ISDS 2019 Conference Abstracts

Multidimensional Semantic Scan for Pre-Syndromic Disease Surveillance

Mallory Nobles[1], Ramona Lall[2], Robert Mathes[2], Daniel B. Neill[3]

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States, [2]NYC Dept. of Health and Mental Hygiene, New York, New York, United States, [3]Center for Urban Science and Progress, New York University, New York, New York, United States

Objective

We present a new approach for pre-syndromic disease surveillance from free-text emergency department (ED) chief complaints, and evaluate the method using historical ED data from New York City's Department of Health and Mental Hygiene (NYC DOHMH).

Introduction

An interdisciplinary team convened by ISDS to translate public health use-case needs into well-defined technical problems recently identified the need for new "pre-syndromic" surveillance methods that do not rely on existing syndromes or pre-defined illness categories [1]. Our group has recently developed Multidimensional Semantic Scan (MUSES), a pre-syndromic surveillance approach that (1) uses topic modeling to identify newly emerging syndromes that correspond to rare or novel diseases; and (2) uses multidimensional scan statistics to identify emerging outbreaks that correspond to these syndromes and are localized to a particular geography and/or subpopulation [2,3]. Through a blinded evaluation on retrospective free-text ED chief complaint data from NYC DOHMH, we demonstrate that MUSES has great potential to serve as a "safety net" for public health surveillance, facilitating a rapid, targeted, and effective response to emerging novel disease outbreaks and other events of relevance to public health that do not fit existing syndromes and might otherwise go undetected.

Methods

Multidimensional semantic scan uses topic modeling to learn illness categories directly from the data, eliminating the need for pre-defined syndromes. Topic models are a set of algorithms that automatically summarize the content of large collections of documents by learning the main themes, or topics, contained in the documents [4]. Our method learns two sets of topics: a set of topics over the historical data designed to capture common illnesses, and a set of emerging topics over only the most recent data that are optimized to capture any new illnesses not captured by the historical topics. We then use multidimensional scan statistics to identify clusters of cases isolated to a certain topic, hospital, and/or demographic group of patients [5].

To evaluate the ability of MUSES to detect a diverse set of emerging patterns relevant to public health in large and complex data, we apply our algorithm to historical chief complaint data from NYC. This dataset has over 28 million ED cases from 53 NYC hospitals during 2010-2016. For each hospital we have data on the patients' free-text chief complaint, date and time of arrival, age group, gender and discharge ICD-9 diagnosis code. Public health practitioners at NYC DOHMH performed a blinded evaluation of the top 500 highest-scoring clusters detected by our method and by a competing state of the art keyword-based approach [6-8]. For each of these clusters, the evaluators indicated if the cluster (1) represents a meaningful collection of cases and (2) is, in their judgement, of significant interest to public health.

Results

The blinded evaluation by NYC DOHMH demonstrated that our method correctly identifies a larger number of events of interest to public health than the baseline keyword-based scan method. 320 (64%) of the top 500 results from MUSES corresponded to meaningful health events, while the keyword-based method only detected 246 such events (49.2%). MUSES also identified 6 more highly relevant events and 74 less meaningless clusters than the keyword-based method. Figure 1 shows that for any fixed number of clusters that public health officials choose to examine, MUSES identifies more meaningful events than keyword-based scan. Alternatively, for any desired number of true clusters detected, MUSES exhibits substantially higher precision: for example, in order to identify 100 true clusters, it had to report 159 total clusters (precision = 63%) as compared to 225 total clusters (precision = 44%) for the keyword-based scan. This corresponds to a 53% reduction in the number of false positive clusters.

Additionally, to determine how our approach might provide situational awareness of emerging health concerns following a natural disaster, we examined the clusters identified by our approach in the week following October 29, 2012, when Hurricane Sandy struck New York City and caused a historic level of damage. These results show a progression of clusters from acute cases related to falls and shortness of breath, to mental health issues like depression and anxiety, to chronic health issues that require maintenance procedures, like dialysis and methadone distribution. It is of note that public health officials manually inspected emergency room data immediately following Hurricane Sandy and noticed an increase in the words "methadone", "dialysis" and "oxygen" [7]. The ability of MUSES to automatically identify similar symptoms as human experts highlights its ability to learn meaningful but novel combinations of symptoms.

Conclusions

Our MUSES system offers a novel method for pre-syndromic surveillance that achieves the goals set forth by public health practitioners during the ISDS Consultancy. When evaluated against a state of the art baseline, MUSES identifies a larger number of events of interest, has a lower false positive rate, and produces more coherent results. This ability to report newly emerging case clusters of high relevance to public health, without overwhelming the user with a large number of false positives, suggest high potential utility of the approach for day-to-day operational use as a "safety net" for public health surveillance, complementing existing syndromic surveillance approaches. We are currently building a pre-syndromic surveillance system based on the MUSES approach and plan to make this software widely available to public health partners in the near future.

Acknowledgement

We wish to thank the BCD Syndromic Surveillance Unit at NYC Department of Health and Mental Hygiene for providing retrospective data for this study and for participating in the blinded evaluation, and the Department of Homeland Security Hidden Signals Challenge for providing funding support for system development.

References

1. Faigen Z, Deyneka L, Ising A, et al. 2015. Cross-disciplinary consultancy to bridge public health technical needs and analytic developers: asyndromic surveillance use case. *Online J Public Health Inform*. 7(3), e228. PubMed https://doi.org/10.5210/ojphi.v7i3.6354
2. Maurya A, Murray K, Liu Y, Dyer C, Cohen WW, et al. Semantic scan: detecting subtle, spatially localized events in text streams. 2016. arXiv preprint arXiv:1602.04393.
3. Nobles M, Deyneka L, Ising A, Neill DB. 2015. Identifying emerging novel outbreaks in textual emergency department data. *Online J Public Health Inform*. 7(1), e45. https://doi.org/10.5210/ojphi.v7i1.5710
4. Blei D, Ng A, Jordan M. 2003. Latent Dirichlet allocation. *J Mach Learn Res*. 3, 993-1022.
5. Neill DB. 2012. Fast subset scan for spatial pattern detection. *J R Stat Soc B*. 74(2), 337-60. https://doi.org/10.1111/j.1467-9868.2011.01014.x
6. Burkom H, Elbert Y, Piatko C, Fink C. 2015. A term-based approach to asyndromic determination of significant case clusters. *Online J Public Health Inform*. 7(1), e11. https://doi.org/10.5210/ojphi.v7i1.5675
7. Lall R, Levin-Rector A, Mathes R, Weiss D. 2014. Detecting unanticipated increases in emergency department chief complaint keywords. *Online J Public Health Inform*. 6(1), e93. https://doi.org/10.5210/ojphi.v6i1.5069
8. Walsh A, Hamby T, St John TL. 2013. Identifying clusters of rare and novel words in emergency department chief complaints. *Online J Public Health Inform*. 6(1), e146.

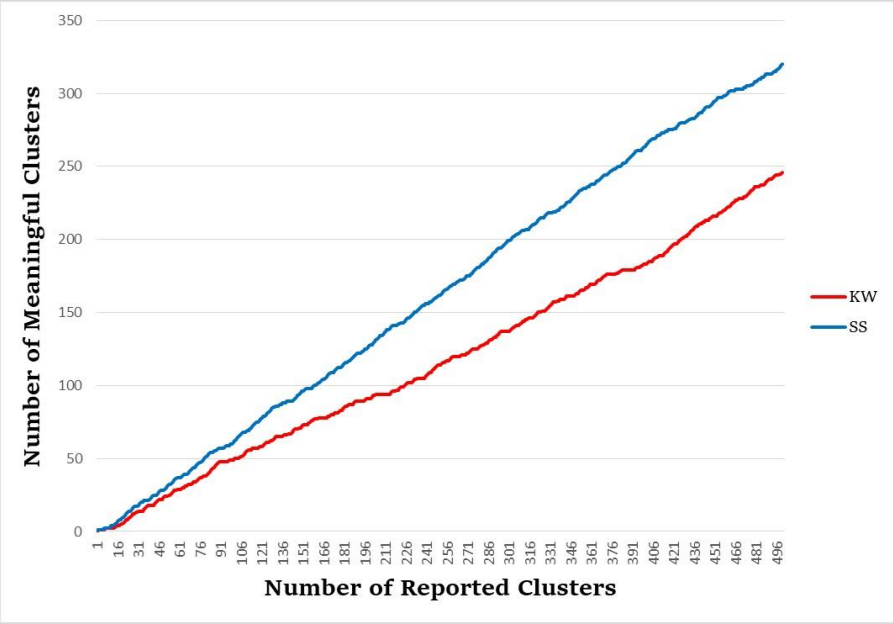

Figure 1: Performance comparison