Text-Processing of VA Clinical Notes to improve case detection models for Influenza-Like-Illness

Brett R South, MS[1], Shobha Phansalkar, MS, RPh[1], Ashwin Deepak Swaminathan, BS[1] Jill Anthony, MS, Ph.D.[2], Sylvain Delisle, MD, MBA [3], Trish Perl, MD, MSc[4], 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]Johns Hopkins University, Department of Molecular Microbiology and Immunology, Bloomberg School of Public Health, [3]VA Maryland Health Care System and University of Maryland, School of Medicine, Baltimore, Maryland, USA, [4]The Johns Hopkins Hospital, Baltimore, Maryland, USA

OBJECTIVE

There were two objectives of this analysis. First, apply text-processing methods to free-text clinician notes extracted from the VA electronic medical record for automated detection of Influenza-Like-Illness (ILI). Secondly, determine if use of data from free-text clinical documents can be used to enhance the predictive ability of case detection models based on coded data.

METHODS

A randomized sample of 15,377 outpatient encounters from the VA Maryland Health Care system (VAMHCS) and VA Salt Lake City Health Care system (VASLCHCS) was selected for chart review during the study period 10/01/03 to 3/31/04. ILI Cases were identified using an explicit definition of ILI. Following chart review all electronic notes for positive cases and a random sample of all negative cases (N=2,414 cases) were selected for text-processing. ILI concepts from the case definition were mapped to UMLS concepts using the National Library of Medicine's Metathesaurus® search tool. Three automated text-processing techniques were applied: string matching for relevant UMLS concepts; UMLS concepts in conjunction with a negation algorithm called NegEx[1]; and an NLP system called MedLEE[2]. For all methods presence of any two non-negated UMLS concepts per note denoted ILI. Predictive values of each of the text-processing techniques alone and with addition of ESSENCE based ICD9 codes (VAMHCS 197) and clinical data (e.g. expectorant orders) were estimated in comparison to chart review.

RESULTS

Chart review identified 279 true positive cases. Automated search using string matching for UMLS concepts, NegEx and MedLEE identified 263, 247 and 255 cases respectively. Sensitivity, specificity, and PPV for each of these methods was: UMLS concepts (94%,71%,6%), NegEx (88%,94%,21%) and MedLEE (91%,91%,16%). Combining text-processing methods with ESSENCE based ICD-9 codes or adding expectorant orders improves model performance. String matching methods alone provide high sensitivity, but low specificity. When compared against each other, NegEx is as performant as MedLEE. However, models combining NegEx with ESSENCE based ICD-9 codes or expectorant orders provide high specificity and sensitivity for detection of ILI cases. (Table 1).

Table 1. Predictive values for Case Detection Models*

Model	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	ROC area % (95% CI)
1. One or more VAMHCS 197 ICD-9 codes	79 (74,84)	97 (96,98)	31 (26,36)	88 (85,90)
2. Model 1 OR expectorant	84 (79,88)	95 (95,96)	25 (22,29)	90 (88,92)
3. UMLS	94 (91,97)	71 (67,73)	6 (5,6)	82 (81,84)
4. Model 2 and 3	84 (79,88)	96 (95,96)	25 (22,29)	90 (88,92)
5. UMLS+NegEx	88 (84,92)	94 (93,95)	21 (19,24)	91(89,93)
6. Model 2 and 5	97 (95,99)	91 (89,92)	16 (14,18)	94 (93,95)
7. MedLEE	91 (87,94)	91 (90,93)	16 (14,18)	91 (90,93)
8. Model 2 and 7	96 (93,98)	89 (87,90)	13 (12,15)	92 (91,94)

* Negative predictive value for Model 6 was 98% and 99% for Models 1,2,3,4,5,7,8.

CONCLUSION

Text-processing techniques provide a means of utilizing electronic clinical documents as an additional data source for syndromic surveillance case detection models. Application of text-processing methods on VA electronic notes provides high specificity and sensitivity for detection of ILI cases.

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.