Using A Natural Language Processing System to Extract and Code Family History Data from Admission Reports

Jeff Friedlin, DO,[1] Clement J. McDonald MD.[1]

[1]Regenstrief Institute, Inc, and Indiana University School of Medicine, Indianapolis, IN.

Abstract

We developed a rule-based natural language processing (NLP) system for extracting and coding clinical data from free text reports. We studied the systems ability to accurately extract and code family history data from hospital admission notes. The system searches the family history for 12 diseases (and relative degree). It achieved a sensitivity of .96 and a PPV of .97 for disease extraction, and .96 and .93 respectively for relative categorization.

Introduction

Physicians prefer the freedom and ease of free speech dictation when creating patient reports. The admission note (or history and physical) frequently contains the patient's family history. Knowledge of a patient's family history can be important when making clinical decisions, including the timing of preventive screening procedures such as mammograms and colonoscopies. It can also be important in assessing a patient's risk for certain inheritable diseases, such as coronary artery disease[1]. Because it is usually stored in free text format, the family history is rarely used by applications such as clinical decision support systems. A need exists to accurately extract and code family history data from free text reports to enable its use by such systems.

Methods

We used admission notes dictated at St. Francis Hospital during the period from December 2005 to January 2006 as our data source. St. Francis Hospital is a multi-center 434-bed primary care general hospital system in Indianapolis, Indiana.

The NLP system we used is the REgenstrief data eXtraction tool (REX), developed at the Regenstrief Research Institute in Indianapolis, Indiana. REX first locates and extracts the family history section (if present) from the admission notes, and then attempts to find the following 12 diseases: coronary artery disease, diabetes, colon, breast and ovarian cancer, hypertension, stroke, sickle cell disease, asthma, depression, Alzheimer's disease, and osteoporosis. Using proximity phrase searching, it determines negation and in what context the disease is found. If a disease is found, it categorizes the relative mentioned in connection with the disease as either primary, secondary or unknown. An experienced board certified family physician reviewed all admission notes to create the gold standard.

Results

REX analyzed a total of 1337 admission notes, and detected a family history section in 878 (66%) reports. We excluded family histories that contained only nonspecific data, such as "noncontributory" or "unobtainable". This left 613 family histories for analysis. Mention of one of the 12 diseases occurred 522 times, a primary relative was noted 158 times, and a secondary relative 25 times. Table 1 shows the positive predictive value (PPV) and sensitivity measurements of REX compared to the gold standard. REX determined that a disease occurred in the family history a total of 518 times. Of these 518 instances, 15 were deemed false by the gold standard (97% correct). The gold standard found that a disease occurred in the family history in 522 instances. Of these 522, REX missed 19 (4% missed). Regarding relative categorization, REX coded a relative as either primary or secondary a total of 178 times. Seven of these were considered errors by the gold standard (96% correct). The gold standard found 183 instances of either a primary or secondary relative. REX correctly categorized all but 12 of these (7% missed).

	PPV	Sensitivity
Diseases[a]	.97	.96
Relative catagorization[b]	.96	.93

Table 1. Outcome measures of REX compared to gold standard

[a]an aggregate of all 12 diseases

[b]aggregate of both primary and secondary relatives

Conclusion

Admission notes contain large amounts of often untapped data, including family history. REX is easy to implement, highly accurate, and cost effective for the extraction of this data. This valuable information can then be used for research, genetic counseling and real-time clinical decision support.

References

1. Fiszman M, et al. The Family History -- More Important Than Ever N Engl J Med 2004 351: 2333-2336