Automated Detection of GI Syndrome using Structured and Non-Structured Data from the VA EMR

Brett R South, MS[1], Adi V Gundlapalli, MD, PhD[1], Shobha Phansalkar, RPh, MS, PhD[1], Shuying Shen, MS[1], Sylvain Delisle, MD, MBA [2], Trish Perl, MD, MSc[3], Matthew H Samore, MD [1]

[1]VA Salt Lake City Health Care system and the Department of Medicine, University of Utah, School of Medicine, Salt Lake City, UT, USA, [2]VA Maryland Health Care System and University of Maryland, School of Medicine, Baltimore, Maryland, USA, [3]The Johns Hopkins Hospital, Baltimore, Maryland, USA

OBJECTIVE

We performed a gold-standard manual chart review for gastro-intestinal (GI) syndrome to evaluate automated detection models based on both structured and non-structured data extracted from the VA electronic medical record (EMR).

METHODS

We randomly sampled 15,377 of 253,818 outpatient visits to the VA Maryland Health Care system (VAMHCS) and the VA Salt Lake City Health Care system (VASLCHCS) during the 10/01/03 to 3/31/04 study period. "GI syndrome" cases were defined as follows: vomiting or diarrhea or abdominal pain lasting less than 7 days AND illness not attributable to a non-infectious etiology. For automated case detection, we used provider-assigned ICD-9 diagnostic codes and their free-text documentation of index outpatient encounter extracted from the VA EMR. ICD-9 detection models included "GI" ICD-9 code sets used in the "ESSENCE" and the "BioSense" syndromic surveillance systems. Case detection based on text-processing methods began by mapping symptoms from the case definition to UMLS concepts. We then used the NegEx[1] negation algorithm adapted to VA notes to identify "Cases" to determine if the full text of any notes written on the day of the sampled patient encounters (n=76,500) included at least one non-negated UMLS GI concept. Notes were also processed using MedLEE[2] a natural language processing (NLP) system to identify epidemiologic factors useful for case investigation such as previous exposure to infection or duration of illness. Additionally, we searched for documentation on the index visit day of fever $\geq$ 37.8ºC.

RESULTS

The ESSENCE and BioSense ICD-9 code sets detected 242 GI syndrome cases (sample prevalence: 1.57%). The NegEx algorithm for text-processing detected 2,338 visits with non-negated vomiting or diarrhea or abdominal pain. Altogether, 43 visits met the GI clinical case definition on the basis of chart review (sample prevalence: 0.28%). ICD-9 codes alone had higher specificity, but lower sensitivity than text-processing for ascertainment of clinically defined GI syndrome cases (Table). Use of the "OR" operator in combined models improved sensitivity and the area under the ROC. Use of the "AND" operator in combined models enhanced specificity and positive predictive value. MedLEE identified exposure to infection and illness duration in 19 and three of the GI syndrome cases respectively. Only four of the clinically defined GI syndrome cases had fever.

Summary GI Case Detection Models

Case detection Model	Sens (%) (95% CI)	Spec (%) (95% CI)	ROC (95% CI)
M1. Any NegEx	95 (84,99)	85 (84,86)	90 (87,93)
M2. "GI" ICD-9	54 (38,69)	99 (98,99)	76 (69,84)
M3. M2 AND M1	51 (36,67)	98 (98,99)	75 (67,82)
M4. M2 OR M1	100 (92,100)	84 (83,84)	92 (92,92)

CONCLUSIONS

Case detection models based on text-processing alone or combined text and ICD-9 code models out-performed ICD-9 code based models alone. The greatest precision can be achieved when combined models using the "AND" operator are used for case detection. Combining non-structured with structured data sources could serve as a useful screening method to identify cases for further epidemiologic investigation. However, improvements in clinical documentation of exposure to infection and illness duration are needed. Future efforts will include building and statistically validating case detection models based on an expanded group of clinical data elements relevant to GI syndrome.

REFERENCES

1. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.

2. Friedman C. Towards a comprehensive medical language processing system: methods and issues.Proc AMIA Annu Fall Symp. 1997; :595-9.