Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure

Jennifer H Garvin,[1,2,3] Scott L DuVall,[1,2] Brett R South,[1,2,3] Bruce E Bray,[3] Daniel Bolton,[1,2] Julia Heavirland,[1] Steve Pickard,[1,2] Paul Heidenreich,[4,5] Shuying Shen,[1,2,3] Charlene Weir,[1,6,3] Matthew Samore,[1,2,3] Mary K Goldstein[4,5]

[1]IDEAS Center, SLC VA Healthcare System, Salt Lake City, Utah, USA
[2]Division of Epidemiology, University of Utah School of Medicine, Salt Lake City, Utah, USA
[3]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA
[4]VA Palo Alto Health Care System, Palo Alto, California, USA
[5]Center for Primary Care and Outcomes Research (PCOR), Stanford University, Stanford, California, USA
[6]GRECC, SLC VA Healthcare System, Salt Lake City, Utah, USA

Correspondence to
Dr Jennifer H Garvin, Salt Lake City VA IDEAS Center 660/151, George E. Wahlen VA Medical Center, 500 Foothill Dr., Salt Lake City, UT 84148, USA; jennifer.garvin@va.gov

This study is undertaken as part of the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374; and CHIR Project: VA HSR HIR 08-374 (CHIR) HIR 09-007. The views expressed are those of authors and not necessarily those of the VA or affiliated institutions.

Received 9 August 2011
Accepted 27 February 2012

ABSTRACT

Objectives Left ventricular ejection fraction (EF) is a key component of heart failure quality measures used within the Department of Veteran Affairs (VA). Our goals were to build a natural language processing system to extract the EF from free-text echocardiogram reports to automate measurement reporting and to validate the accuracy of the system using a comparison reference standard developed through human review. This project was a Translational Use Case Project within the VA Consortium for Healthcare Informatics.

Materials and methods We created a set of regular expressions and rules to capture the EF using a random sample of 765 echocardiograms from seven VA medical centers. The documents were randomly assigned to two sets: a set of 275 used for training and a second set of 490 used for testing and validation. To establish the reference standard, two independent reviewers annotated all documents in both sets; a third reviewer adjudicated disagreements.

Results System test results for document-level classification of EF of <40% had a sensitivity (recall) of 98.41%, a specificity of 100%, a positive predictive value (precision) of 100%, and an F measure of 99.2%. System test results at the concept level had a sensitivity of 88.9% (95% CI 87.7% to 90.0%), a positive predictive value of 95% (95% CI 94.2% to 95.9%), and an F measure of 91.9% (95% CI 91.2% to 92.7%).

Discussion An EF value of <40% can be accurately identified in VA echocardiogram reports.

Conclusions An automated information extraction system can be used to accurately extract EF for quality measurement.

OBJECTIVES

Effective quality and performance monitoring and feedback requires complete and accurate information, a goal congruent with the 'meaningful use' concept promulgated by the American Recovery and Reinvestment Act.[1–6] The VA External Peer Review Program (EPRP) provides metrics of accountability for heart failure to determine if clinical guidelines are being used by treatment providers appropriately to facilitate high-quality care for veterans.[3–5 7] These performance measures are formulated from evidence-based clinical practice guidelines developed by the American College of Cardiology and the American Heart Association.[8 9] A key clinical component of the guideline is assessment of the left ventricular ejection fraction (EF) and the determination of whether the EF is <40%, as this identifies patients eligible for life-prolonging treatments.[9]

The VA has used automated information systems since 1985 in the form of the Veterans Health Information Systems and Technology Architecture (VistA) and the Computerized Patient Record System.[10–12] Methods of automating the performance measurement from available structured data using the VA electronic health record (EHR) have been developed by Goulet *et al*[13] for other measures; however, the EF is not consistently available in a structured form that can be used for quality measurement. Our research addresses this key gap in automated data collection for heart failure and provides mechanisms to further national efforts to develop EHRs that facilitate quality care.

BACKGROUND AND SIGNIFICANCE

Heart failure is the number one reason for discharge of veterans treated within the VA healthcare system.[14] Approximately 5.8 million people in the USA have experienced heart failure and each year 670 000 people are diagnosed with it for the first time. In 2010, heart failure cost the USA $39.2 billion[15]; these costs were associated with healthcare services, medications, and lost productivity of workers with heart failure.[16]

Multiple life-prolonging treatments for patients with heart failure have been demonstrated in randomized trials that enrolled patients with an EF of <40%.[9] Thus, the EF measurement helps clinicians determine a patient's treatment. Accordingly, the VA and other organizations, such as the Joint Commission, have used assessment of EF as a measure of quality and performance. Currently, VA EPRP abstractors manually abstract data from the Computerized Patient Record System for this purpose.[7]

Human abstraction is time consuming and expensive.[17] Furthermore, readily adaptable information technologies exist to automate and reliably extract this type of information from free-text data.[17 18] Natural language processing (NLP) approaches have been developed to identify patients with heart failure for prospective clinical trials[19] and to extract clinical concepts related to heart failure from radiology[20] and non-VA echocardiogram reports.[21] We believe that similar NLP approaches could be used to extract concepts from VA echocardiograms.

MATERIALS AND METHODS

Setting and context

The Consortium for Healthcare Informatics Research (CHIR)[22] is a major VA research initiative, formed to create new tools and methods for NLP. VA Informatics and Computing Infrastructure (VINCI)[23] facilitates the use of data in a secure environment and serves as both a software development environment and a secure location to store and analyze data. This study used VINCI and was conducted within CHIR. The VA's Corporate Data Warehouse (CDW)[24] is a national repository comprising data from many VA clinical and administrative systems which provided data for this work in association with VINCI. The documents used in this research were obtained from the VA CDW and housed within VINCI to develop the NLP system.

Document acquisition and document formats

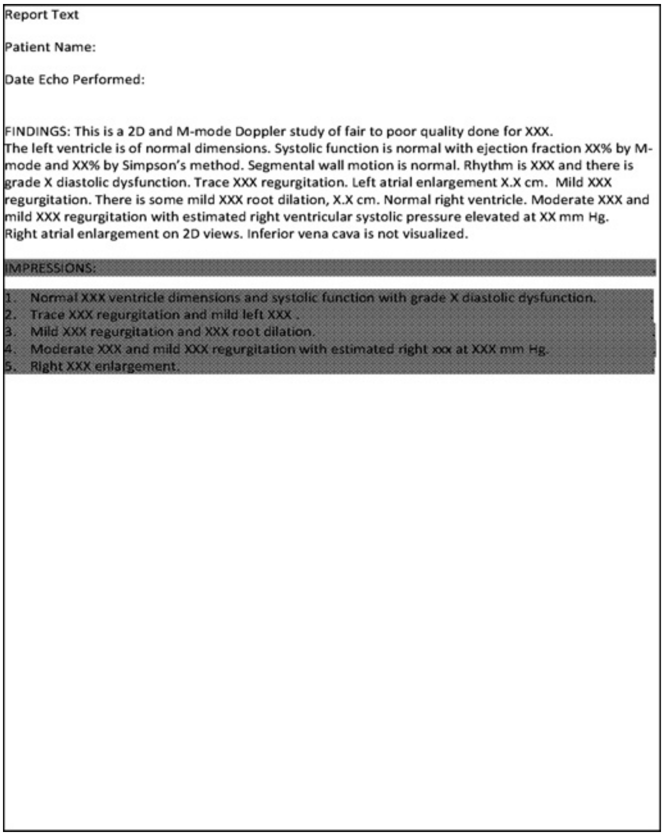
Echocardiogram text documents can be found in a variety of locations within the VistA files, including the Text Integration Utilities (TIU) note file, radiology file, and medicine file. TIU is a set of software tools used in the VA to create and manage clinical documents in a standardized manner[25] and is also the label of one file within VistA, where documents developed by TIU software are stored. The principal investigator and clinical domain experts developed a query for the CDW tables associated with the VA EHR for echocardiogram reports from April 1, 2007 through September 30, 2008 from one VA medical center and from January 1, 2006 to December 31, 2008 for the remaining centers. Data were available for the majority of medical centers in our study through VINCI, but for one medical center there were limited data. We found 14 118 documents that met the target date range.

The 14 118 documents were grouped into sets based on originating site and placed in ascending length order. To eliminate incomplete notes, the research team visually examined notes and determined the minimum number of characters constituting complete notes. This number varied by site (500, 550, 600, 1200, 1500) due to document formatting and inclusion of header and footer text. Selecting only reports longer than the designated minimum reduced the number to 12 002.

The team visually analyzed a sample of 15–20 documents from each medical center to determine the degree of document structure as follows.

We found that the characteristics of the documents varied by VA location. We categorized the document based on the degree of structure noted above as illustrated in figures 1–3. We also categorized documents with similar characteristics, such as use of an outline, large headers, small headers, and the location of the conclusion section, into five distinct document formats, numbering each document according to its format. We identified the relevant conclusion section for each numbered format from which the EF was to be derived and the number of sections in each document as described in table 1.

The 12 002 reports in the document set were then placed in random order by medical center, and a random selection of 765 echocardiogram reports was drawn based on our sampling strategy described below. We then removed all obvious identifiers not needed for the research in accordance with IRB/human subject approval requirements.



Report Text

Patient Name:

Date Echo Performed:

FINDINGS: This is a 2D and M-mode Doppler study of fair to poor quality done for XXX.
The left ventricle is of normal dimensions. Systolic function is normal with ejection fraction XX% by M-mode and XX% by Simpson's method. Segmental wall motion is normal. Rhythm is XXX and there is grade X diastolic dysfunction. Trace XXX regurgitation. Left atrial enlargement X.X cm. Mild XXX regurgitation. There is some mild XXX root dilation, X.X cm. Normal right ventricle. Moderate XXX and mild XXX regurgitation with estimated right ventricular systolic pressure elevated at XX mm Hg.
Right atrial enlargement on 2D views. Inferior vena cava is not visualized.

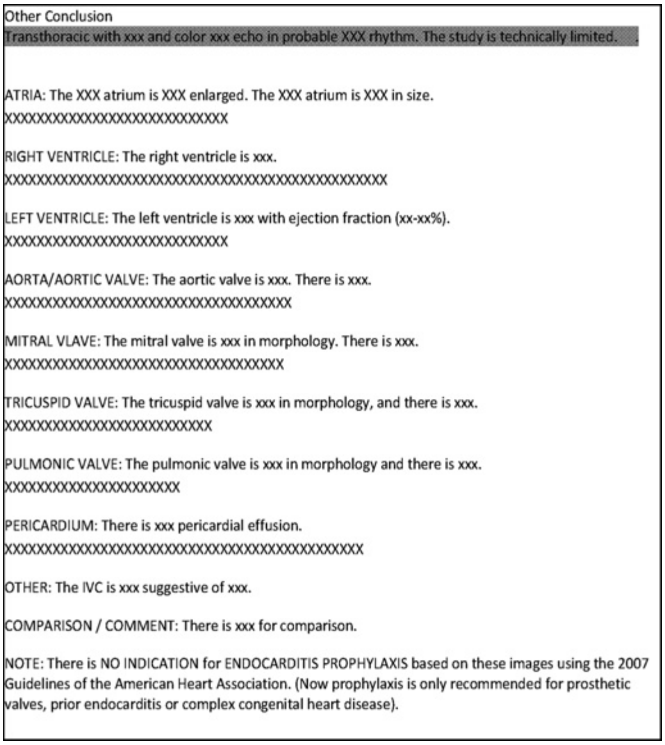
IMPRESSIONS:

1. Normal XXX ventricle dimensions and systolic function with grade X diastolic dysfunction.
2. Trace XXX regurgitation and mild left XXX .
3. Mild XXX regurgitation and XXX root dilation.
4. Moderate XXX and mild XXX regurgitation with estimated right xxx at XXX mm Hg.
5. Right XXX enlargement.

Figure 1 Degree of structure: unstructured (synthetic document).

Sampling strategy for document determination

The minimum sample size requirements were determined to provide specified confidence and power to assess recall in the test set of documents. Adjustments to the base sample size were



Other Conclusion
Transthoracic with xxx and color xxx echo in probable XXX rhythm. The study is technically limited.

ATRIA: The XXX atrium is XXX enlarged. The XXX atrium is XXX in size.
XXXXXXXXXXXXXXXXXXXXXXXXXXXX

RIGHT VENTRICLE: The right ventricle is xxx.
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

LEFT VENTRICLE: The left ventricle is xxx with ejection fraction (xx-xx%).
XXXXXXXXXXXXXXXXXXXXXXXXXXXX

AORTA/AORTIC VALVE: The aortic valve is xxx. There is xxx.
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

MITRAL VLAVE: The mitral valve is xxx in morphology. There is xxx.
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

TRICUSPID VALVE: The tricuspid valve is xxx in morphology, and there is xxx.
XXXXXXXXXXXXXXXXXXXXXXXXXX

PULMONIC VALVE: The pulmonic valve is xxx in morphology and there is xxx.
XXXXXXXXXXXXXXXXXXXXXX

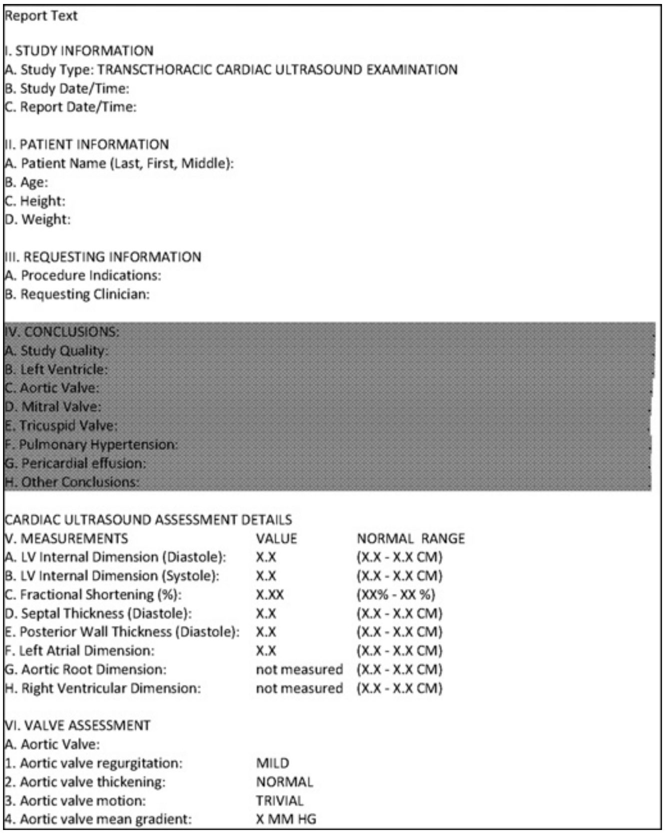
PERICARDIUM: There is xxx pericardial effusion.
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

OTHER: The IVC is xxx suggestive of xxx.

COMPARISON / COMMENT: There is xxx for comparison.

NOTE: There is NO INDICATION for ENDOCARDITIS PROPHYLAXIS based on these images using the 2007 Guidelines of the American Heart Association. (Now prophylaxis is only recommended for prosthetic valves, prior endocarditis or complex congenital heart disease).

Figure 2 Degree of structure: semi-structured (synthetic document).



Report Text

I. STUDY INFORMATION
A. Study Type: TRANSCTHORACIC CARDIAC ULTRASOUND EXAMINATION
B. Study Date/Time:
C. Report Date/Time:

II. PATIENT INFORMATION
A. Patient Name (Last, First, Middle):
B. Age:
C. Height:
D. Weight:

III. REQUESTING INFORMATION
A. Procedure Indications:
B. Requesting Clinician:

IV. CONCLUSIONS:
A. Study Quality:
B. Left Ventricle:
C. Aortic Valve:
D. Mitral Valve:
E. Tricuspid Valve:
F. Pulmonary Hypertension:
G. Pericardial effusion:
H. Other Conclusions:

CARDIAC ULTRASOUND ASSESSMENT DETAILS

V. MEASUREMENTS	VALUE	NORMAL RANGE
A. LV Internal Dimension (Diastole):	X.X	(X.X - X.X CM)
B. LV Internal Dimension (Systole):	X.X	(X.X - X.X CM)
C. Fractional Shortening (%):	X.XX	(XX% - XX %)
D. Septal Thickness (Diastole):	X.X	(X.X - X.X CM)
E. Posterior Wall Thickness (Diastole):	X.X	(X.X - X.X CM)
F. Left Atrial Dimension:	X.X	(X.X - X.X CM)
G. Aortic Root Dimension:	not measured	(X.X - X.X CM)
H. Right Ventricular Dimension:	not measured	(X.X - X.X CM)

VI. VALVE ASSESSMENT
A. Aortic Valve:

1. Aortic valve regurgitation:	MILD
2. Aortic valve thickening:	NORMAL
3. Aortic valve motion:	TRIVIAL
4. Aortic valve mean gradient:	X MM HG

Figure 3 Degree of structure: structured (synthetic document).

made to account for documents that did not have EF information, for clustering of documents within the seven VA facilities, and for oversampling of unstructured and semi-structured document formats (figures 1 and 2).

For a non-clustered design, the minimum required sample size for recall was obtained using exact power calculations for a one-sided binomial test. For this application, the alternative hypothesis was defined as the proportion of positive cases (EF <40%) correctly identified, and recall as being above the specified threshold with at least 80% probability (power) that the 95% lower confidence bound exceeds the specified threshold of 0.90. This calculation resulted in a minimum of 179 documents. It was estimated that 80% of documents would contain EF information. In all we used 224 (179/0.8) documents.

In this study, each VA site was a cluster. We assumed that documents from the same cluster were correlated. We first calculated the required sample size assuming documents were independent, and then adjusted the sample size based on the clustering effect. We calculated the number of documents required per site based on the number of sites available and the intra-class correlation (ICC). To account for correlation, we used the design effect (Deff) to modify the usual method to calculate the number of records required per site: Deff=1+(m−1)×ICC, where m is the minimum sample size divided by the number of sites.

For 32 documents per cluster (224/7) and an ICC of 0.005, the design effect was 1.155. Therefore, the minimum number of positive cases in the clustered design was 224×1.155=259. Dividing these cases evenly among facilities resulted in 37 per facility. Doubling the sample size for the three facilities with un- or semi-structured formats resulted in a minimum of 367 documents in the test set. Because we had a total of 765 documents, 398 documents were available for training; we assigned the 765 documents randomly to test or training.

Reference standard development

Our reference standard was produced by human annotators using an annotation schema and guidelines. The schema and guidelines were pilot tested and iteratively developed (figures 1–3).[26–28] Annotation guidelines provided explicit examples and were used to train the annotators.[29] Several members of the research team conducted annotator training on documents reserved for that purpose until all three pairs of annotators achieved an inter-annotator agreement (IAA) of 90%,[30] and they began to annotate the 765 documents in the training and test sets. The IAA was assessed to determine the reliability of the reference standard.[31] Cohen's κ was used to assess pair-wise reliability.

There were three annotators in total who rotated in and out of the role of adjudicator, taking turns as both an annotator and an adjudicator. If there were differences between the two annotators' findings, the third annotator, serving in an adjudicator role, resolved the discrepancy. All three annotators were nurses trained to abstract quality measurement information, including EF, for the VA EPRP program. They were also given training based on the written annotation guidelines and using annotation software. We used Knowtator as the annotation software, a general-purpose text annotation tool developed to assist in evaluation of NLP systems,[30 32–34] and to develop a reference standard for NLP training and evaluation. The co-investigator responsible for developing the system did not participate in annotation development and did not have access to annotated test documents. Further, the test set was annotated independently of system development.

The Knowtator schema for this project included three classes of concepts, each corresponding to a concept in the domain.[34] The first class of concepts was 'all mentions of EF of the left ventricle.' The second class of concepts was 'all mentions of the quantitative value associated with the EF.' The third class of

Table 1 Characteristics of documents obtained for the research

VA medical center location	Degree of structure	Format number	Location of relevant section/ number of sections	Minimum number of characters considered to be complete	Number of documents used in the research
Location 1	Unstructured	Format 2	2/2	600	190
Location 2	Structured	Format 3	4/4	500	75
Location 3	Structured	Format 4	4/10	500	68
Location 4	Structured	Format 3	4/4	550	72
Location 5	Semi-structured	Format 5	4/13	1200	145
Location 6	Semi-structured	Format 1	9/9	1500	145
Location 7	Structured	Format 3	4/4	700	70
				Total documents	765

concepts was qualitative assessment of EF and the left ventricular (LV) systolic function. Note that moderate or worse LV systolic dysfunction was considered functionally equivalent to EF <40%. The method used to measure EF and the section of the report in which the EF mention was found were also annotated. All EF mentions were annotated.

Measurements

In addition to providing descriptive statistics, we used measurements (not clustered by site) for NLP system training and testing based on partial and exact matching of the reference standard and NLP system output. These included precision (equivalent to positive predictive value), recall (equivalent to sensitivity or true positive rate), accuracy, and the F measure, the harmonic mean of recall and precision for concept classes. Accuracy was calculated by dividing the number of correct identifications by the NLP system, when compared to the reference standard, by the number of documents the system reviewed. We used Cohen's κ to determine the IAA when there was comparison of two annotators during the development of the reference standard, and the multi-rater Fleiss' κ was used to establish agreement among all three annotators. In addition, we determined false positives and false negatives and conducted detailed failure analyses at each iteration of system development during training.

SYSTEM DEVELOPMENT: NLP TRAINING AND TESTING FOR CONCEPT EXTRACTION AND RULE DEVELOPMENT FOR INFERENCE

The system was designed within the Unstructured Information Management Architecture (UIMA) framework.[35] Determining the document-level classification of EF greater than or less than 40% was achieved in two steps, first by extracting concepts from the report text (NLP development) and then by using this information to determine the binary classification of whether or not the document contained an EF of <40% (inference).

NLP development for concept extraction

Once the training and test sets were established, we divided the training set reports into batches. The batches were determined by selecting 15–45 documents for each batch starting with the first document in the training set. Based on a review of the first batch, we developed an initial set of patterns for the NLP system to extract concepts and classify them (figure 4).

Regular expressions, string matching, and filters were used to extract each of the concept classes (figure 5, items A–F). The report was split into sections by regular expressions that identified numbered or lettered lists and paragraphs that started with words followed by a colon (figure 5, item A). The concepts were found using multi-part regular expressions that allowed for variation in expression (figure 5, item B). For example, introductory text such as 'estimated' or 'visual estimate of' could have been followed by modifiers like 'global,' 'LV,' or 'left-sided,' which could have been followed by the concept 'EF,' 'function,' 'systolic function/dysfunction,' or 'contractility.' Because of the general nature of some of the concepts (such as 'function'), the section in which the concept was found and the preceding text to each mention of the concept was used as a filter. Previous exams and reference to the right ventricle would exclude a concept. A list of possible ways that concepts could be expressed in the text was developed during the iterative failure analysis process during training. After each new document was added to the training set, new concepts were added to the list.

Qualitative values were found using number patterns, but allowed for modifiers such as 'near,' '>,' '%,' and ranges of values (figure 5, item C). Quantitative assessments were found in a similar manner (figure 5, item D). The method by which the EF value was measured in the report (usually M-mode or Simpson's method) was also identified (figure 5, item E). The reason for finding the method information was to validate the text when two different values were found in the same sentence. The individual instances of each class were then combined so that the EF mention, qualitative and qualitative assessments, and context could be evaluated together (figure 5, item F). To do this, patterns were created that used the individual concept classes instead of the actual text.

The original report text was temporarily replaced by tags that represented the concept classes. This two-phase pattern building allowed complex regular expressions to be used for the individual classes along with rules and filtering. However, only the

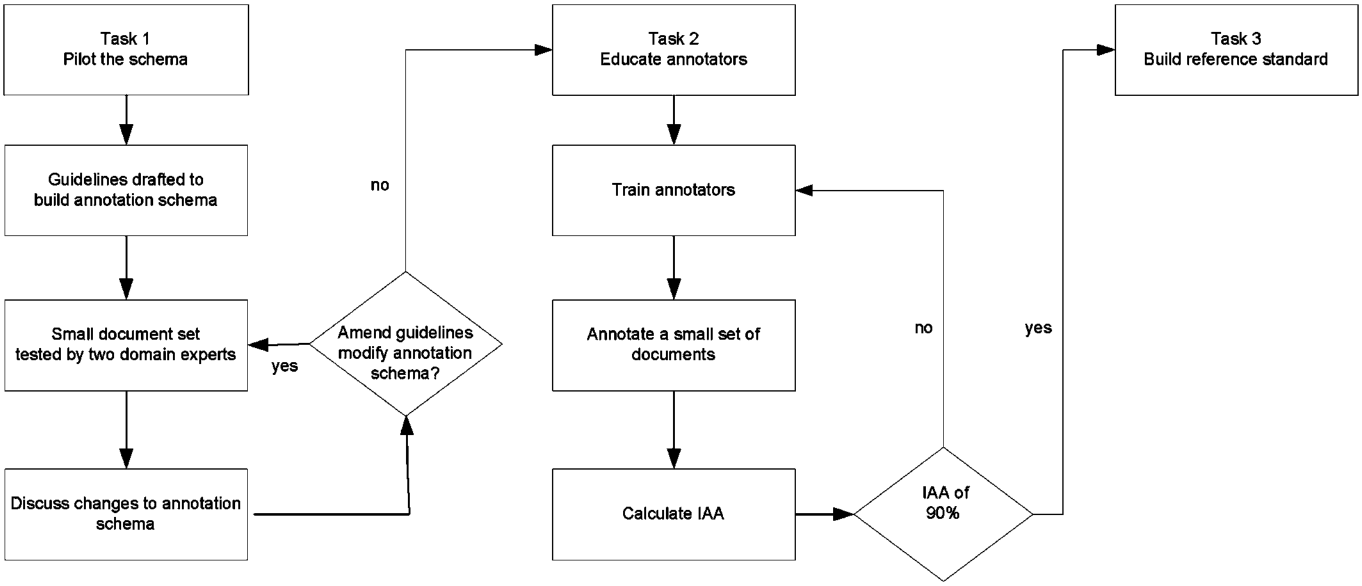

Task 1
Pilot the schema
Guidelines drafted to build annotation schema
Small document set tested by two domain experts
Discuss changes to annotation schema
Amend guidelines modify annotation schema?
yes
no
Task 2
Educate annotators
Train annotators
Annotate a small set of documents
Calculate IAA
IAA of 90%
no
yes
Task 3
Build reference standard

Figure 4 Development of annotation guideline and schema for establishment of a reference standard. IAA, inter-annotator agreement.

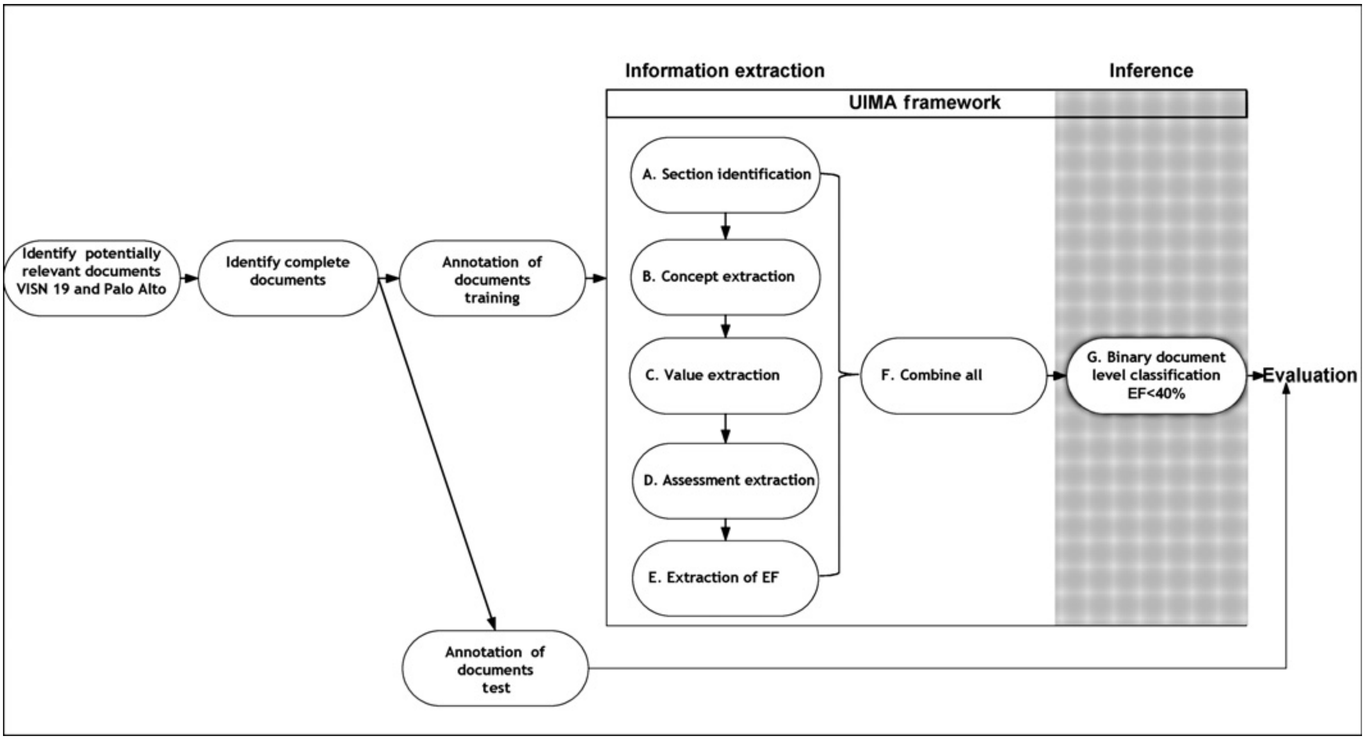
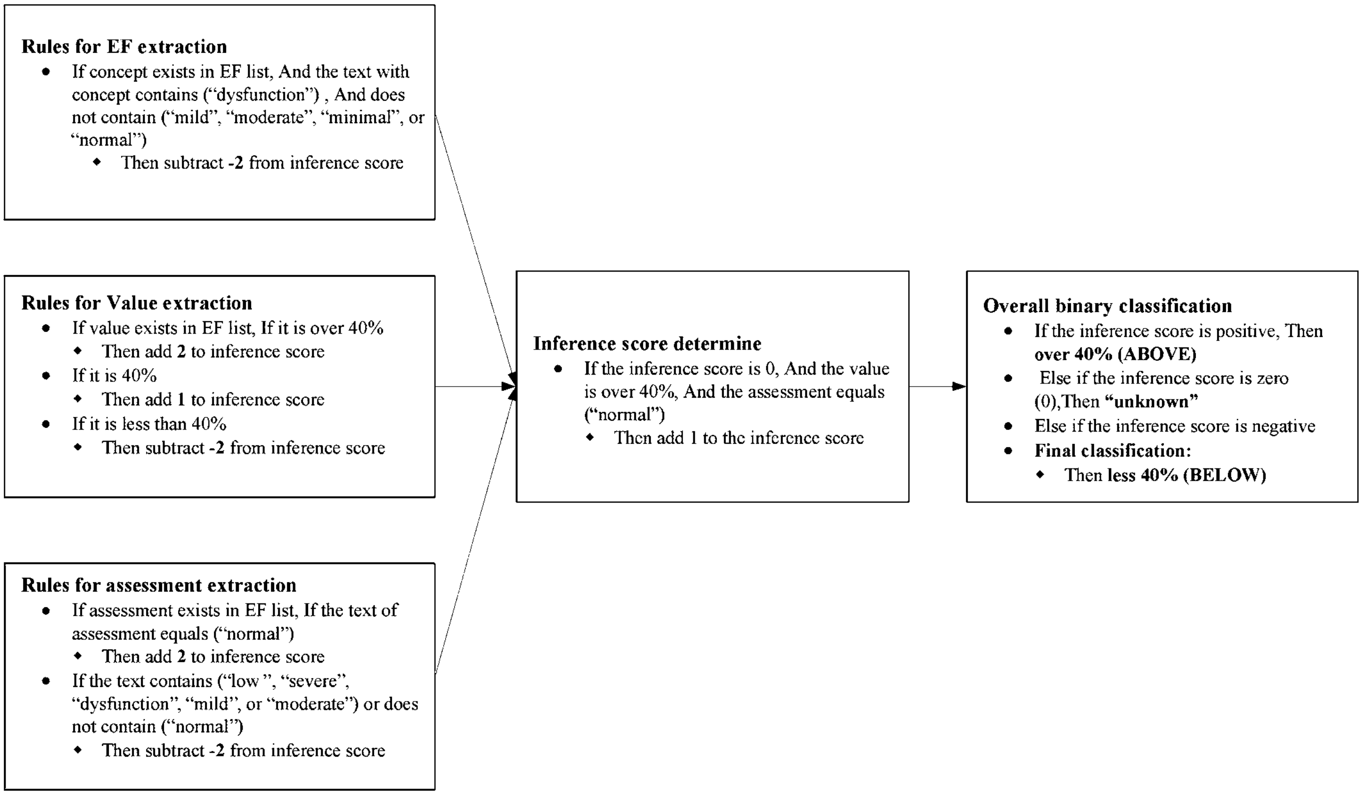

Information extraction
Inference
UIMA framework
Identify potentially relevant documents VISN 19 and Palo Alto
Identify complete documents
Annotation of documents training
A. Section identification
B. Concept extraction
C. Value extraction
D. Assessment extraction
E. Extraction of EF
F. Combine all
G. Binary document level classification EF<40%
Evaluation
Annotation of documents test

Figure 5 EF system components and use. EF, ejection fraction.

end result (represented as tags in the text) was used for the next set of patterns. Because the combination of values was prone to much more variation than the individual concept classes, the list of patterns used to combine the concepts had the most number of regular expressions and took the most number of iterations to converge on a final pattern set. The patterns included the individual concept class tags in different combinations and order and included connecting text (such as 'appears to be,' 'was about,' 'around') that was often interspersed. As part of the NLP training process, we conducted a detailed failure analysis to



Rules for EF extraction
- If concept exists in EF list, And the text with concept contains ("dysfunction") , And does not contain ("mild", "moderate", "minimal", or "normal")
  - Then subtract **-2** from inference score

Rules for Value extraction
- If value exists in EF list, If it is over 40%
  - Then add **2** to inference score
- If it is 40%
  - Then add **1** to inference score
- If it is less than 40%
  - Then subtract **-2** from inference score

Rules for assessment extraction
- If assessment exists in EF list, If the text of assessment equals ("normal")
  - Then add **2** to inference score
- If the text contains ("low ", "severe", "dysfunction", "mild", or "moderate") or does not contain ("normal")
  - Then subtract **-2** from inference score

Inference score determine
- If the inference score is 0, And the value is over 40%, And the assessment equals ("normal")
  - Then add 1 to the inference score

Overall binary classification
- If the inference score is positive, Then **over 40% (ABOVE)**
- Else if the inference score is zero (0),Then **"unknown"**
- Else if the inference score is negative
- **Final classification:**
  - Then less **40% (BELOW)**

Figure 6 Rules and sequence of use. EF, ejection fraction.

Table 2 Cohen's κ (bootstrap 95% confidence intervals) for the documents

Annotator pairs	Agreements/total number of documents	κ (95% CI)
A1 and A2	291/292	0.91 (0.84 to 0.97)
A2 and A3	277/281	0.76 (0.67 to 0.85)
A1 and A3	241/243	0.83 (0.71 to 0.91)

identify needed changes and update or add to the patterns set. We repeated these steps using the training batches, modifying the patterns iteratively, until the system reached a recall and precision target level of 90% for identifying relevant concepts. Please see table 3 for a description of the results of each of the iterations of the patterns and examples of what changes were made in the patterns.

Rule development for inference

Rules were used to assess each of the combinations of concepts in the document and gave positive weight toward a final score of combinations providing evidence of the classification of EF >40% and negative weight to combinations providing evidence of EF <40%. Each assessment was mapped as positive or negative evidence and given the same weight as quantitative values. After all mentions had been assessed, documents with a final score >0 were classified as EF >40%, <0 were classified as EF <40%, and 0 were classified as unknown EF. The patterns for concept extraction and rules for inference and classification were developed for use across all of the VA sites in the study and were not specific to any one site. We named this system Capture with UIMA of Needed Data using Regular Expressions for EF (CUIMANDREef) (figure 6).

The section heading and location in the document was used to determine the sections of the echocardiogram report, such as measurements taken by the ultrasound machine and the conclusion section. This was done so that when there were multiple mentions of EF, the system could obtain the EF from the section that represented the synthesis of the assessment of the heart, many times titled 'conclusion' section, or commonly, but not always, following the imaging and measurement sections. During training, the performance of the system reached the pre-specified level of accuracy without using all of the training documents. The 123 unused documents were added to the original test set of 367 documents, giving us a total of 490 documents in the test set.

RESULTS

Reference standard development: results of inter-rater agreement

The prevalence of EF of <40% was calculated based on the test set annotation (reference standard) and found to be 13.5%. Because the prevalence was neither very high nor very low, no adjustment for prevalence was necessary to assess reliability and Cohen's κ was used. The reliability between pairs of annotators ranged from substantial to almost perfect agreement (table 2) for the binary classification of EF <40% or not. The κ values show very high overall agreement between each annotator pair. Each pair showed nearly perfect agreement with the least agreeing pair, A2 and A3, agreeing on 98.6% of documents. The high agreement among the pairs of annotators makes the κ statistic unusually sensitive to any disagreements. The lower κ value for A2 and A3 is a result of this sensitivity. A small set of 28 documents were rated by all three annotators (A1, A2, A3) to establish agreement among all annotators, for which the multi-rater Fleiss' κ was calculated to be 0.608 ($p<0.0001$), constituting moderate to substantial agreement.

System development and training

As part of the training, we used an initial set of regular expressions, strings, and filters for concept extraction purposes and a set of rules to use the extracted concepts to infer the binary classification of whether the echocardiogram had an EF of <40% or not from documents in the training set. Seven iterations of the regular expressions pattern set development were undertaken before testing (see table 3). The baseline performance of the initial pattern set was recall of 55.56%, precision of 36.25%, and an F measure of 43.87%. Following iterative development, we stopped the revision of pattern sets (training process) when we achieved, at the concept level, a sensitivity (recall) value of 91.83%, positive predictive value (precision) of 92.27%, and an F measure of 92.05% using the last training batch of documents (batch 6). We undertook a detailed failure analysis of the training set at the concept level and found no themes for the variation in false positives. After one additional pattern set revision, the final performance of the pattern sets (prior to test) was a recall of 97.48%, a precision of 90.23%, and an F measure of 93.71%. We also undertook a failure analysis of the inference process to classify documents with an EF of <40% in batch 6 of the training set, resulting in one false negative associated with format 2, and two false positives associated with format 5. At the document level for the test set one false negative was found, resulting in one misclassification; it was associated with format 3.

CUIMANDREef system testing

The CUIMANDREef test output consisted of correct and incorrect concepts when compared to the reference standard. For the test set of documents, at the document level, the NLP system accuracy was 98.2%. This represents 481 records matched out of 490 records. Table 4 includes the document-level classification for the reference standard and the NLP system in the test set. Note that 'positive' means EF is <40%; this is the condition we wanted to detect.

Table 3 Regular expression pattern set development for concept extraction

Iteration	Recall	Precision	F measure	Examples of pattern improvement
Initial pattern set/baseline	55.55%	36.25%	43.87%	
1	77.77%	88.42%	82.75%	Assign meaning to qualitative terms
2	94.96%	98.50%	96.70%	Improve exclusion of mentions referring to right ventricle
3	94.15%	96.02%	95.08%	Split mentions in conjunction and tie to appropriate mention of method
4	95.34%	96.09%	95.71%	Handle articles better, add new qualitative terms
5	94.22%	96.80%	95.49%	Adjust fuzzy mentions of concepts such as 'appears to be'
6	91.83%	92.27%	92.04%	Expand recognized patterns of concept, value, method of measurement
7 Final pattern set before test	97.48%	90.23%	93.71%	

Table 4 Comparison of reference standard and NLP system document-level classification as ejection fraction (EF) <40% for the test set

	Reference standard		
	Positive	Negative	Total
NLP system results			
Positive	62	0	62
Negative	1	412	413
Total	63	412	475

Table 4 shows a high level of accuracy: (62+412)/475=99.8% correct for the NLP extraction system when compared to the reference standard at the document-level classification. Fifteen documents from the total test set of 490 do not appear in this table because a classification was not determined by the annotators, the NLP system, or both.

As compared with the reference standard, the NLP performance on the document-level classification task was as follows: sensitivity (recall) 98.4% (95% CI 91.5 to 100.0), specificity 100% (95% CI 98.7 to 100.0), positive predictive value (precision) 100% (95% CI 91.5 to 100.0), and F measure 99.2% (95% CI 98.4 to 100.0). Specificity and positive predictive value are 100% because there were no false positives in the results from the test set. There was only one misclassification in the form of one false negative.

At the concept level, we assessed the performance of concept extraction across all classes, resulting in an overall sensitivity of 88.9% (95% CI 87.7 to 90.0), a positive predictive value of 95% (95% CI 94.2 to 95.9), and an F measure of 91.9%, (95% CI 91.2 to 92.7). Table 5 includes the concept-level analysis from the test documents. A detailed failure analysis of the NLP test set at the concept level was undertaken; no systematic reasons for the false positives were found.

Document results

Incomplete documents

There were 2116 documents from seven medical centers over an 18-month period that did not meet the minimum number of characters required for a document to be considered complete. Incomplete documents represented 15% of the echocardiogram reports initially obtained. The system ignored the incomplete reports because using them can provide inaccurate information. Incomplete reports may have some EF mentions but will not contain the most clinically relevant EF measurement, the measurement in the conclusion section of the report.

DISCUSSION

We successfully developed and used an NLP system (CUIMANDREef) to accurately classify VA echocardiogram reports at the document level according to whether or not they include an EF of <40%. The development of CUIMANDREef furthers the intention of meaningful use of EHRs as defined by the Department of Health and Human Services[1 5 36] which includes automatically extracting data from EHRs to improve quality of care.[25] Policy changes over the last 5 years requiring meaningful use of EHRs necessitate automation of clinical information not routinely found in structured data, such as the EF, through a means other than manual review of clinical records.[35 36]

We used CUIMANDREef on a variety of VA echocardiogram report formats with good performance. In the combined training and test sets, there were four misclassifications at the document level: in the training set, one false negative was associated with format 2 and two false positives were associated with format 5, and in the test set, one false negative was associated with format 3. In the varying format types, the location of the conclusion section and a larger number of sections may have been potential causes of failure in the training and test sets; however, we cannot know this conclusively from our data. We conclude that the system performs exceptionally well across all sites and formats while noting that in this application, performance differed slightly by format.

Using this system in real time would require several considerations. We found a significant number of incomplete documents and the possibility of eliminating or navigating them would need to be addressed. For example, part of the proposed plan would be to remove documents below a minimum number of characters.

We acknowledge several limitations of this work. We used echocardiograph reports from a limited number of medical centers, so we may not have encountered all possible document formats within the VA. CUIMANDREef only identifies 88.9% of patients with EF <40. Yet within the VA we do not currently have an automated method to obtain EF values from text, so having 88.9% of the values from available text provides greater access to this data. We plan to further generalize CUIMANDREef with a greater variety of document formats to improve its performance. We are conducting research using machine learning to gain better performance to obtain EF in another study.

In addition, the study was conducted entirely with VA documents, which raises a question about generalizability outside the VA. However, the cardiologists reading and reporting on echocardiograms typically have been trained and/or currently work in non-VA as well as VA settings, and there is no reason to believe that VA echocardiography reports differ systematically from non-VA reports. Future work will sample documents more widely across VA medical centers.

CONCLUSION

We built CUIMANDREef which successfully extracted EF concepts and associated EF values from free-text echocardiogram reports. This system has a sensitivity (recall) of 88.9% and a positive predictive value (precision) of 95% at the concept level, and a sensitivity of 98.41% and a positive predictive value of 100% at the document level. Document format may affect system performance for document-level classification because of

Table 5 Comparison of reference standard and NLP system at the class level for all concepts

	Test set performance by class for all concepts for all documents		
Concept class	Recall (95% CI)	Precision (95% CI)	F measure (95% CI)
Ejection fraction (EF) mentions	89.5 (87.1 to 91.6)	90.9 (88.6 to 92.9)	90.2 (88.7 to 91.8)
Quantitative values	91 (89.3 to 92.6)	95.5 (94.1 to 96.6)	93.2 (92.2 to 94.3)
Qualitative assessment of EF	84.9 (82.1 to 87.4)	99.5 (98.6 to 99.9)	91.6 (90.1 to 93.2)
Overall recall, precision, and F measure	88.9 (87.7 to 90)	95 (94.2 to 95.9)	91.9 (91.2 to 92.7)

varying relevant section location and a larger number of sections, but this will need to be further explored in future research. However, knowing the EF helps providers determine treatment decisions. Using an automated system to capture this data is more efficient than manual review and improves access to data. Heart failure experts may identify important additional use cases for this system, and future research may study implementation methods for this useful tool. Automation of classification of EF <40% is an important first step in providing critical information for quality improvement within the context of meaningful use of EHRs for patients with heart failure.

Acknowledgments We thank James Potter, Willy Ray, Allison Lynn Weaver, and Youngjun Kim for their work on this project.

Funding This study was supported by VA HSR&D, grant number: A HSR HIR 08-374; and CHIR Project: VA HSR HIR 08-374 (CHIR) HIR 09-007 (PJT).

Competing interests None.

Ethics approval Ethics approval was provided by University of Utah IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. *Being a Meaningful User of Electronic Health Records*. http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__meaningful_use_-providers/2998 (accessed 7 Jul 2011).
2. **Vest JR,** Jasperson J. What should we measure? Conceptualizing usage in health information exchange. *J Am Med Inform Assoc* 2010;**17**:302—7.
3. **CMS National Quality Forum.** *Guide for Reading the EHR Incentive Program EP measures*. 2010. http://www.cms.gov/QualityMeasures/Downloads/QMGuideForReadingEHR.pdf (accessed 6 Jan 2011).
4. **Bloomrosen M,** Starren J, Lorenzi NM, *et al*. Anticipating and addressing the unintended consequences of health IT and policy: a report from the AMIA 2009 Health Policy Meeting. *J Am Med Inform Assoc* 2011;**18**:82—90.
5. *The Official Web site for the Medicare and Medicaid electronic health records (EHR) Incentive Programs*. https://www.cms.gov/ehrincentiveprograms/ (accessed 6 Jan 2011).
6. *American Recovery and Reinvestment Act of 2009*. http://www.recovery.gov/About/Pages/The_Act.aspx (accessed 9 May 2011).
7. **Jha AK,** Perlin JB, Kizer KW, *et al*. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. *N Engl J Med* 2003;**348**:2218—27.
8. *Heart failure: Percentage of Patients with Diagnosed Heart Failure (HF) Aged Greater Than or Equal To 18 years with Quantitative Or Qualitative Results Of Left Ventricular Function (LVF) Assessment Recorded*. http://qualitymeasures.ahrq.gov/content.aspx?id=7803 (accessed 7 Jun 2011).
9. **Hunt SA,** Abraham WT, Chin MH. 2009 Focused Update Incorporated Into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation. *Circulation* 2009;**119**:e391—479.
10. *VistA-HealtheVet Monograph, 2008—2009*. http://www.va.gov/VISTA_MONOGRAPH/docs/2008_2009_VistAHealtheVet_Monograph_FC_0309.pdf (accessed 6 Jan 2011).
11. *VA Electronic Health Records (CPRS)*. http://vista.med.va.gov/cprs/ (accessed 6 Jan 2011).
12. **Kolodner RM.** *Computerizing Large Integrated Health Networks*. 1st edn. New York, NY: Springer Verlag, 1997.
13. **Goulet JL,** Erdos J, Kancir S, *et al*. Measuring performance directly using the veterans health administration electronic medical record: a comparison with external peer review. *Med Care* 2007;**45**:73—9.
14. **VA Office of Research and Development,** Health Services Research and Development Service. *Chronic Heart Failure*. Quality Enhancement research initiative (QUERI), 2011. http://www.queri.research.va.gov/about/factsheets/chf_factsheet.pdf (accessed 6 Jan 2011).
15. **Center for Disease Control.** *Heart Failure Fact Sheet*. http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm (accessed 6 Jan 2011).
16. **Massie BM,** Shah NB. Evolving trends in the epidemiologic factors of heart failure: rationale for preventive strategies and comprehensive disease management. *Am Heart J* 1997;**133**:703—12.
17. **Hripcsak G,** Friedman C, Alderson PO, *et al*. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995;**122**:681—8.
18. **Meystre S,** Savova G, Kipper-Schuler K, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008:128—44.
19. **Chung J,** Murphy S. Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. *AMIA Annu Symp Proc* 2005:131—5.
20. **Friedlin J,** McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc* 2006:269—73.
21. **Pakhomov S,** Weston SA, Jacobsen SJ, *et al*. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007:281—8.
22. *The Consortium for Healthcare Informatics Research (CHIR)*. http://www.research.va.gov/funding/solicitations/docs/Consortium-Healthcare-Informatics.pdf (accessed 6 Jan 2011).
23. *VA Informatics and Computing Infrastructure (VINCI)*. http://www.hsrd.research.va.gov/for_researchers/vinci/ (accessed 6 Jan 2011).
24. *VA corporate data Warehouse (CDW)*. http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm (accessed 6 Jan 2011).
25. *Text Integration Utilities (TIU) Technical Manual; Version 1.0*. 1997. Revised June 2010. http://www.va.gov/vdl/documents/Clinical/CPRS-Text_Integration_Utility_(TIU)/tiutm.doc
26. **Ogren PV,** Savova G, Buntrock JD, *et al*. Building and evaluating annotated corpora for medical NLP systems. *AMIA Annu Symp Proc* 2006:1050.
27. **Musen MA,** Gennari JH, Eriksson H, *et al*. PROTEGE-II: computer support for development of intelligent systems from libraries of components. *Medinfo* 1995;**8**:766—70.
28. **Roberts A,** Gaizauskas R, Hepple M, *et al*. The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc* 2007:625—9.
29. **Ashton C,** Kuykendall D, Johnson ML, *et al*. An Empirical assessment of the Validity of explicit and Implicit Process-of-care Criteria for quality assessment. *Med Care* 1999;**37**:798—808.
30. **Ogren PV.** *Knowtator*. http://knowtator.sourceforge.net/ (accessed 6 Jan 2011).
31. **Hripcsak G,** Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform* 2002;**35**:99—110.
32. **Stanford Center for Biomedical Informatics Research.** *What is protégé-owl?* http://protege.stanford.edu/overview/protege-owl.html (accessed 6 Jan 2011).
33. **Ogren PV.** *Knowtator*. http://www.orbitproject.org/resource/knowtator (accessed 6 Jan 2011).
34. **Ogren PV.** *Knowtator: a Protégé plug-in for Annotated Corpus Construction*. http://knowtator.sourceforge.net/docs/Ogren_HLT-NAACL06_Demo_Abstract_Final.pdf (accessed 6 Jan 2011).
35. **Ferrucci D,** Lally A. *'UIMA: an Architectural Approach to Unstructured Information Processing in the Corporate Research Environment,' Natural Language Engineering*. **Vol 10**. 2004:327—48.
36. **Heubusch K.** Meaningful Use White Paper Series. Paper no. 8: Preparing for meaningful use. *J AHIMA* 22 September 2010. http://www.ahima.org/advocacy/arrameaningfuluse.aspx

Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure

Jennifer H Garvin, Scott L DuVall, Brett R South, et al.

J Am Med Inform Assoc published online March 21, 2012
doi: 10.1136/amiajnl-2011-000535