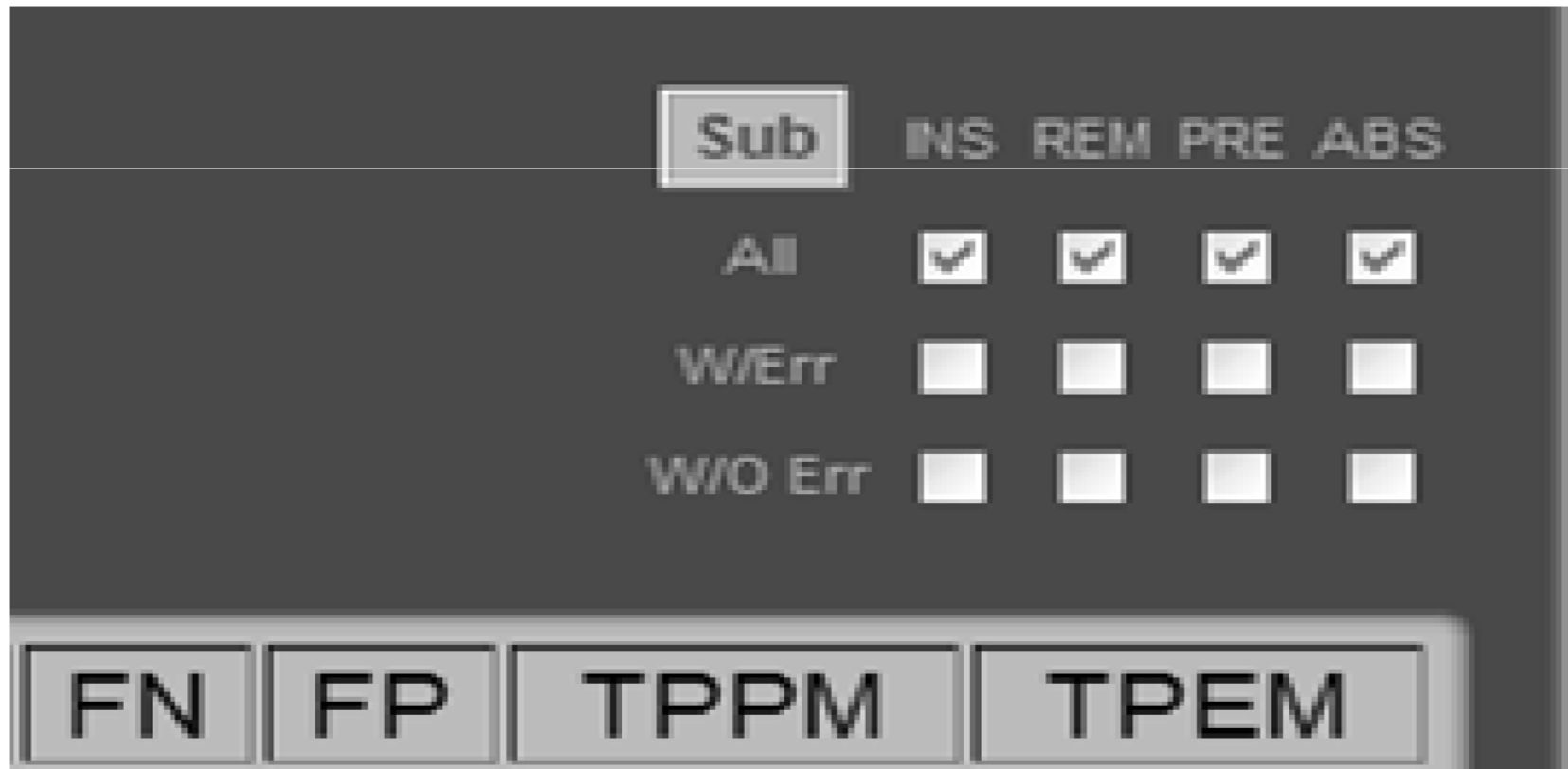
W/O Err

1 2 3    Save To Excel

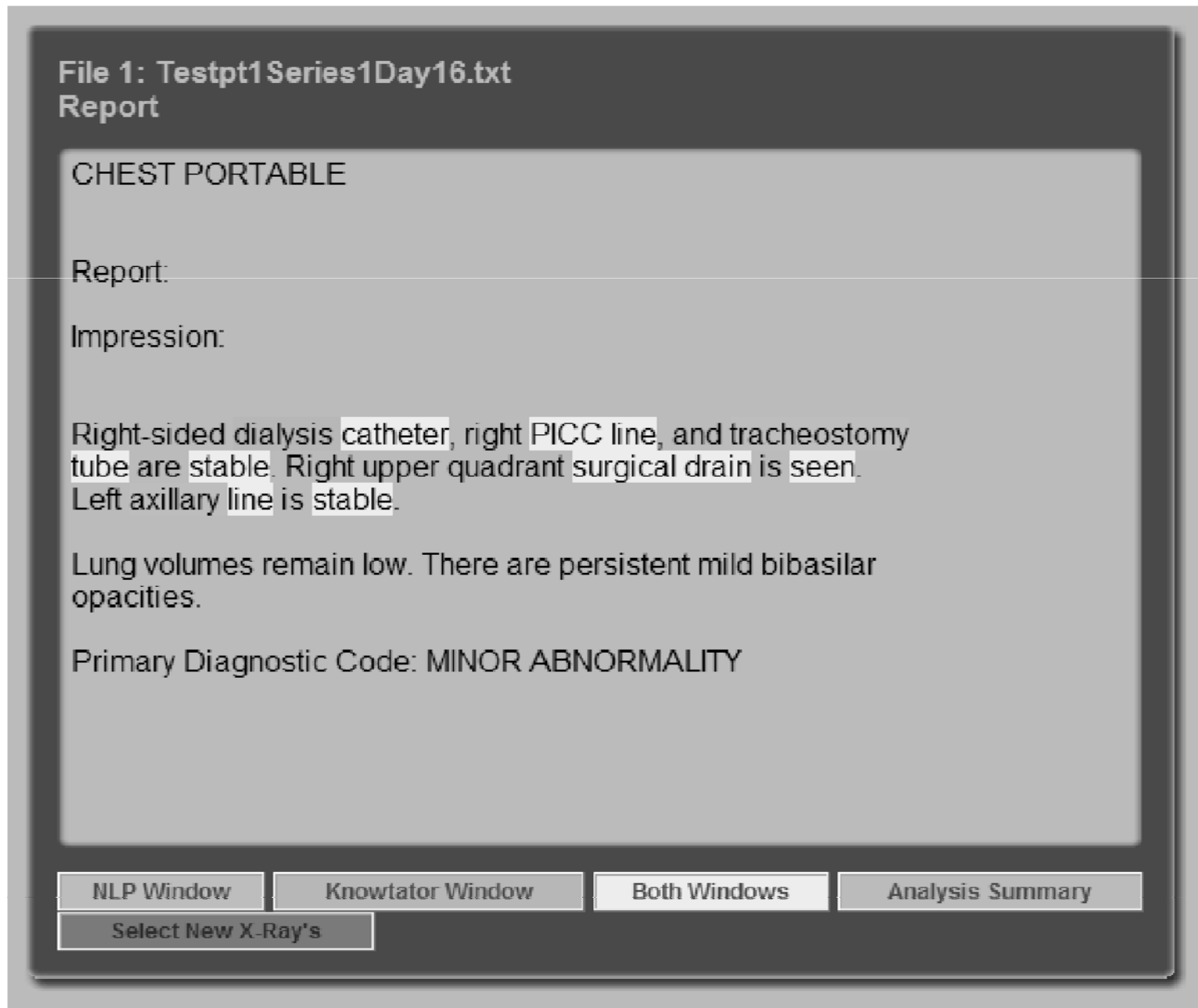
### Analysis Summary

File	Show Mult	XML exist	TP	FN	FP	TPPM	TPEM
Testpt1 Series1Day16.txt	<input checked="" type="checkbox"/>	Yes	10	0	0	3	7
Testpt1 Series1Day17.txt	<input checked="" type="checkbox"/>	Yes	10	0	2	3	7
Testpt1 Series1Day18.txt	<input checked="" type="checkbox"/>	Yes	8	0	0	2	6
Testpt1 Series1Day19.txt	<input checked="" type="checkbox"/>	Yes	4	0	0	0	4
Testpt1 Series1Day2.txt	<input checked="" type="checkbox"/>	Yes	4	0	0	0	4
Testpt1 Series1Day3.txt	<input checked="" type="checkbox"/>	Yes	8	0	4	1	7
Testpt1 Series1Day4.txt	<input checked="" type="checkbox"/>	Yes	8	0	2	0	8
Testpt1 Series1Day5.txt	<input checked="" type="checkbox"/>	Yes	10	2	6	1	9
Testpt1 Series1Day6.txt	<input checked="" type="checkbox"/>	Yes	10	0	0	1	9
Testpt1 Series1Day7.txt	<input checked="" type="checkbox"/>	Yes	12	0	0	1	11

# NLP Training Tool, Debugger Platform



# NLP Training Tool, Debugger Platform



# NLP Training Tool, Debugger Platform

File 1: Testpt1Series1Day16.txt  
Report

CHEST PORTABLE

Report:

Impression:

Right-sided dialysis catheter, right PICC line, and tracheostomy tube are stable. Right upper quadrant surgical drain is seen. Left axillary line is stable.

Lung volumes remain low. There are persistent mild bibasilar opacities.

Primary Diagnostic Code: MINOR ABNORMALITY

**Analysis**

Device	Start	End	Status	Start	End	Inference
catheter	92	100	stable	151	157	Present
PICC line	108	117	stable	151	157	Present
tracheostomy tube	123	146	stable	151	157	Present
surgical drain	180	194	seen	198	202	Present
axillary line	215	228	stable	232	238	Present

**Knowtator**

Device	Start	End	Status	Start	End	Inference
dialysis catheter	83	100	stable	151	157	Present
PICC line	108	117	stable	151	157	Present
tube	142	146	stable	151	157	Present
surgical drain	180	194	seen	198	202	Present
line	224	228	stable	232	238	Present

# Results

# Round 1 Results

Round 1 compared output from the original NLP with output from the original annotated reference standard

Some errors in the reference standard found and corrected:

- 8 devices were mis-annotated by annotators
  - The device was omitted (e.g. “curvilinear line is no longer seen”)
  - Inclusion of inappropriate devices (e.g. “aortic valve”)
- 20 devices inappropriately identified by NLP
  - The device was omitted (i.e. “tracheostomy appliance”)
  - Inclusion of inappropriate devices (i.e. “monitoring devices”)



# Discussion

# Discussion of Main Findings

- The NLP algorithm had excellent recall (95%) and precision (98%) for line/device terms
  - also good performance on inferencing about the terms found
    - insertion (89%) and removal (94%) terms
- It is easy to add additional terms such as abbreviations if they are found
- The CXR Debugger tool was very helpful for improving the NLP in quick rounds of development and testing

# Remaining Challenges

- Could possibly improve insertion and removal detection even further
  - we had relatively few CXR reports with this information
  - we now plan to obtain sample of all CXR reports, not just portable CXR, during the patient stay
    - Expect that this will capture more of the insertions and removals, which often are done outside of ICU stay
- Linkages of reports for same patient through time when the sequence of information in the reports is not as expected
  - will require decision rules about what to impute in the gaps
- Challenge for free-text analyses of radiology reports and to some extent all other reports
  - Identifying the correct set of reports in the absence of structured labels

# Potential Applications of the CXR NLP

- Infection surveillance and epidemiological studies
  - “central line days”
    - Would not require real-time application
- Using NLP to feed into clinical decision support (CDS)
  - Requires real-time application

# References

- Rubin D, Wang D, Chambers DA, Chambers JG, South BR, Goldstein MK. Natural language processing for lines and devices in portable chest x-rays. AMIA Annu Symp Proc 2010; 692-6.
- Goetz AM, Wagener MM, Miler JM, Muder RR. Risk of infection due to central venous catheters: effect of site of placement and catheter type. Infection Control and Hospital Epidemiology. 1998; 19(11): 842-845.
- Cunningham H, Maynard D, Bontcheva K, Tablen V. GATE: an Architecture for development of robust HLT. Applications Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics; 2002, p.168-175
- Tokars JI, Klevens RM, Edwards JR, Horan TC. Measurement of the impact of risk adjustment for central line-days on interpretation of central line-associated bloodstream infection rates. Infection Control and Hospital Epidemiology. 2007; 28(9): 1025-1029.
- Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucl. Acids Res. (2004) 32 (suppl 1):D267-D270
- Rubin DL. Creating and curating a terminology for radiology: ontology modeling and analysis. Journal of Digital Imaging. 2008 Dec; 21 (4): 355-62.
- Ogren PV. Knowtator: a protege plug-in for annotated corpus construction. Proceedings of the 2006 conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics; 2006. p.273-5.

For more information, please contact:

Mary K. Goldstein at [mary.goldstein@va.gov](mailto:mary.goldstein@va.gov)

Dan Wang at [dan.wang@va.gov](mailto:dan.wang@va.gov)

Tammy Hwang at [tammy.hwang@va.gov](mailto:tammy.hwang@va.gov)



## Results for 500 reports: preliminary

	Total	Total(NLP)	TP	FN	FP	Recall	Precision
Lines/devices	799	796	766	33	30	0.958	0.962
status-insertion	97	95	76	21	19	0.783	0.8
status-removal	66	66	59	7	7	0.893	0.893
status-presence	620	635	572	48	63	0.922	0.9
Status-other	16	0	0	16	0	0	n/a



## **NLP Algorithm Refinements:**

- Assumption of nearest device status: good for handling lists.

Example: "The endotracheal tube and nasogastric tube are in satisfactory position."

- "nasogastric tube" has the status of "in satisfactory position". No status term for "endotracheal tube" is found within its own clause. However, "endotracheal tube" is close to the device term: "nasogastric tube", and by our logic, assumes the same status.