Text Mining of Clinical Progress Notes to Predict Future Onset of Sepsis in Hospitalized Patients

[1]Goh Kim Huat,[2]Adrian Yeow Yong Kwang,[1]Wang Le,[3]Hermione Poh,[3]Li Ke,[3]Joannas Yeow Jie Lin,[3]Gamaliel Tan Yu Heng

[1]Division of Information Technology and Operations Management, Nanyang Technological University, Singapore

[2]School of Business, Singapore University of Social Sciences, Singapore

[3]Academic Informatics Office, Medical Informatics, National University Health System, Singapore

Introduction

Sepsis Epidemic—A Clinical Problem

The global epidemiological burden of sepsis is difficult to ascertain. It is estimated to affect more than 30 million people worldwide every year, potentially leading to 6 million deaths. **Early diagnosis and treatment is key.**

Roadblocks

•There is no gold standard in the definition of sepsis

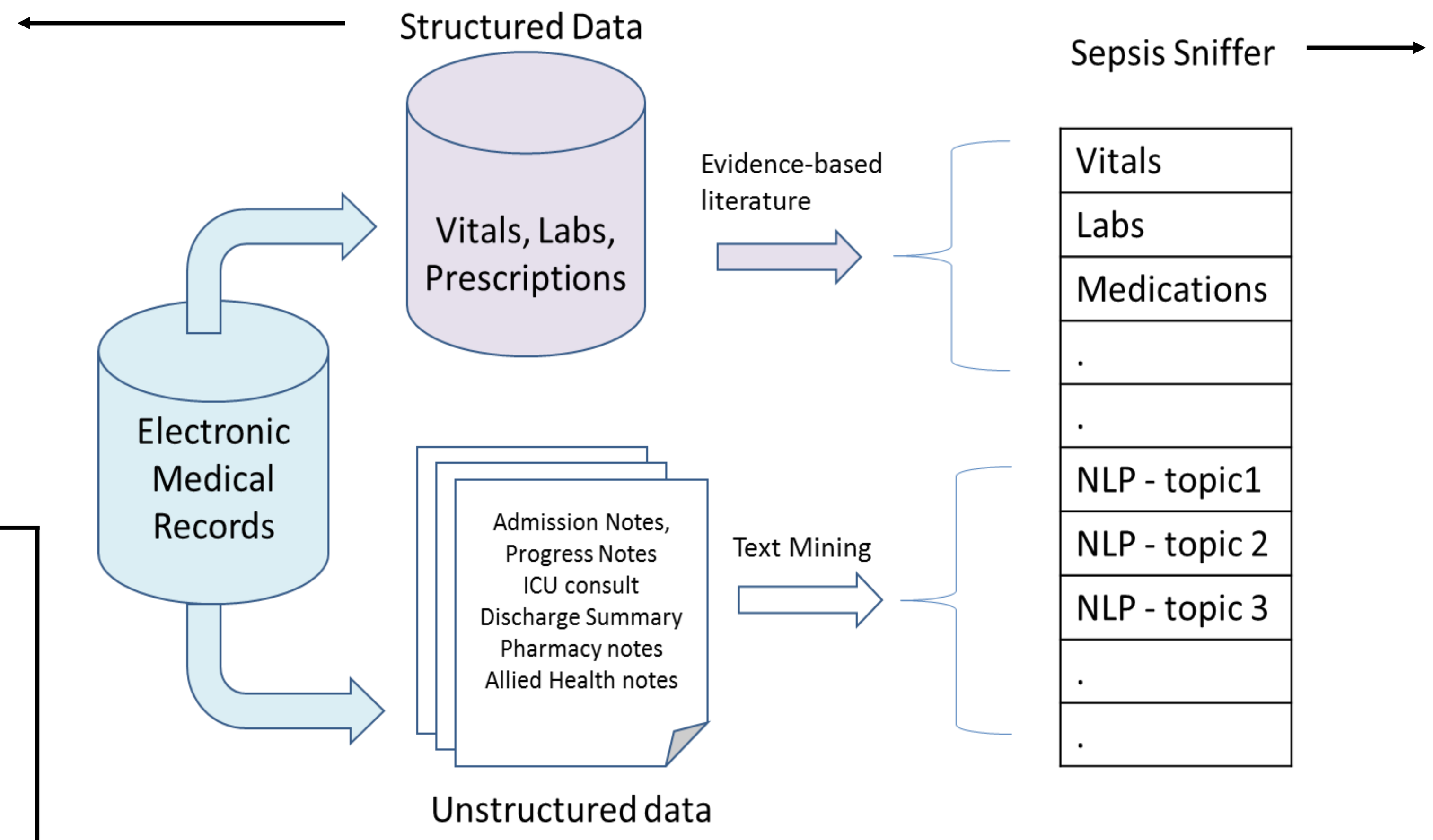
•Sepsis is easily confounded by other diseases, making diagnosis difficult

•Sepsis is a complex condition with many different interactions with other disorders.

Structured Data

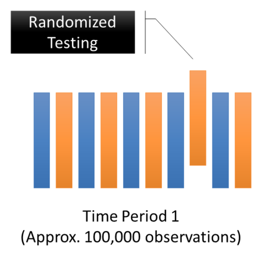
Category	Predictors
Patient information	Age, gender
Vital Sign	Blood pressure, heart rate, temperature, saturation, respiratory rate
Investigations	Total white cell, culture, lactate, high sensitive C-reactive protein, procalcitonin, arterial blood gas
Treatment	Vasopressor, antibiotics

Sepsis Sniffer Design



Model Development

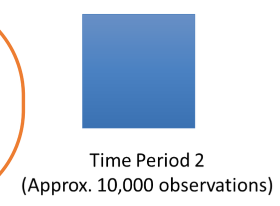
Phase 1 Training and Testing



No. of patients	7004
Age (Mean±SD)	63.71±17.08
LOS (Days)	5.52±14.31
Mortality(%)	4.5
Sepsis(%)	23.87

Medical records of the septic patients **prior** to the onset of sepsis during their hospitalization was used to train and test the predictive model.

Phase 2 External Validation



No. of patients	778
Age (Mean±SD)	63.90±16.81
LOS (Days)	5.17±10.80
Mortality(%)	5.01
Sepsis(%)	24.29

A hold-out sample with patients' septic status masked was used to test the accuracy of the predictive model developed.

Topic Categories

Category	Topic Count	Definition
Clinical Status	28	Routine updates of clinical conditions as well as diagnosis (e.g. vitals) excluding lab and radio-diagnostic tests
Communication	3	Communication between staff
Lab Test	24	Orders and reports of lab or radio-diagnostic test results
Non-Clinical Status	2	Routine updates of non-clinical conditions
Social Relationship	2	Information about family and social aspects of patient
Symptom	10	Clinical symptoms
Treatment	31	Treatment procedure or medication prescribed as well as the status of the treatment/ medication

Latent Dirichlet Allocation (LDA) was used to extract topics from the progress notes.

Natural Language Processing

Topic Clustering
- LDA routine
- 25,50,75,100, 150 topics

Filtering
- Weighting

Parsing
- Lemmatization
- Removal of stop words
- Part of **Speech** Tagging

Cleaning
- Anonymizing
- HIPAA

Rationale

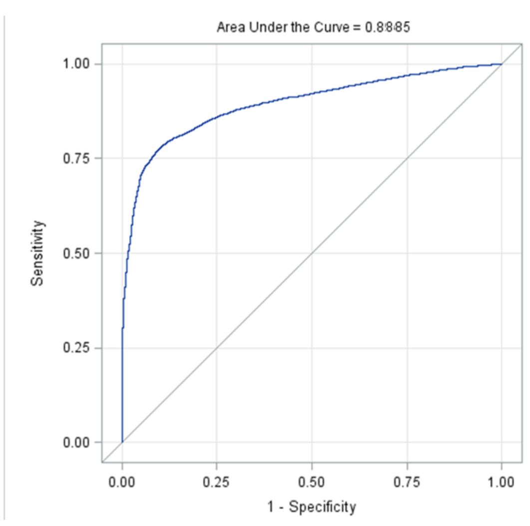
•Develop and test models independently

•Focus on pre-sepsis data and not post sepsis: transfer to ICU as trigger event for onset of sepsis

•Ensure that the lexicon developed from the first phase is *consistent over time* and applicable to the second phase

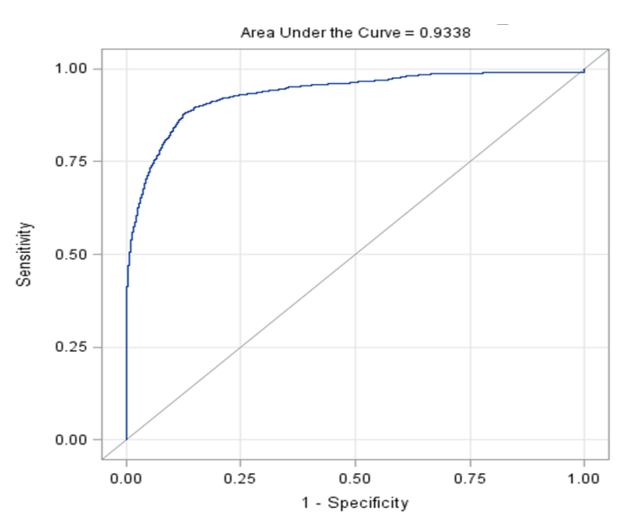
Results

Without Text Mining



Validated: AUC **0.8885**

With Text Mining



Validated: AUC **0.9338**

Comparison

Reference	Data	Method	Sensitivity (%)	Specificity (%)	AUC (%)
Our model	EMR	Logistics	-	-	93
Thiel et al. (2010)	cohort	RPART[1]	17	96[2]	-
Shashikumar et al. (2017)	EMR	Logistics	55	85[3]	78
Lukaszewski et al. (2008)	cohort	Neural networks	-	-	83
Mani et al. (2014)	EMR	Naive Bayes	-	-	78
Dummitt et al. (2018)	EMR	Survival analysis	-	-	87
Pereira et al. (2011)	cohort	Fuzzy C-Means	-	-	90
Futoma et al. (2017)	EMR	Gaussian process	85[4]		64
Desautels et al. (2016)	EMR	Machine learning	-	-	88
Henry et al. (2015)	EMR	Survival analysis	85[4]	67	83

Note: [1] RPART: Recursive Partitioning And Regression Tree; [2] At a specificity of 96%, our model can achieve a sensitivity of 60%, higher than 17%; [3] At a specificity of 85%, our model can achieve a sensitivity of 87%, higher than 55%; [4] At a sensitivity of 85%, our model can achieve a specificity of 86%, higher than 67%.

External Validation

The predictive model with textual information when tested on the hold-out sample produced a ROC AUC of 0.9338 (the predictive model without textual information – ROC AUC of 0.8885). The test results achieved a good balance between sensitivity (89%) and specificity (89%).

Conclusion

Text Mining Adds Value

The improvement in AUC in the sepsis predictor suggests that the addition of topics derived from text-mined progress notes provided additional value to the sepsis predictive model.

More importantly, based on our review of current medical studies, our predictive model achieves significantly better performance when compared to existing sepsis detection models which rely solely on structured variables.

References

Hall, M.J, et al. NCHS data brief, 62. 2011, Am J Respir Crit Care Med. 2016 Feb;193(3):259-72.
Crit Care. 2004; 8(6): R409–R413.,Torio CM and Andrews RM, 2013 (AHRQ HCUP Brief).
Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. 2015. "A Targeted Real-Time Early Warning Score (Trewscore) for Septic Shock," Science Translational Medicine (7:299), pp. 1-9.
Mani, S., Ozdas, A., Aliferis, C., Varol, H. A., Chen, Q., Carnevale, R., Chen, Y., Romano-Keeler, J., Nian, H., and Weitkamp, J.-H. 2014. "Medical Decision Support Using Machine Learning for Early Detection of Late-Onset Neonatal Sepsis," Journal of the American Medical Informatics Association (21:2), pp. 326-336.
Thiel, S. W., Rosini, J. M., Shannon, W., Doherty, J. A., Micek, S. T., and Kollef, M. H. 2010. "Early Prediction of Septic Shock in Hospitalized Patients," Journal of Hospital Medicine (5:1), pp. 19-25.