Automatic Contact Tracing for Outbreak Detection Using Hospital EMR Data

Michael DeWitt*

6/14/2020

Abstract

Contact tracing is a well-known tool for public health professions to trace and isolate contacts of known infectious persons. During a pandemic contact tracing is critical to ending an outbreak, but the volume of cases makes tracing difficult without adequate staffing tools. Hospitals equiped with electronic medical records, can utilise these databases to automatically link cases into possible tranmission chains and surface potential new outbreaks. While this automatic contact tracing does not have the richness of contact tracing interviews, it does provide a way for health systems to highlight potential super-spreader events and support their local health departments. Additionally, these data provide insight into how a given infection is spreading locally. These insights can be used to inform policy at the local level.

Contents

1 Introduction

Contact tracing is an effective way to trace the origin of cases and is a well known tool for public health. However, contact tracing typically relies on small staffs of local public health offices. During the SARS-CoV-2 outbreak, these public health offices were and continue to be saturated with cases. With each positive case making approximately 5-10 contacts per day and a serial interval of 4 days (2-14 days), this means that for each positive case, somewhere between 20-40 unique contacts need to be tracked per new case, assuming that cases are detected at the first sign of symptoms. In the middle of an outbreak the workload is daunting. Effective contact tracing is a part of the CDC's guide to relaxing restrictions on social distancing.

Additionally, analysis of SARS-CoV-2 has shown that the virus typically spreads in so call super-spreader events. This means that one index may generate many secondary cases, while the majority of cases However, contact tracing is limited by several factors including staffing to

*Cone Health, michael.dewitt@conehealth.com

2 Methods

Data from the Electronic Medical Record is processed into a local enterprise data warehouse (EDW). For SARS-CoV-2 the critical data collected were positive test results. These positive tests were recorded both for on-site rapid tests and external laboratory results. These positive test results were then paired with general patient information collected about the patients within the EMR as part of the normal intake process. These data include the patient's reported address, their primary spoken language, and employer. Employment history was also supplemented with insurance information (e.g. plan numbers) if the patient had commercial insurance.

The addresses on file for those patients were then geocoded using an on-premises secure software. Prior to geocoding the addresses, a program was run to clean the addresses in order to make the probability of a matching address being found higher (e.g. consistent abbreviations for streets, removal of apartment numbers). These geocoded addresses were then truncated to four decimal places in order to represent a 30 foot radius around the reported address. A similar approach was used for employment information with self-reported employment information being cleaned for consistency as well as compared against a list of locally developed employer alias (e.g. Proctor & Gamble = Proctor And Gamble, etc) in order to account for inconsistencies in naming.

Using understanding of local demography, a select group of languages were generated. It is known that certain immigrant communities settled in well defined areas of the service region. Because of the close-knit nature of these communities and the relative small number of speakers of these languages, these languages were also used to examine for potential outbreaks.

In order to establish possible linked cases, the positive patients were grouped on each of the three main criteria, location, employer, and language. For those persons that shared a unique location, a "cluster" was formed. The patient that recorded the first positive case was declared the index case for the cluster, and the difference from the index case was calculated for each cluster. This process was repeated for employers and languages. If the time between cases was greater than a threshold (e.g. 30 days), then the index case was recalculated iteratively for a given cluster (i.e. one household might have more than one index case).

The derived cluster information was combined from the three different sources and paired with other demographic information like age, race, and gender. If an individual appeared in more than one contact network, the earliest known transmission chain was retained. This data was then combined into a line list, or list of all positive cases, and a contact list that contained the above information regarding who likely was the index case for each infection. All data cleaning and aggregation took place in the R Statistical Environment (R Core Team 2019). The Epicontacts package (Nagraj et al. 2017) was utilised in order to visualise the tranmission chains under a variety of different views.

3 Results

4 Discussion

References

Nagraj, VP, Thibaut Jombart, Nistara Randhawa, Bertrand Sudre, Finlay Campbell, and Thomas Crellen. 2017. *Epicontacts: Handling, Visualisation and Analysis of Epidemiological Contacts.* https://CRAN.R-project.org/package=epicontacts.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.