OJPHI

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

ISDS 2016 Conference Abstracts

Predicting Acute Respiratory Infections from Participatory Data

Bisakha Ray*[1] and Rumi Chunara[2]

[1]New York University School of Medicine, New York, NY, USA; [2]New York University, New York, NY, USA

Objective

To evaluate prediction of laboratory diagnosis of acute respiratory infection (ARI) from participatory data using machine learning models.

Introduction

ARIs have epidemic and pandemic potential. Prediction of presence of ARIs from individual signs and symptoms in existing studies have been based on clinically-sourced data[1]. Clinical data generally represents the most severe cases, and those from locations with access to healthcare institutions. Thus, the viral information that comes from clinical sampling is insufficient to either capture disease incidence in general populations or its predictability from symptoms. Participatory data — information that individuals today can produce on their own — enabled by the ubiquity of digital tools, can help fill this gap by providing self-reported data from the community. Internet-based participatory efforts such as Flu Near You[2] have augmented existing ARI surveillance through early and widespread detection of outbreaks and public health trends.

Methods

The GoViral platform[3] was established to obtain self-reported symptoms and diagnostic specimens from the community (Table 1 summarizes participation detail). Participants from states with the most data, MA, NY, CT, NH, and CA were included. Age, gender, zip code, and vaccination status were requested from each participant. Participants submitted saliva and nasal swab specimens and reported symptoms from: fever, cough, sore throat, shortness of breath, chills, fatigue, body aches, headache, nausea, and diarrhea. Pathogens were confirmed via RT-PCR on a GenMark respiratory panel assay (full virus list reported previously[3]).

Observations with missing, invalid or equivocal lab tests were removed. Table 2 summarizes the binary features. Age categories were: ≤ 20, > 20 and < 40, and ≥ 40 to represent young, middle-aged, and old. Missing age and gender values were imputed based on overall distributions.

Three machine learning algorithms—Support Vector Machines (SVMs)[4], Random Forests (RFs)[5], and Logistic Regression (LR) were considered. Both individual features and their combinations were assessed. Outcome was the presence (1) or absence (0) of laboratory diagnosis of ARI.

Results

Ten-fold cross validation was repeated ten times. Evaluations metrics used were: positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity[6]. LR and SVMs yielded the best PPV of 0.64 (standard deviation: ±0.08) with cough and fever as predictors. The best sensitivity of 0.59 (±0.14) was from LR using cough, fever, and sore throat. RFs had the best NPV and specificity of 0.62 (±0.15) and 0.83 (±0.10) respectively with the CDC ILI symptom profile of fever and (cough or sore throat). Adding demographics and vaccination status did not improve performance of the classifiers. Results are consistent with studies using clinically-sourced data: cough and fever together were found to be the best predictors of flu-like illness[1]. Because our data include mildly infectious and asymptomatic cases, the classifier sensitivity and PPV are low compared to results from clinical data.

Conclusions

Evidence of fever and cough together are good predictors of ARI in the community, but clinical data may overestimate this due to sampling bias. Integration of participatory data can not only improve population health by actively engaging the general public[2] but also improve the scope of studies solely based on clinically-sourced surveillance data.

Table 1. Details of included participants.

Flu and Cold Season	Signups	Linked Symptom and Virus Reports
2013–2014	404	71
2014–2015	623	82

Table 2. Coding of binary features.

Feature	0	1
Sex	Male	Female
Symptoms	Absent/Unreported/ Unavailable	Present
Vaccination	Don't Know/Not Vaccinated/Missing	Vaccinated

Keywords

ARI; machine learning; participatory epidemiology

Acknowledgments

NSF grant 1343968.

References

1. Monto AS, et al. Clinical signs and symptoms predicting influenza infection. Arch Intern Med. 2000;160(21):3243-7.
2. Chunara R, et al. Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. Online Journal of Public Health Informatics. 2012;5(1).
3. Goff J, et al. Surveillance of Acute Respiratory Infections Using Community-Submitted Symptoms and Specimens for Molecular Diagnostic Testing. PLoS currents. 2014;7.
4. Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2(2):121-67.
5. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
6. Parikh R, et al. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45.

***Bisakha Ray**
E-mail: bisakha.ray@nyumc.org