Generalization of Machine Learning Approaches to Identify Notifiable Diseases Reported from a Statewide Health Information Exchange

Gregory Dexter[a], Suranga N. Kasthurirathne[a,b], Brian E. Dixon[a,b,c], Shaun J. Grannis[a,d]

[a] *Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA,*
[b] *Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA,*
[c] *Center for Health Information and Communication, Department of Veterans Affairs, HSR&D Service, Indianapolis, IN, USA,*
[d] *School of Medicine, Indiana University, Indianapolis, IN, USA*

Introduction

Machine learning (ML) presents much potential to operationalize secondary uses of healthcare data for public health informatics. Previous efforts have demonstrated the ability to use ML for better population surveillance and reporting [1]. Unfortunately, there is a dearth of methodological studies that assess the feasibility and performance characteristics of practical, generalizable ML methods that can be applied to use cases covering large geographical regions and a multitude of hospital systems. Public health informatics must embrace diversity of data from clinical settings and other sources to be useful in practice. This is particularly relevant given that restrictions on cross-organizational data sharing and variations in disease reporting rates [2] may impede the generalizability of ML solutions across locations. One noteworthy use case is government designated notifiable condition detection, where infectious disease cases are reported to public health authorities. We seek to evaluate the generalizability of ML solutions to predict cases of public health concern using data collected from a Health Information Exchange (HIE) network.

Methods

We extracted 1.7 million laboratory messages reported during 2016-2017 to the Indiana Network for Patient Care (INPC), a statewide HIE that facilitates interoperability among 117 hospitals and other free-standing laboratories and physician practices in the state of Indiana. Of these, we identified messages pertaining to Syphilis, Salmonella and Histoplasmosis. Each message was manually labelled as either positive or negative. Next, each message was vectorized with a bag of words approach for ML purposes. We used these vectors to train ML models, and assessed their predictive performance and generalizability. To assess overall model performance, we built a series of Random Forest decision models to predict each condition using 80/20% train/test of all messages. To assess generalizability, we iteratively withheld all messages reported by each of the larger integrated lab systems from our test dataset. This dataset was used to train a series of Random Forest classification models by condition. Each model was tested using the lab system specific holdout data. ML models were evaluated using sensitivity and specificity.

Results

The dataset consisted of 2,701 Syphilis results (24% positive), 6,790 Salmonella results (18% positive) and 3,310 Histoplasmosis results (21% positive). Condition specific models reported the following outcomes; (sensitivity : specificity); Syphilis (0.91 : 0.96), Salmonella (0.95 : 0.99), and Histoplasmosis (0.96 : 0.96). ML models trained using different lab systems as holdout data (lab system-specific models) reported varying performance depending on the lab system used as the holdout dataset. We list the range of sensitivity and specificity measures for each disease; Syphilis ([0.11-0.95], [0.70-1.00]), Salmonella ([0.14-1.00], [0.00-1.00]), and Histoplasmosis ([0.18-1.00], [0.86-1.00]).

Discussion

Condition specific ML models yielded high performance measures using train/test datasets uniformly sampled from all lab systems. However, predictive performance across each condition varied significantly when models were trained on data from a subset of lab systems and tested on the complementary (holdout) lab system data, suggesting low generalizability. These results are of significant value to population surveilance and public health efforts. They highlight the need to; (a) train ML models using datasets representative of all data sources that models will be applied to, and (b) standardize reporting of notifiable conditions across different laboratory systems. Next steps involve efforts to develop better methods to train generalized ML models using representative datasets and minimal human intervention for manual review.

References

[1] S.N. Kasthurirathne, B.E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin, and S.J. Grannis, Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection, *Journal of biomedical informatics* **60** (2016), 145-152.

[2] B.E. Dixon, Z. Zhang, P.T.S. Lai, U. Kirbiyik, J. Williams, R. Hills, D. Revere, P.J. Gibson, and S.J. Grannis, Completeness and timeliness of notifiable disease reporting: a comparison of laboratory and provider reports submitted to a large county health department, *BMC Medical Informatics and Decision Making* **17** (2017), 87.

Address for correspondence:
Gregory Dexter; grdexter@iu.edu