An algorithm for identifying potential COVID in medical records data

Data

We are working with both claims and medical records from a COVID hot spot location. The records data covers October 1, 2017 to March 20th, 2020.

Code Sets

The first challenge to identifying vulnerability to COVID-19 is linking medical histories with infection. This early in the outbreak there does not appear to be significant usage of the CoV-SARS-2 specific codes. The ICD-10-CM code designated for the disease – U07.01 – is not intended for use until April 1, 2020, and we found no evidence of the use of procedure codes specific to specimen collection for diagnosis of COVID-19 (U0001, U0002, 87635, 99091).

We used lists of published codes [1,2] along with anecdotal evidence suggesting potential connection viral conjunctivitis and diarrhea [3] to identify potential cases of COVID based on disease symptoms. We hypothesize that, in the absence of specific direction or codes specific to the disease, providers are simply (and appropriately) describing the manifestations of the disease in each patient.

Identifying potential cases

The algorithm we used to identify potential cases of COVID-19 consists of

1. Merging visits that are separated by 1 or fewer days
2. Identifying visits that contain at least two unique codes from the list in appendix A
3. Removing visits that contain codes for influenza (J09, J10, J11)

Unsurprisingly, this algorithm identifies numerous cases throughout the time we observe this patient population. In order to assess the extent to which we are identifying COVID specific cases – as opposed to other causes – we modeled the event rate of these visits per day as a function of (i) the total number of visits in the record (ii) the day of the week (iii) the number of visits with a diagnosis of Influenza. The details of this model are in appendix B.

The relationship between the expected number of visits that qualify as potential COVID and our model's estimated number of such visits is show in Figure 1. In order to smooth natural within-week fluctuations, the y-axis shows the ratio of observed events to the 97.25% quantile of expected events. An unscaled version of this graph is available in Appendix B. The graph shows a drastic uptick in COVID-similar cases in the last 6 days of observation. These correspond to March 12, and March 15-19 (the weekend days are not shown).

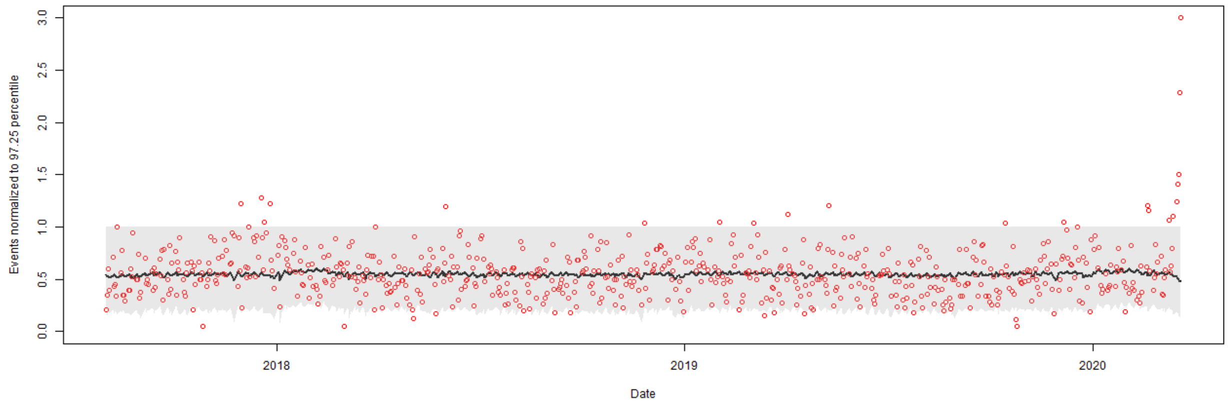

Figure 1 – Cases per day presenting with symptoms similar to COVID. The large spike at the right end corresponds with outbreak of the disease at this location.

Characteristics of patients

The number of observed COVID-similar cases is double the expected number beginning on March 13 and remains so throughout the observation window. We are interested in assessing the demographic characteristics of the patients presenting with COVID-similar complaints on or after that date compared to (i) the general population and (ii) the population of patients presenting with these symptoms before the outbreak began (prior to January 1, 2020). There are 178 such cases in that time window (175 unique patients), and we estimate – based on the difference in expected and actual case load – that approximately 70% of them correspond to COVID itself. We have 9082 such cases in the record from before January 1, 2020 corresponding to 6898 patients; we assume that none of these are COVID.

References

1. American Academy of Pediatrics (2020) "How to use ICD-10-CM, new lab testing codes for COVID-19", *AAP News,* March 12, 2020, https://www.aappublications.org/news/2020/03/12/coding031220
2. American College of Cardiology (2020) "Coding Updates for COVID-19", *ACC News,* March 11, 2020, https://www.acc.org/latest-in-cardiology/articles/2020/03/11/14/19/coding-updates-for-coronavirus-disease-2019-covid-19
3. "Alert: Important coronavirus updates for opthalmologists", *The Opthalmic News and Education Network,* March 25, 2020, https://www.aao.org/headline/alert-important-coronavirus-context
4.

Appendix A

ICD-10 codes used to identify COVID-19

- J12.89 – Other viral pneumonia
- J20.8 – Acute bronchitis due to other specified organisms

- J22 – Unspecified acute lower respiratory infection
- J40 – Bronchitis, not specified as acute or chronic
- J80 – Acute respiratory distress symdrome
- J98.8 – Other specified respiratory disorders
- R05 – Cough
- R06.02 – Shortness of breath
- R50.9 – Fever unspecified
- R09.2 – Respiratory arrest
- B97.29 – Other coronavirus as the cause of diseases classified elsewhere
- H10.3 – Unspecified acute conjunctivitis
- R19.7 – Diarrhea, unspecified

Appendix B

We modeled the number of visits qualifying as potential COVID but occurring before the outbreak (prior to 1/1/2020). We used negative binomial regression in order to account for overdispersion. There is a significant effect on both overall volume and the number of COVID qualifying cases on weekends, so Saturdays and Sundays were removed from the sample. Model parameters are listed in the table below:

Table 1 – Parameters of model predicting the expected number of COVID-similar cases per day.

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	1.348E+00	1.799E-01	7.494E+00	6.668E-14
N	1.545E-04	1.958E-05	7.891E+00	3.010E-15
J09	-4.511E-02	5.867E-02	-7.688E-01	4.420E-01
J10	6.697E-03	3.395E-03	1.972E+00	4.857E-02
J11	7.068E-03	2.197E-03	3.217E+00	1.296E-03
day1	4.773E-02	4.878E-02	9.786E-01	3.278E-01
day2	-1.093E-01	5.211E-02	-2.098E+00	3.593E-02
day3	-5.336E-02	5.310E-02	-1.005E+00	3.150E-01
day4	-3.100E-02	6.867E-02	-4.514E-01	6.517E-01

The estimated size parameter (accounting for overdispersion was 14.86). After fitting the model, we found that 1.96%, and 1.31% of observations in the training data fall above or below respective the 95% confidence bound. This suggests that the overdispersion parameter is appropriately accounting for uncertainty in the model. The figure below shows model estimated event counts by day throughout the range of dates for which we have data.

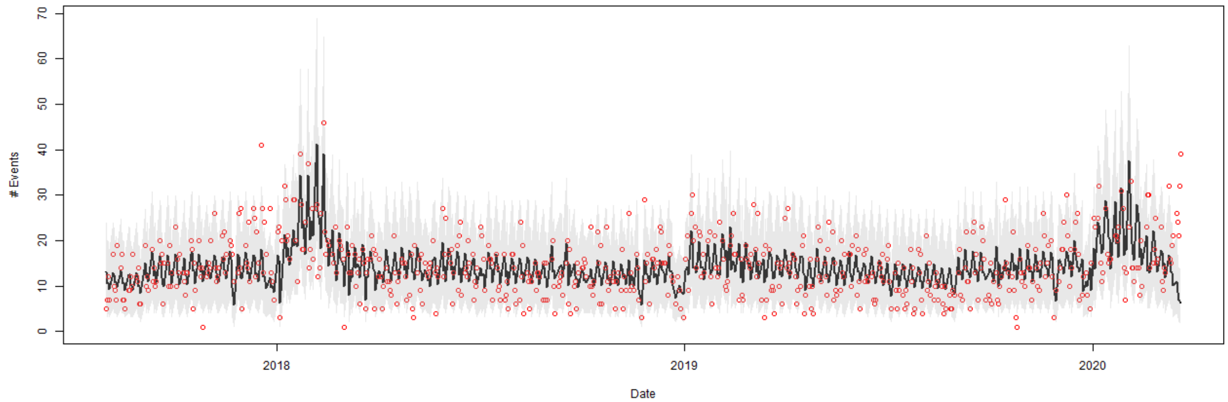

Figure 2 – Number of COVID-similar per day throughout our observation window.