Identifying Case Onset Points for Early Detection of Influenza-like Illness

Shannon Rumsey
Edward Ho
Chunting Zheng
Nealson Setiawan
shannonrumsey@ucsb.edu
Department of Statistics and Applied Probability, University of California, Santa Barbara
Santa Barbara, CA, USA

Jennifer Park
Meghan Elcheikhali
Trevor Ruiz
tdr@ucsb.edu
Department of Statistics and Applied Probability, University of California, Santa Barbara
Santa Barbara, CA, USA

Arinbjörn Kolbeinsson
Eric J. Daza
edaza@evidation.com
Evidation Health
San Mateo, CA, USA

ABSTRACT

Early detection of infectious diseases can accelerate case isolation and break chains of infection; however, inconsistent criteria for defining case onset can compromise data-driven approaches to early detection. This pilot research study combines qualitative symptom surveys, biometric data from wearable devices, and lab results collected according to standardized criteria in an effort to identify onset points for influenza-like illness. We use data collected from 5,229 study participants over a three-month interval at the outset of the COVID-19 pandemic in 2020 to develop a predictive model of the daily probability of illness given symptom reports and a basic biometric profile. To identify onset points, we conduct change-point detection on sequences of estimated probabilities. We find that onset points coincide approximately with positivity periods in 40-50% of cases. Broadly, our work underscores the potential of readily available biometric data for improved and personalized early detection technologies.

KEYWORDS

health analytics, infectious disease prediction, wearable device data, multiple time series, change point detection

ACM Reference Format:
Shannon Rumsey, Edward Ho, Chunting Zheng, Nealson Setiawan, Jennifer Park, Meghan Elcheikhali, Trevor Ruiz, Arinbjörn Kolbeinsson, and Eric J. Daza. 2023. Identifying Case Onset Points for Early Detection of Influenza-like Illness. In *Proceedings of KDD'23 Southern California Data Science Day*. ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Early detection of infectious diseases can play a large role in preventing widespread transmission and expediting interventions, thereby improving both individual and community health outcomes [2, 5]. While test results are key tools for identifying or confirming illnesses, they leave out essential context such as the time of onset and expected recovery period. Wearable devices record an array of biometric data related to an individual's health that potentially capture physiological correlates of illness onset. This data has strong potential to improve and individualize early detection and monitoring [4]. However, substantial missing data and the sparsity of studies that combine wearable device outputs with lab results and other indicators of illness present complications.

This research leverages biometric data from wearable devices together with lab results and qualitative surveys collected from a study population during the COVID-19 pandemic to develop a model that accurately predicts whether a participant is sick on a certain day given their symptoms and, in cases of illness, identifies the onset date and recovery period. We train two candidate predictive models to estimate one's daily probability of illness given survey responses and wearable device recordings, and apply change-point detection to the sequence of estimated probabilities to identify candidate onset points. We compare predictions from the candidate models, and we examine the estimated onset points relative to the first date of lab-confirmed illness ("trigger date") to provide assessments of the accuracy of onset point estimation.

2 DATASETS

The study data was collected by Evidation Health in partnership with Fitbit [3]. Study participants were pre-existing Evidation and Fitbit customers who already owned a Fitbit watch model meeting eligibility criteria. A total of 5,229 individuals participated in the study from February 2020 to May 2020. Each day, the participants were asked to answer a survey regarding whether or not they had experienced any flu-like symptoms in the last 48 hours. If the participant indicated two or more symptoms, they would be asked to take a self-administered PCR test provided to them at the beginning of the study. The PCR test tested for three viral illnesses: (i) Influenza A; (ii) Influenza B; and (iii) Respiratory Syntactical Virus. Participants were asked to wear their Fitbit device continuously except during charging for the duration of the study period.

For this research, we utilize daily symptom presence and severity for several common flu-like symptoms (fever, headache, sore throat, cough) recorded from the survey data, and a simple daily biometric profile (resting heart rate, activity level and caloric expenditure, and hours spent in bed) recorded from the Fitbit device. A participant

KDD'23 Southern California Data Science Day, August 7, 2023, Long Beach, CA

was considered to have lab-confirmed illness until self-indicating that they had fully recovered.

3 METHODS

Our overall methodology involves: (i) imputing missing biometric measurements; (ii) standardizing the biometric profile for each participant to adjust for differing individual baselines; (iii) fitting predictive models to estimate the probability of illness; and (iv) applying change-point detection to the sequences of estimated probabilities on a per-participant basis to identify candidate onset points.

During charging and at any other times the device might be taken off, the Fitbit watch records zeroes across all measurements. These missing values constitute roughly 47% of all observations recorded during the entire study period. To handle the problem of missing data, we imputed missing values using Multivariate Imputation by Chained Equations (MICE) [6]. The technique iteratively updates conditional mean imputations on a per-variable basis. We performed five iterations to obtain the imputed data.

We implemented two machine learning techniques for predicting illness: gradient boosting and a simple neural network. The boosted model was trained on all of the predictors in our (imputed) dataset, including survey responses. Observations were weighted using weights that decay exponentially away from the trigger date in order to increase sensitivity during training to data near confirmed cases of illness. The neural network, on the other hand, was trained on only three predictors: resting heart rate, total minutes spent in bed, and calories exerted. We found that performance deteriorated with additional predictors.

We applied Bayesian online change-point detection [1] to the sequences of estimated probabilities for each study participant produced by the best-predictive model to determine candidate intervals in which a participant may have been sick. These intervals were compared with the trigger dates for each participant.

4 RESULTS

The performance of the machine learning models was measured by the area under the ROC curve (AUC-ROC) computed based on predictions for a set of 364 participants whose data was held out during model training. The models were comparable but the boosted model yielded slightly better accuracy: the boosted model had an AUC-ROC score of 0.788 whereas the neural network had a score of 0.747.

Fig. 1 shows examples of alignment between change-points and positivity intervals. One or more change-points aligned with the positivity interval in about half of held-out cases: for 53.6% of patients whose data was withheld for validation, change-points appeared to coincide well with expected onset (29% of cases) or recovery (30% of cases) based on the positivity period. Depending on how coincidence is defined, this figure ranges from 40-56%. For most participants, several "false" change-points were detected: the median number of change-points detected per patient was 4.

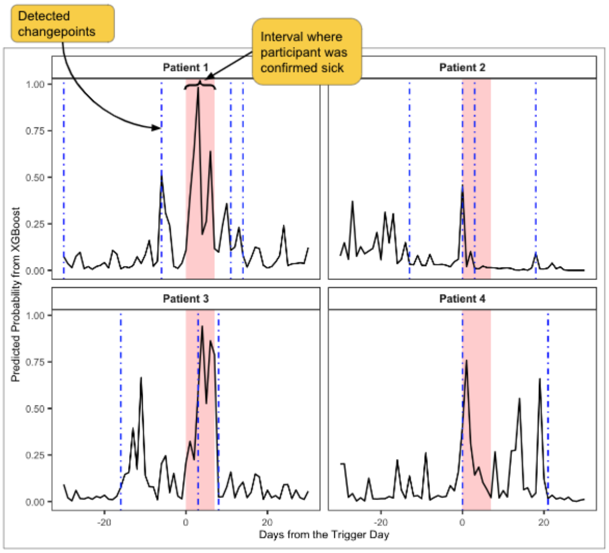

Figure 1: Predicted probability of illness 30 days before and after the trigger date, with inferred change-points, for four patients. Patients shown exemplify: (i) alignment with onset (right); (ii) misalignment with onset (left); (iii) many superfluous change-points (top); few superfluous change-points (bottom); (v) alignment with recovery (bottom left).

5 DISCUSSION

Models of illness probability based on biometric profiles and symptom reports resulted in high-sensitivity but low-specificity predictions that were promising but left room for improvement, possibly by including a more robust biometric profile. Change-points inferred from predicted illness probabilities correctly identified case onset or recovery about half of the time, but also identified superfluous time points that did not correspond to illness. Thus, as a predictive method, our approach could be improved predominantly from better control of false positives at both stages.

Interestingly, in instances where case onset was accurately identified, the lag between the change point and trigger date varied by 1.85 days on average in either direction. One possible explanation is that "physiological onset" according to bodily changes may differ from case positivity and/or symptom onset by up to several days.

It is important to note that the study data relied strongly on participant compliance and primary measurements were all self-administered. Thus, measurement error may be substantial. Even so, our results underscore the promise of biometric data from wearable devices for early and individualized detection of influenza-like illnesses and suggest a potential framework for identifying case onset.

ACKNOWLEDGMENTS

This project is partially supported with fellowships from NSF under Award No. 1924205.

REFERENCES

[1] Ryan Prescott Adams and David JC MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742* (2007).

[2] Arash Alavi, Gireesh K Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A Metwally, et al. 2022. Real-time alerting system for COVID-19 and other stress events using wearable data. *Nature medicine* 28, 1 (2022), 175–184.

[3] Arinbjörn Kolbeinsson, Piyusha Gade, Raghu Kainkaryam, Filip Jankovic, and Luca Foschini. 2021. Self-supervision of wearable sensors time-series data for influenza detection. *arXiv preprint arXiv:2112.13755* (2021).

[4] Mika A Merrill and Tim Althoff. 2023. Self-Supervised Pretraining and Transfer Learning Enable\titlebreak Flu and COVID-19 Predictions in Small Mobile Sensing Datasets. In *Conference on Health, Inference, and Learning*. PMLR, 191–206.

[5] Caitlin K Monaghan, John W Larkin, Sheetal Chaudhuri, Hao Han, Yue Jiao, Kristine M Bermudez, Eric D Weinhandl, Ines A Dahne-Steuber, Kathleen Belmonte, Luca Neri, et al. 2021. Machine learning for prediction of patients on hemodialysis with an undetected SARS-CoV-2 infection. *Kidney360* 2, 3 (2021), 456.

[6] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, Peter Solenberger, et al. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* 27, 1 (2001), 85–96.

Received 30 June 2023