CUIMANDREef (Capture with UIMA of Needed Data using Regular Expressions for EF)

Jennifer Garvin PhD, MBA, RHIA, CPHQ, CCS, CTR, FAHIMA

PI of EF Project A 1st Year Translational Use Case Project (TUCP) of the CHIR

Acknowledgement for Efforts of the EF Team

Brett South, Shuying Shen, Dan Bolton,

Scott DuVall, Jim Potter, Julia Heavirland,

Bruce Bray, Steve Pickard

Paul Heidenreich, Willie Ray

Polling Question

- What do you most want to learn about during this session (please choose the most important)
  1. Investigator interested in learning more about CHIR
  2. Investigator working with NLP/IE research interested in the methods used in this project
  3. Clinician interested in being able to use the results of this research
  4. Clinician/Investigator/ Administrator/ Staff member wanting to learn how NLP/IE is being developed and used in VA
  5. Other- does not fit above categories

Overview

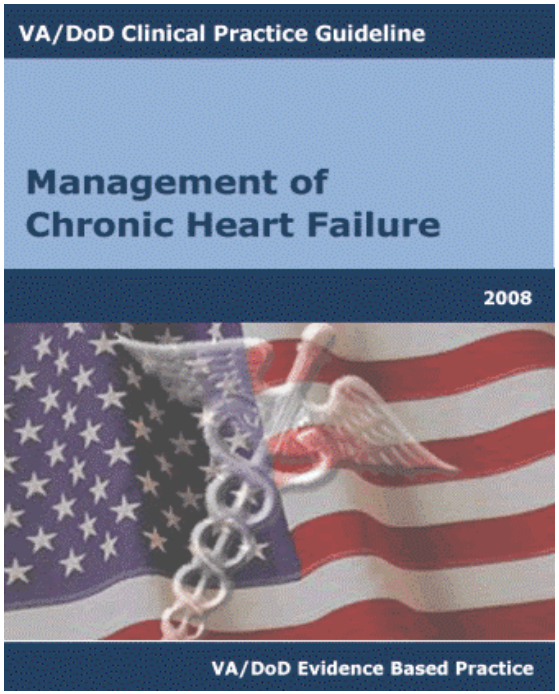
- Describe one VA applied informatics research project with regard to:
  - Team of researchers
  - Applied use case (purpose and research objectives)
  - Highlights of using the VA Informatics and Computing Infrastructure (VINCI)
  - Highlights of steps developing this natural language processing (NLP) tool
  - Results and Findings
  - Conclusions and Directions for future research

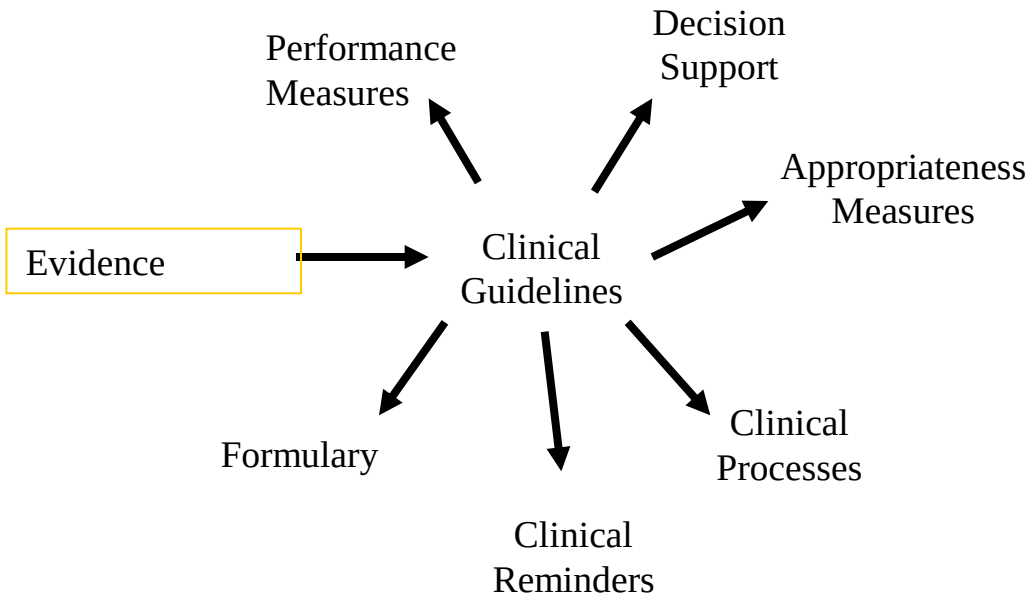
Ejection Fraction Team

- PI
- Informatics Co-Investigators:
  - Statistician expert in Annotation and System Development
  - Expert in Annotation and System Development
  - Expert in System Development and VINCI
  - Expert in Clinical, Research Informatics and Clinical Domain Expert (Cardiology)
- CHF Heart QUERI Co-investigators
- SAS and SQL Expert Database Developer
- Research Coordinator and Statistician

Applied Use Case
Purpose & Background

American Heart Association
Learn and Live®





VA Informatics and Computing Infrastructure- VINCI

- IRB approvals from all sites
- Obtain NDS approval
- Assigned a research folder on VINCI
- All approved investigators and staff can access data
- Used a SAS program (PHI Hunter) to remove all obvious identifiers not needed for the research (undertaken after documents were determined and prior to undertaking research)

VINCI- Document Acquisition

- Initial request to VINCI staff
- Text notes related to echos can be in several locations of the VISTA tables because each medical center can have a unique location
- Iterative document review to ensure that the right notes are available for the research
- Documents available for research and are present in VINCI research folder

Other Preparation

- Workflow analysis- visitation and discussion with 2 echo laboratories
- Understanding how the documents are developed- variety of approaches
- Initial review of document structure to inform our sampling strategy- **we planned to oversample free text and semi-structured text**

Methods-EF Sampling Strategy

- Timeframes and locations- (7 medical centers):
- In this study, each VA site is considered a cluster. We assumed that documents from the same cluster are correlated
- For the clustered design we obtained the sample size required to establish recall (sensitivity) of at least 90%
- We first calculated the required sample size assuming documents are independent, and then adjusted the sample size based on the clustering effect
- To account for clustering, the sample size was increased 29 positive cases per facility
- **Doubling the sample size for the three facilities with free- or semi-structured text** resulted in a minimum of 367 documents in the test set.

EF Sampling Strategy

- Because we had a total of 765 documents, 367 minimum documents for the test set, the remaining 398 documents were available for training.
- However, if during system training, the performance of the system reaches the pre-specified level of accuracy without using all available training documents, the remaining unused documents in the training set could be added to the test set.
- The 765 documents were randomly assigned to a training and test set in preparation for annotation.

NLP Development Methods- Reference (Gold) Standard Development

- All documents in the training and test sets must have an accompanying reference (gold) standard so that the accuracy of the system can be measured during training and testing
- Software program called Knowtator was used for annotation
- Two independent reviewers with a third adjudicator when disagreement occurred

NLP Development Methods- Training and Testing

- Training
  - Developed an initial rule set or process of extracting concepts
  - Separated the training documents into batches and ran the system
  - Evaluated false positives and false negatives based on comparison to the reference (gold) standard
  - Reprogram the system and run against the next batch
  - When pre-specified level of accuracy reached, measure accuracy at the last iteration
- Testing
  - The system is run on the sequestered documents and the output received by the statistician

Results- Reference Standard

- 765 documents annotated – reference/gold standard for testing binary classification of EF less than 40% or not (Use-case)
- The prevalence of EF less than 40% was found to be 13.5

Results Inter-annotator Agreement (IAA)

- The IAA at the use-case level was assessed using Cohen's kappa.
- The agreement between pairs of annotators was high -0.94-0.99

Results- System Development for Classification of EF less than 40%

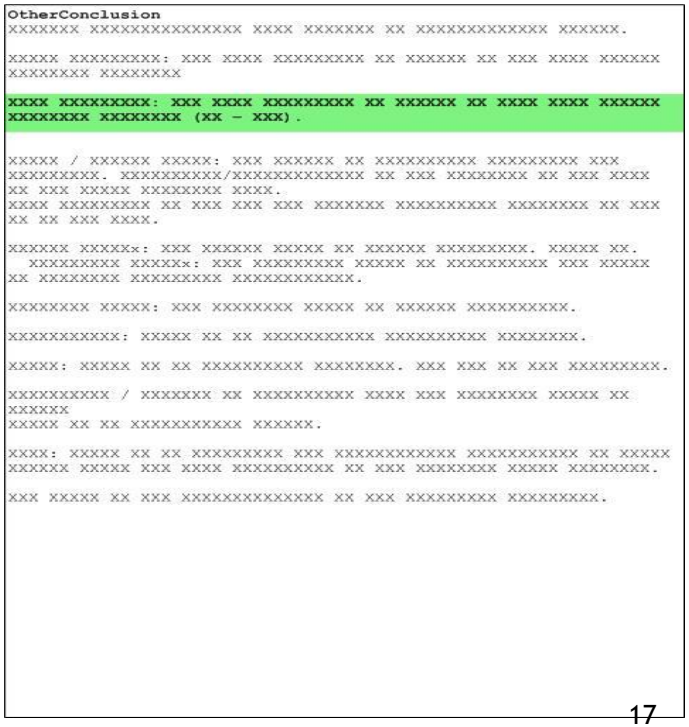
- **Training set – at the last iteration the** sensitivity (recall) (based on comparison between reference standard and system output) was 91.38%, one more revision of the programming was done before the test.
- **Test set** [1]Sensitivity (recall):98.41% (CI 91.47-99.95)

[1] Note that values of Unknown and Undetermined have been classed as negative for this analysis.

Results- Document Format (Fake Texts using XXX in Place of Letters)

- A significant finding was that the format and location of the relevant section was associated with false negatives in the training set (6/7 associated with format #5) and the test set (4/4 format 3)


ReportText
I. STUDY INFORMATION
A. XXXXXX:
B. XXXXXXXXXX:
C. XXXXXXXXXX:
II. PATIENT INFORMATION
A. XXXXXXXXXXXXXXXXXX:
B. XXX:
C. XXXXXX:
III. REQUESTING INFORMATION
A. XXXXXX XXXXX:
B. XXXXXXX XXXXXX:
IV: CONCLUSIONS
A. XXXXXXXXXX: XXXXXX XXXXXXX XXXXX XXXX XXXXX x XXXXXXXXX XXXXXXXX
B. XXXXXXXXXX XX: XXXX XXXX XXXXX XXXX XXXXXXXXX XXXXXX XXXXXX X XXX
C. XXXX XXXXXX: XX X XXXXXXX XXXX XXXXXX XXXXX XX X XXXXXX X XXXXXX X
D. XXXXXX XXXXX: XXXX XXXXXXXXXXX; XX XXXXXXXX
E. XXXXXXXXX XXXXX: XXXXXXX XXXXXXXXXXXX
F. XXXXXXXXX XXXXXXXXXXXX: XXXXXXXXX
1. XXXXXX XXXX XXXXXXXXXX XXXXXXX XXXXXXX: XX XXXXXXXXX XXXX.
2. XXXXXX XXXXXXXXX XXXXXXX XXXXXXX. XXXXXX XXX XXXXXXX.
3. XXXXXX XXXXXXXXXX XXXXXX XXXXX. XX XXXXXXXXXX XXXXXXX XXXX XXXXX XXXXX XX XX/XXXX.
CARDIAC ULTRASOUND ASSESSMENT DETAILS
VI. MEASUREMENTS VALUE NORMAL RANGE
A. XX XXXXXXX XXXXXXXXX (XXXXX): X.X (X.X - X.X XX)
…
H. XXXXX XXXXXXXXXX XXXXXXXXX: X.X (X.X - X.X XX)
VII. VALVE ASSESSMENT
A. XXXXXX XXXXX
1. XXXXXX XXXXX XXXXXXXXXXX: XXXXXXX
…
D. XXXXXXX XXXXX XXXXXXXXXXX: XXXXXXX
IX. OTHER FINDINGS
A. XXXXX XXXXXXXXX (XX)|
1. XX XXXXXXXX XXXXXXX: XXX XXXXXXXX
…
D. XXXXXXXXX
X. CARDIAC ULTRASOUND LABORATORY INFORMATION
A. XXXXXXX XX XXXXXXXXXX: XX XXX/XX


OtherConclusion
XXXXXXX XXXXXXXXXXXXXX XXXX XXXXXXX XX XXXXXXXXXXXX XXXXXX.

XXXXX XXXXXXXXX: XXX XXXX XXXXXXXXX XX XXXXXX XX XXX XXXX XXXXXX XXXXXXXX XXXXXXXX

XXXX XXXXXXXXX: XXX XXXX XXXXXXXXX XX XXXXXX XX XXXX XXXX XXXXXX XXXXXXXX XXXXXXXX (XX – XXX).

XXXXX / XXXXXX XXXXX: XXX XXXXXX XX XXXXXXXXXX XXXXXXXXX XXX XXXXXXXXX. XXXXXXXXXX/XXXXXXXXXXXX XX XXX XXXXXXXX XX XXX XXXX XX XXX XXXXX XXXXXXXX XXXX.
XXXX XXXXXXXXX XX XXX XXX XXX XXXXXXX XXXXXXXXX XXXXXXXX XX XXX XX XX XXX XXXX.

XXXXXX XXXXXx: XXX XXXXXX XXXXX XX XXXXXX XXXXXXXXX. XXXXX XX. XXXXXXXXX XXXXXx: XXX XXXXXXXXX XXXXX XX XXXXXXXXXX XXX XXXXX XX XXXXXXXX XXXXXXXXX XXXXXXXXXXX.

XXXXXXXX XXXXX: XXX XXXXXXXX XXXXX XX XXXXXX XXXXXXXXX.

XXXXXXXXXX: XXXXX XX XX XXXXXXXXXX XXXXXXXXX XXXXXXXX.

XXXXX: XXXXX XX XX XXXXXXXXX XXXXXXX. XXX XXX XX XXX XXXXXXXXX.

XXXXXXXXX / XXXXXXX XX XXXXXXXXX XXXX XXX XXXXXXX XXXXX XX XXXXXX
XXXXX XX XX XXXXXXXXXX XXXXXX.

XXXX: XXXXX XX XX XXXXXXXXX XXX XXXXXXXXXX XXXXXXXXX XX XXXXX XXXXXX XXXXX XXX XXXX XXXXXXXXX XX XXX XXXXXXX XXXXX XXXXXXX.

XXX XXXXX XX XXX XXXXXXXXXXXX XX XXX XXXXXXXX XXXXXXXX.

Limitations

- Limited number of echocardiogram formats
- We used a simple regular expressions approach rather than more complex approaches such as concept-based indexing
- We used one of many document sources for ejection fraction and the associated value

Conclusions and Future Directions

- We can efficiently obtain the classification of ejection fraction less than 40% value for quality measurement from echocardiogram reports using this system
- The format of the echocardiogram report affects the performance of the NLP system
- It is important to further generalize and develop this tool by using it in a greater variety of echocardiogram documents within the VA

Acknowledgements and Thank you!

- Other comments and questions? Contact e-mail Jennifer.garvin@va.gov
- Special Thanks to the Research Team listed in first slide and special acknowledgements for mentorship by Matt Samore M.D., Mary Goldstein M.D. , M.S., Bruce Bray M.D. and Paul Heidenreich M.D.
- This study is undertaken as part of the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the 1st year Translational Use Case Project, grant #: HIR -09-007
- Views expressed are those of the speaker and not necessarily those of the Department of Veterans Affairs