Automated Identification of Episodes of Asthma Exacerbation for Quality Measurement in a Computer-Based Medical Record

David B. Aronow, MD, MPH,* ** James R. Cooley, MD,* Stephen Soderland, MS**
* Clinical Quality Management, Harvard Community Health Plan, Brookline MA
** Center for Intelligent Information Retrieval, University of Massachusetts, Amherst MA
{aronow soderlan}@cs.umass.edu

Harvard Community Health Plan and the Center for Intelligent Information Retrieval are developing tools to support automated quality of care measurement from clinical text data. A statistically based text classification system uses semantic features in computerized encounter notes to identify acute exacerbations of asthma. Individual encounter notes are sorted in bins of highly likely, highly unlikely and uncertain likelihood of documenting exacerbation, and then aggregated into episodes of exacerbation for frequency analysis. It is estimated that this approach could reduce the burden of manual chart review by 65%.

INTRODUCTION

Harvard Community Health Plan (HCHP) has benefited from computerized medical records since its founding 25 years ago. The data within the Automated Medical Record System (AMRS), in addition to serving as a patient's medical record, is a strategic resource for both quality improvement and utilization management.[1] Coded portions of AMRS are used extensively, however text portions, which include both dictated and hand-written provider notes, are relatively inaccessible, except by resource-intensive manual chart review.

The Center for Intelligent Information Retrieval (CIIR) is a National Science Foundation supported State-Industry-University consortium located in the Computer Science Department of the University of Massachusetts in Amherst. CIIR was established to focus research in text-based information systems and to facilitate transfer of new technologies to industry. FIGLEAF is a CIIR-developed text classification system based on statistical analysis of semantic features. HCHP joined CIIR to develop state-of-the-art tools to access clinical text data resources by computerized means.[2]

Quality Measurement at HCHP and the Asthma Care Project

HCHP recognizes that measurement of key components of health care processes and outcomes is essential to enable both the management and the improvement of these clinical processes. Corporate focus has been directed at clinical topics which are high frequency, high cost, high variation in current clinical management, or for which clinical guidelines have been developed.

Asthma is the most frequent diagnosis for emergency department visits and hospitalizations for children at HCHP, and national medical literature is documenting increasing rates of morbidity and mortality from asthma across the US.[3] Recent asthma care guidelines emphasize the use of anti-inflammatory medications to ameliorate the asthma course by decreasing the number of asthma exacerbations, which should decrease the need for emergency department visits and hospitalizations.[4]

In 1993, HCHP initiated an Asthma Care Project to measure key components of asthma care of children. Most specific measures were accomplished through automated review of coded clinical data. The most difficult measure, requiring significant medical chart review by trained coders, was identification of the number of episodes of asthma exacerbation per patient during the study year. The study team reasoned that as the percent of asthmatics routinely using anti-inflammatory medications increased, the frequency of acute exacerbations would decrease. Thus an initial exacerbation rate would serve as a baseline to which future measures, performed after guideline implementation, would be compared.

Exacerbation measurement required review of the context and content of each encounter to decide (a) if the encounter was for asthma and (b) if the asthma encounter note documented an acute exacerbation. Because this task was so time and effort consuming, the Asthma Care Project was selected as a case study for the pilot project of the HCHP-CIIR collaboration.

0195–4210/95/$5.00 © 1995 AMIA, Inc.

The general question HCHP posed to CIIR was: to what extent can automated information systems replace manual review of medical record encounter notes in support of quality measurement. Specifically for this project, can computer systems identify those encounters of pediatric asthmatics which concern acute exacerbations, approximately 5% of their encounters. Further, how well do these results aggregate to an exacerbation episode count, which is the manual chart review task.

This is a classification problem, in that a large collection of documents need to be sorted according to a specified set of characteristics. In this case, medical record encounter notes need to be sorted according to evidence that an acute exacerbation of asthma has been documented. We defined the process as a three bin sort: one bin containing encounter notes classified as very probably exacerbation, a second bin as very probably not exacerbation , and a third bin for encounters about which the classifier was uncertain. Intensive manual chart review could then be directed to the much smaller volume of notes in the Uncertain Bin.

FIGLEAF Inductive Text Classification System

The FIGLEAF (FIne Grained Lexical Analysis Facility) text classification system bases its classification on statistics derived from a set of training documents.[5] FIGLEAF computes weights based on the relative probability that a semantic feature (a term, bigram or phrase) occurs in a positive or a negative document. Documents are ranked according to the sum of the weights of features found in the document. Higher weights indicate a greater belief that the document is positive for the classification being trained. Thresholds may be set on the document weights to sort documents into three bins: positive, negative, and uncertain.

Adding evidence from each feature will only give an accurate ranking if each piece of evidence is independent. For each document, FIGLEAF analyzes its table of feature frequencies to identify sequences of terms that should be considered as single units when computing document weights.

METHODS

Clinical data preparation and benchmarking

In the 1993 Asthma Care Project, a group of 76 patients aged 2 to 18 years was randomly selected from the 6,744 asthmatics identified in this age range. The complete medical records of these children included 965 HCHP encounters (12.7 per patient) in the study period (10/1/91-9/30/92) and serve as the test collection for these experiments.

A training collection of 10/1/91-9/30/92 encounter notes was prepared by extracting the complete records of a different 100 children randomly selected from the same pool of 6,744 asthmatics. These children had 1,368 encounters (13.7 per patient).

Approximately 5% of the visit encounter notes in AMRS are transcribed dictations. The rest are hand-written provider notes, manually entered letter-for-letter as written by trained inputters. Figures 1 and 2 present typical encounter notes for acute asthma exacerbation and documentation of non-acute respiratory care respectively.

Figure 1. Acute Asthma Exacerbation
Typical AMRS Encounter Notes

G100 ASTHMA
COUGH & WHEEZE X1-2D HX PNEU 4/92 AFEB ACTIVE RR=48 W/MOD RETRAX CHEST DIFFUSE EXP WHEEZING & RHONCHI ONLY SL. CLEARING AFTER 2 NEBS O2 SATS 93 -> HOSP ER.

G100 ASTHMA
S/T,WHZ,SL COUGH,AFEB.H/O ASTHMA/PNEUMONIA 6/92.PE:AFEB, AUDIBLE WHZ,TM'S CLR,PHAR ERYTH,NO EDEMA OR EXUD.1(+)ADENOP CHEST-DIFFUSE WHZ,NO RALES,RHONCHI.0.25 VENT BY NEB -> LUNGS CLR.REFR 320.T/C SENT.

Figure 2. Non-Acute Respiratory Care
Typical AMRS Encounter Notes

G103 BRONCHOSPASM
HX OF NIGHT TIME WHEEZE,CLEARS W/VENT INHALER,LUNGS CLEAR NOW.DISC ENVIRONMENTAL CHANGES,COVER MATTRESS,& DUST ETC.

G100 ASTHMA
NOTES SOB W/VIGOROUS RUNNING SPORTS DESPITE GD CONDITION MOM HAS ASTHMA TRIAL VENTOLIN B/4 EXERCISE

In order to mimic the expected implementation of a classifier at HCHP, an automated code filter was implemented to reduce the number of grossly irrelevant encounter notes. Based on the provider assigned codes recorded at each encounter, both the training and test collections had removed from them encounter notes containing no occurrence of any of:

- a code for Asthma or asthma-like conditions (Acute Bronchitis, Bronchiolitis and Bronchospasm)
- the code Diagnosis Deferred (meaning no definitive assessment or diagnosis was made)
- any out-of-Plan referral or hospital arrangements

This left 231 encounter notes (24% of the original and 3.0 per patient) in the test collection and 357 in the training (26% of the original collection and 3.6 per patient).

To validate the code filtration process, all encounter notes which were filtered out of the test collection were reviewed for evidence of being asthma exacerbations. Four notes were found to be positive exacerbations, one with each of the following coded assessments: Respiratory Disorder (Not Specifically Coded), Pleurisy, R/O Pneumonia, and Cough.

Finally, benchmarks were established in order to train and test FIGLEAF. Every encounter in both collections was reviewed independently by two clinicians and scored as exacerbation, uncertain or not exacerbation, in accordance with the definitions used by the 1993 project. All uncertains and inter-reviewer disagreements were rescored as a positive or a negative in reconciliation meetings. In the end, the training collection had 133 of 357 encounters scored positive for exacerbation (37%) and the test collection had 99 of 231 encounters positive (43%). Note that prior to the code filtration, approximately 10% of encounters were positive exacerbations .

FIGLEAF Classification of Training and Test Encounter Notes

A feature frequency table was generated for the entire training collection. Features which occurred in fewer than six documents were discarded. Using FIGLEAF's algorithms, feature weights were then calculated. Figure 3 shows a sample of feature frequencies and weights. The numbers following "+" and "-" are the number of positive and negative documents in which the feature occurred.

Figure 3. Sample Features with Frequency in Positive and Negative Documents and Weight

WORSE :: +6 -9	0.191406
PRED :: +11 -2	56.9735
LUNGS_CLEAR :: +2 -15	-46.8144
WHEEZE :: +94 -57	15.8711
MILD_WHEEZE :: +10 -2	53.9307
EXP_WHEEZE :: +27 -4	63.1
EVALUATION :: +0 -15	-100
BNFT_COORD_APPROVE_BNFT :: +3 -13	24.6721

The training collection of 357 documents was rank ordered by FIGLEAF using ten-fold cross-validation. This involved sequentially assigning document weights to each one tenth of the documents in the training collection, based on feature weights derived from the remaining nine-tenths of the collection. The ten lists of document weight for partial collections were combined to rank the entire collection.

To sort the documents into bins, document weight cut-offs for the positive bin and the negative bin were determined such that the positive bin contained no more than a target percentage of negative documents (False Positives) and the negative bin contained no more than a target percentage of positive documents (False Negatives).

Using feature weights derived from the training collection, document weights were calculated for the test documents. Documents with weight above the positive bin cut-off are assigned to the Positive Encounter Bin, those below the negative bin cut-off to the Negative Encounter Bin, and the remainder to the Uncertain Encounter Bin.

Aggregation of Encounter Notes into Episodes

The quality of care metric used in the Asthma Care Project was the number of episodes of asthma exacerbation per patient, rather than the number of exacerbation encounters. This is because the episode is believed to be a more clinically significant unit of measure of the severity of asthma and the impact of interventions. To aggregate individual encounters into episodes, the manual chart reviewers used the rule that exacerbation encounters within seven days of each other were to be considered one episode.

To aggregate encounters for the automated classification episode benchmark, the positive exacerbation encounters in the test collection were manually mapped on timelines for each patient and the exacerbation episodes verified by review of the encounter notes. Sixty five exacerbation episodes were identified in the seventy six test collection patients. The number of episodes per patient ranged from zero to four, with an average of 0.86. The number of positive benchmarked encounters per exacerbation episode ranged from one to seven, with an average of 1.5 encounters per episode.

Converting FIGLEAF's classification of the test collection encounter notes into the newly defined exacerbation episodes was accomplished using the following rules:

1. if any encounter note in an episode was classified as positive in the test, the episode is classified as positive
2. in remaining episodes, if any encounter note was classified as uncertain in the test, the episode is classified as uncertain
3. remaining episodes had all encounter notes classified as negative and were themselves classified as negative

RESULTS

The result of the FIGLEAF classifier with target cut-off values of 10% for the allowable false positive and false negative are shown in Table 1. Forty six of 99 notes (46%) were correctly sorted into the Positive Encounter Bin, with two false positives. Forty five of 231 notes (24%) were correctly sorted into the Negative Encounter Bin, with 10 false negatives. If reviewers accept the encounter notes in the Positive and Negative Encounter Bins as exacerbations and not-exacerbations respectively, leaving only the Uncertain Encounter Bin notes to be scored by hand, the burden of manual review is reduced by 45%.

Table 1. Three Bin Sort of Asthma Exacerbation Encounter Notes

	Positive Encount Bin	Uncertain Encount Bin	Negative Encount Bin	
Exacerb	46 TP=96%	43	10 FN=18%	99
Not Exacerb	2 FP=4%	85	45 TN=82%	132
	48 (21%)	128 (55%)	55 (24%)	231

The effect of converting the classified encounter notes into episodes is shown in Table 2. Of 65 benchmark exacerbation episodes, 40 (62%) were correctly sorted to the Positive Episode Bin, 23 (35%) were Uncertain Episodes, and 2 (3%) were incorrectly sorted to the Negative Episode Bin. The false positive rate on the episode level is 5%.

While only two episodes were incorrectly classified as negative, the false negative rate is unknown. In a sense, all non-exacerbation encounters, except for those in the two false positive episodes, were corrected grouped in the Negative Episode Bin. However, the false negative rate cannot be calculated due to the lack of satisfactory definition of episodes of care without exacerbations. If record reviewers accept the episodes in the Positive and Negative Episode Bins as exacerbations and not-exacerbations respectively, leaving only the encounter notes in the Uncertain Bin to be read and scored by hand, the burden of manual chart review is reduced by 65%.

Table 2. Three Bin Sort of Asthma Exacerbation Episodes

	Positive Episode Bin	Uncertain Episode Bin	Negative Episode Bin	
Exacerb	40 TP=95%	23	2	65
Not Exacerb	2 FP=5%	unknown	unknown	

DISCUSSION

The first task in the Asthma Care Project, defining what we mean by an exacerbation of asthma, remained a challenge throughout this study. Wheezing and fever in a child less than two years may be labeled "asthma" by one clinician, "bronchiolitis" by another, and "acute bronchitis" by a third. Treatment may be similarly variable, with one clinician giving albuterol, another albuterol and steroids, a third mist by nebulizer only, and a fourth antibiotics only. Whether the classification is being done by a manual coder or automated review, the task is to code to a common standard, applying a consistent definition of acute asthma exacerbation to encounter notes written by many different clinicians. Even with manual record review by an experienced clinician, there will be some uncertainty about exacerbation and episode classification of records written by another clinician.

Manual reviewers followed guidelines such as presence in the encounter note text of words, or ad hoc provider abbreviations for words, such as "attack", "worse wheezing", "respiratory distress" and "retractions", the use of medications such as albuterol by nebulizer and oral steroids, and referral to an Emergency Department or hospitalization.

FIGLEAF uses a process conceptually analogous in that words, word abbreviations and phrases are statistically identified as providing evidence that an encounter note does or does not record an acute exacerbation. FIGLEAF's table of features included many of the manual reviewer guidelines terms and phrases, as well as several unanticipated by the clinician reviewers, such as "SAME DAY" as the value for the field Encounter Type, indicating an urgent visit.

One of the challenges to automated statistical classification systems is to sufficiently analyze context to differentiate phrases such as "WHEEZING LAST NIGHT" from "WHEEZES AT NIGHT". In our study, the false negative rate at the encounter note classification level of 18% exceeded our target of 10% and would not be acceptable in a production environment. The failure of the training to produce the intended bin cut-offs in the test collection is in large part due to subtleties in context that occur at different frequencies in the two collections. Research groups within CIIR are refining document classification algorithms and experimenting with alternative statistical classification approaches.

The second challenge from the Asthma Care Project was the identification of exacerbation episodes through time-based association of individual encounters. The one week time window used by the manual reviewers would have correctly grouped 62 of the 65 benchmark episodes and could serve as an acceptable approximation.

To more fully automate episode level classification we are considering a set of programmable rules to aggregate encounters. The "Aggregator Module" would roll a one week time window through encounter dates, and trigger an episode bundler at each new (i.e. unbundled) positive or uncertain classified encounter. Encounters classified positive would trigger the bundling of all other encounters within seven days into a positive episode. Unbundled uncertain encounters trigger an uncertain episode, subject to manual review, unless a positive encounter was found within seven days. As documented exacerbation episodes do sometimes extend for up to two weeks, the logic rules would be elaborated to accommodate encounters in days eight through fourteen from the episode triggering encounter.

CONCLUSIONS

The CIIR developed FIGLEAF text classification system can be applied to a clinical quality measurement problem involving encounter note coding. Reduction in the burden of charts requiring manual review would allow cost savings and/or larger study sizes. Chart reviewers could direct their efforts to training set annotation and the coding of uncertain episodes.

Acknowledgments

This work was supported by the Harvard Community Health Plan Foundation and by the National Center for Intelligent Information Retrieval at the University of Massachusetts, under National Science Foundation Grant EEC-9209623.

References

1. Schoenbaum SC, Barnett GO. Automated ambulatory medical records systems: An orphan technology. *International Journal of Technology Assessment in Health Care.* 8:598-609, 1992.
2. Aronow DB, Soderland S, Ponte JM, Feng F, Croft WB, Lehnert WG. Automated classification of encounter notes in a computer based medical record. In *Proceedings of MEDINFO '95 8th World Congress on Medical Informatics (in press).* 1995.
3. Evans R, Mullaly DI, Wilson RW, et al. National trends in asthma: Morbidity and mortality of asthma in the United States. *Chest.* 91:658, 1987.
4. National Asthma Education Program. *Guidelines for the Diagnosis and Treatment of Asthma.* Office of Prevention, Education and Control, National Heart, Lung and Blood Institute; August 1991. NIH Publication No. 91-3042.
5. Lehnert W, Soderland S, Aronow D, Feng F, Shmueli A. Inductive text classification for medical applications. *Journal of Experimental and Theoretical Artificial Intelligence,* 7:49-80, 1995.