Early warning signals derived from COVID-19 testing data

Christopher M. Hoover, Joshua Schwab, Maya Peterson, Mark van der Laan, others

Abstract

COVID19 testing data is essential to monitor the status of ongoing epidemics across the U.S. and the world. Because of the high rate of asymptomatic and mildly symptomatic cases that contribute to transmission, widespread testing and subsequent contact tracing and isolation are necessary to reduce epidemic spread. The effective reproductive rate, $\mathcal{R}_e$, is a simple measure of transmission interpreted as the expected number of new infections arising from a single infection at a given time, that can be used to monitor the status of an outbreak. Here we discuss estimation of $\mathcal{R}_e$ from currently available testing data and proprose methods to correct for inherent biases that currently limit the utility of testing as a reliable indicator of the status of the COVID19 pandemic.

TODO:

- Update DAG based on Maya and Mark email chain, incorporate time ordering
  - How does $\mathbb{E}(Y(t))/\mathbb{E}(Y(t-7))$ (or something similar) relate to $\mathcal{R}_e$?
- Decide on data generating mechanism (ABM?)
- **With line list data** adjust testing for demographics, zip code, SES, age
- Estimate $\mathcal{R}_e$ from testing data adjusted for underreporting using Hospitalizations/Deaths data (LSHTM method)

Background

COVID19

Areas across the United States and the world are beginning to reopen following the unprecedented pandemic caused by the emergence of SARS-CoV2. Widespread availability of testing for active and past infections remains a critical component of the ongoing response to the COVID19 pandemic…

Testing

- Background on types of tests
- What tests are actually doing (e.g. distingiush PCR = active infections, antibody = past infections)
  - Sensitivity issues

Table 1: Methods to estimate E(Y)

Method	Pros	Cons	Reference
Number of positive tests	real-time, simple, no analysis required	Heavily biased by number of tests conducted, test-seeking behaviors, etc	
Percent of tests conducted that are positive	Real-time, simple, removes bias from number of tests conducted	Heavily biased by test-seeking behaviors, etc.	
Adjustment from additional epidemiological data	Removes bias associated with test-seeking	Delay between infection and hospitalization/death means unable to estimate in real-time	LSHTM
Adjustment from additional demographic/survey? data	Simple and straightforward	Does not adequately control for confounding by W2, requires linelist testing data	
Fitting dynamic models	Real-time, draws on multiple data sources to control for biases	Identifiable?, complex and computationally intensive	

Testing can be earliest indicator of rising infections if conducted/analyzed appropriately

Hospitalizations and deaths are more what we're worried about, but these are preceded in time by increase in cases. Since tests can detect infection before these more severe outcomes, they can be an earlier indicator of increases in transmission. A useful summary statistic to estimate the current status of an outbreak is $\mathcal{R}_e$, the expected number of additional cases caused by a single case at any time during the outbreak [1].

Testing cannot currently be used to reliably estimate $\mathcal{R}_e$ because of bias due to variability in test-seeking behaviors, test availability (geographically, socioeconomically, etc.)

Testing can be used in either passive or active surveillance to monitor the status of ongoing outbreaks. Especially early in the outbreak when testing was not widely available, the set of individuals allowed (wc?) to be tested was restricted. To date, most COVID19 testing data has been passively collected: those with suspected COVID19 exposure or symptoms seek out testing from a healthcare provider, rather than actively being pursued to be tested in a more systematic fashion, e.g. as part of a representative cohort of the population of interest. This introduces considerable bias into epidemiological estimates derived from the testing data, such as $\mathcal{R}_e$. [Review CITEs on passive surveillance and bias] The number of tests available, test-seeking behaviors of different populations/individuals, and [something else] all influence the number of positive tests and the positive test percent in a given location and time (Figure 1, W2; **Maybe could add $\mathcal{R}_e$ to the DAG as a function of Y?**).

Crude metrics from testing data such as the number of positive tests reported or the percent of tests conducted that are positive have been/are being used to monitor the status of outbreaks across the US/world. However, such metrics are heavily biased due to reasons outlines above, and may be flawed representations of true underlying infection dynamics such as $\mathcal{R}_e$ [**Think this sentence/point is really important and could use some work**]. Here we propose/review methods for correcting for these underlying biases in testing data, discuss their pros and cons, and evaluate their ability to a) estimate $\mathcal{R}_e$ in real time and b) predict a future increase in hospitalizations that will overwhelm capacity. We conclude by proposing methods to generate unbiased testing data prospectively using adaptive design (active surveillance, embedding prospective cohorts within passive testing, all/none of the above?) and [hopefully] show that these methods outperform those currently available that rely on inherent flaws in currently available testing data.

Correcting for bias/Accurate estimation of $E(Y)$/Adjusting for $W2$

Approaches

Data-generating process

Needs to:

- Accurately represent influence of W1 and W2 on transmission
- Accurately represent influence of W1 and W2 on testing (both simulated and observed)
- Allow easy manipulation/implementation of changes in $\mathcal{R}_e$
- Track infection and testing status across different (SES, race, demographic, occupational?) groups

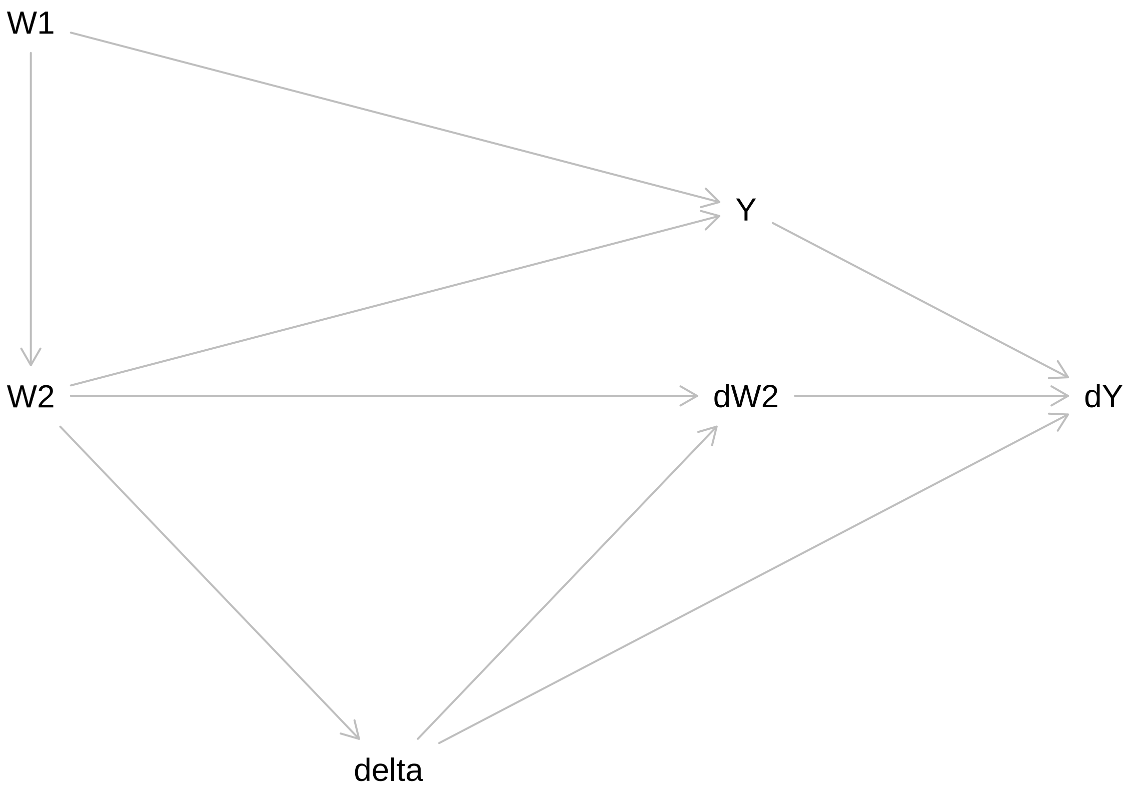


Figure 1: TODO: Make this longitudinal? Data generating process for COVID19 testing data where W1 is demographic covariates (age, race, SES), W2 is risk covariates (symptoms, contacts), Y is SARS-CoV2 infection status, delta is testing status. The observed data O consists of $(W_1, delta, dW2, dY)$, but conditioing on W_2 is necessary to identify $E(Y)$.

Candidates:

- Agent-based model
- Branching-process model with network influence
- Something simpler may be adequate for the point that testing isn't great and could be improved

Estimation from raw testing data

Either number of positive tests or pct positive tests

Adjustment from additional data sources (demographic [Age, Race, Zip, SES] and/or epidemiologic [hospitalizations and deaths])

LSHTM method (Case-fatality ratio adjustment?)

- Time-delay limits utility
- See this LSHTM report for one method that seems to be widely accepted
- I worked briefly on trying to derive a max likelihood estimate of COVID19 cases on day t from hospitalizations and deaths, but got stuck on right censoring. Could revive that effort though

Adjustment built into model fitting

Big question: How inherently different is this really from methods above? Just a combination of them?

Approach Chris has been working on

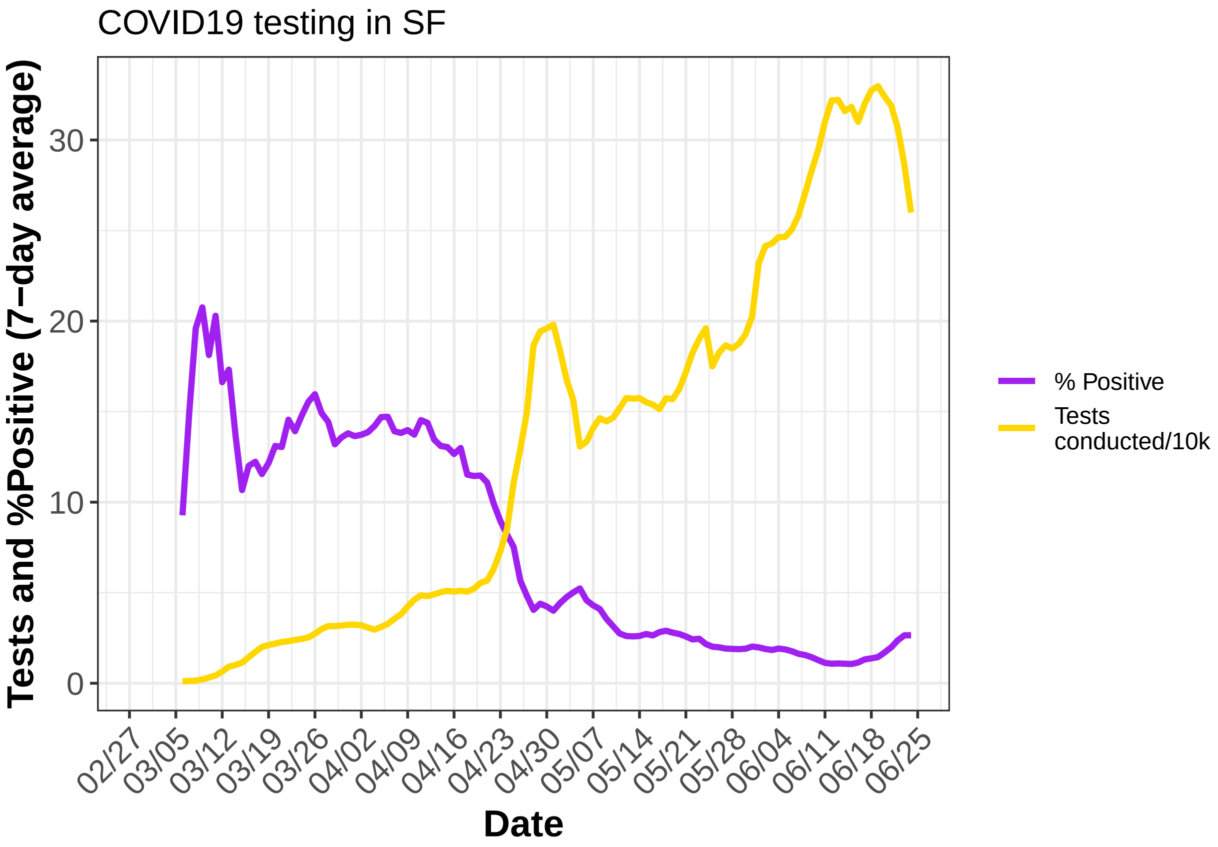
- Still relies on additional data sources for fitting
- Projections require assumptions of number of tests, bias in testing in the future

Improving testing (Is this another paper even??)

- What additional data can be reported to improve epi estimates (Symptom status, age, etc.)
- Can additional data be collected in order to improve the utility of available testing resources
  - Intuition is that smart reallocation of testing to do some active surveillance will improve estimation even more than just doing more tests, which I think could/should be one of the main messages
  - Could maybe start with data from the Mission Study, use it to derive age/race/income? adjusted rates for all of SF, compare to testing data reported for all of SF county at the same time. Could provide an initial estimate of testing bias. Would be even better if we could do record linkage between Mission Study and other testing data

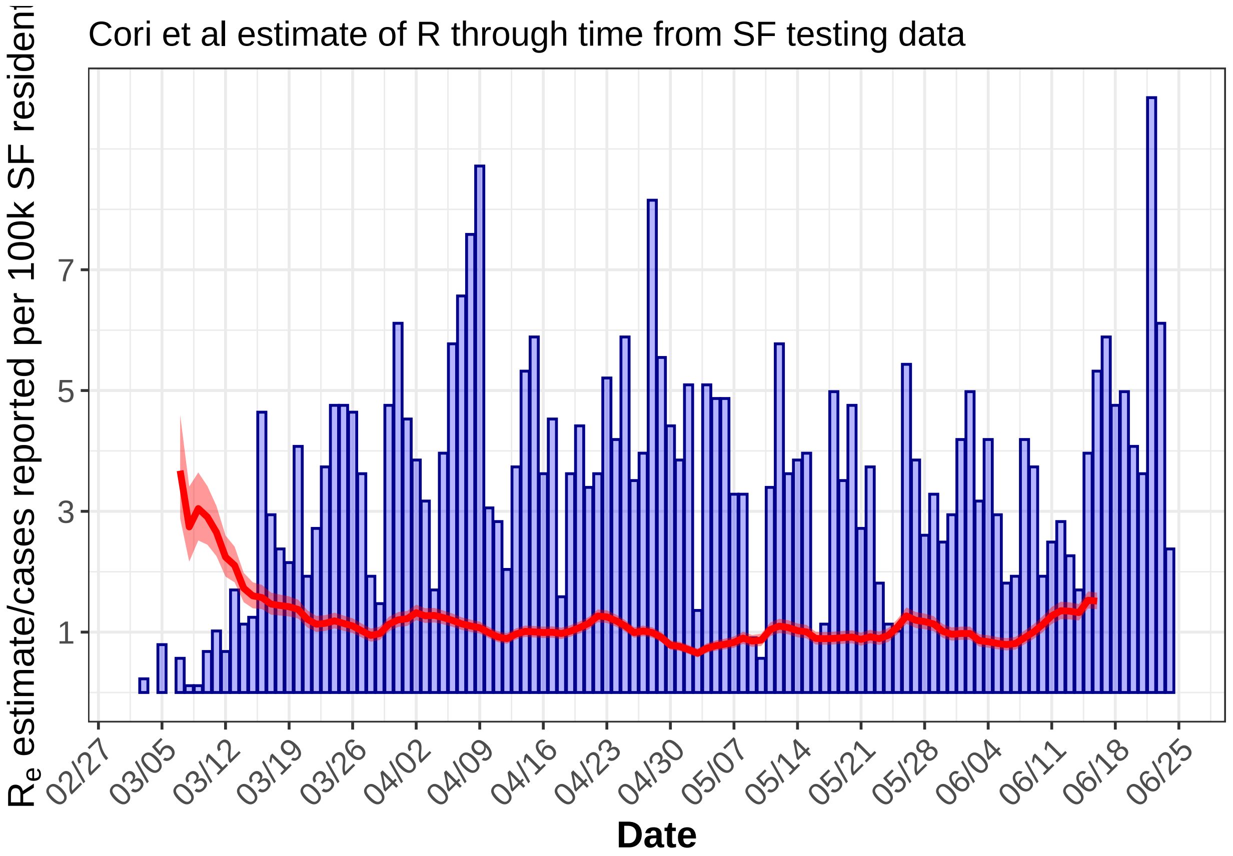
Potential Figures

SF Testing

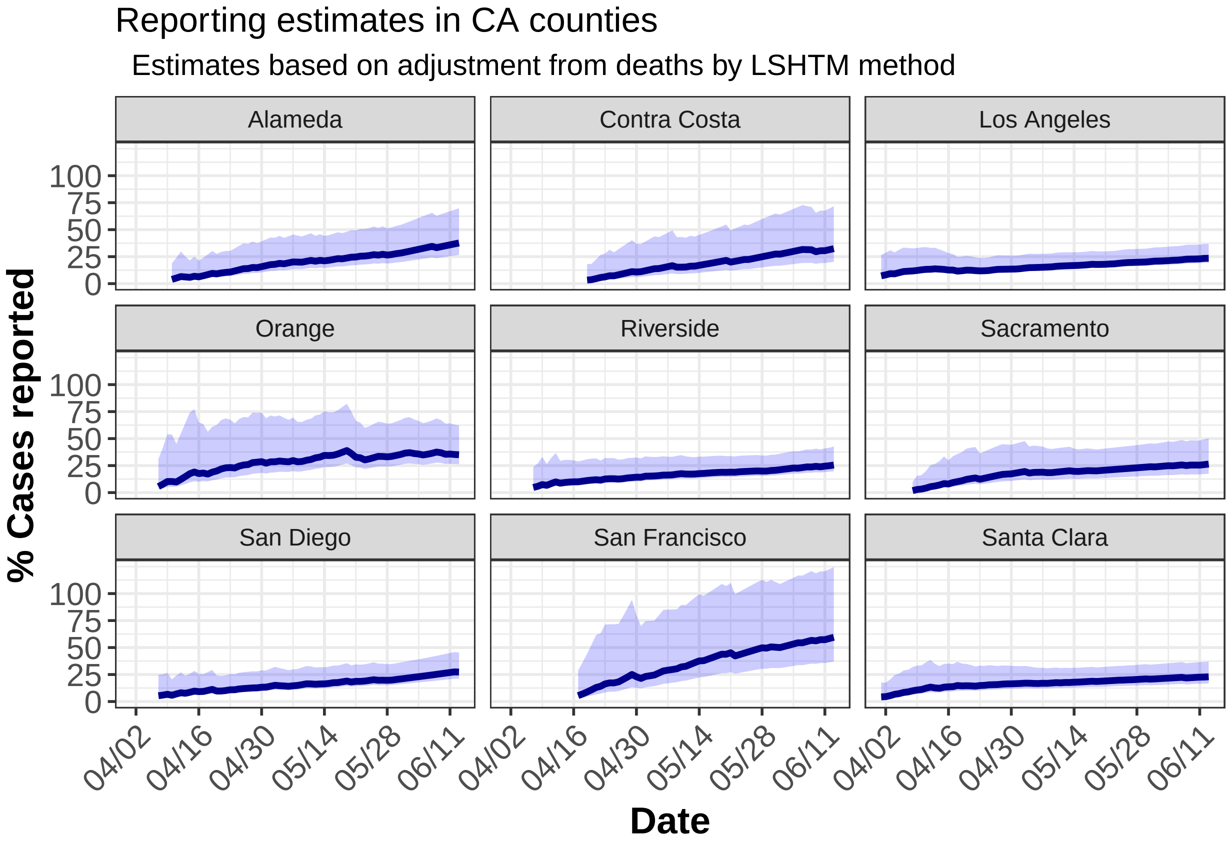


$\mathcal{R}_e$ estimation

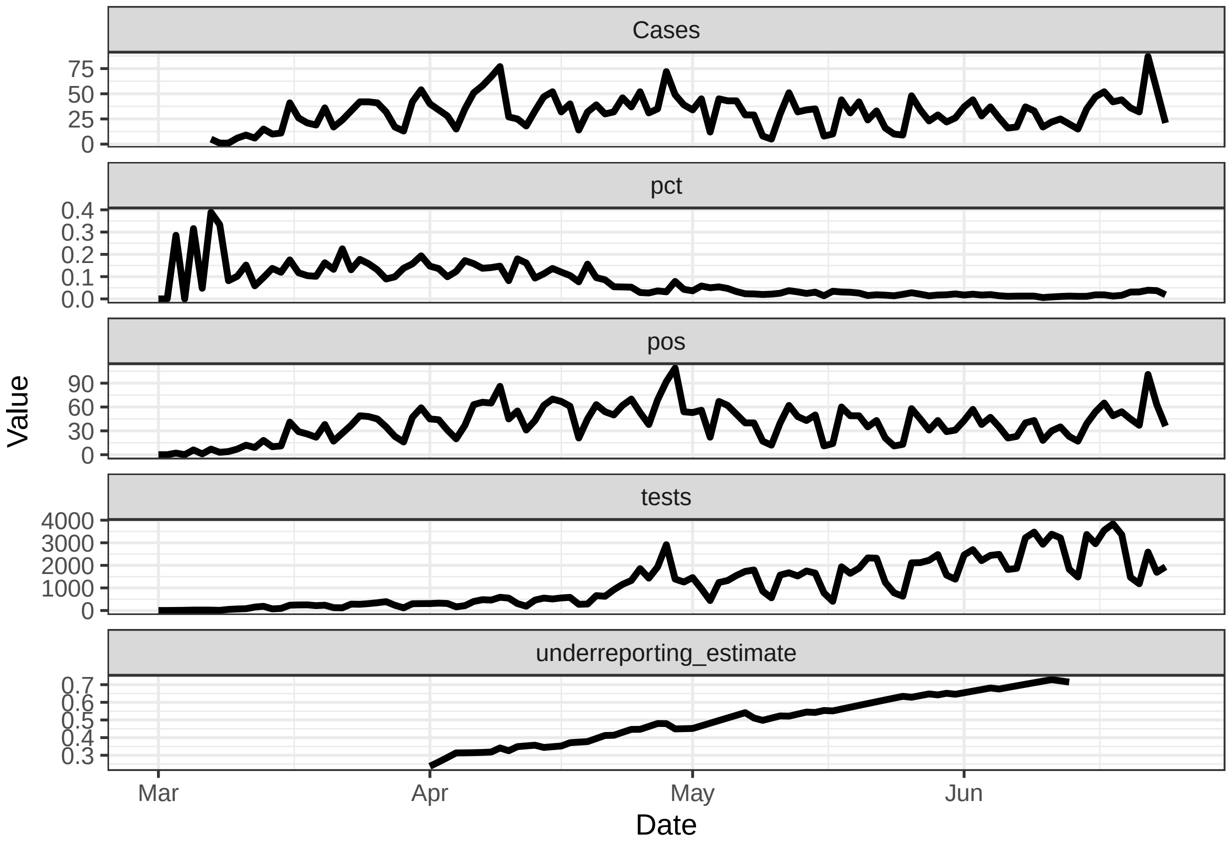
Cori et al method [1,2]

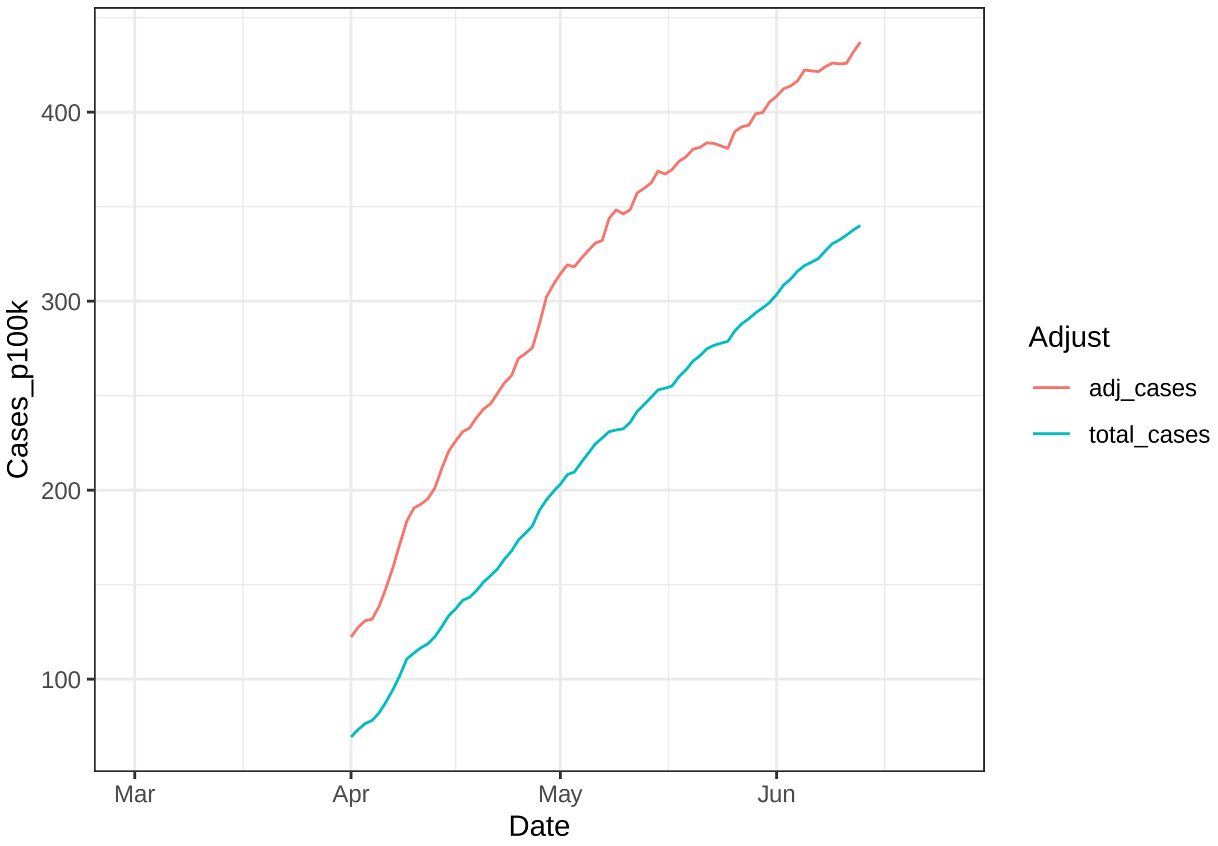

Cori et al estimate of R through time from SF testing data
R_e estimate/cases reported per 100k SF resident
1
3
5
7
02/27
03/05
03/12
03/19
03/26
04/02
04/09
04/16
04/23
04/30
05/07
05/14
05/21
05/28
06/04
06/11
06/18
06/25
Date

Underreporting estimates via LSHTM method



These estimates come from data from the statewide database, but some wonkiness in there (e.g. negative case counts over the weekends sometime), so will reestimate for SF using the county data.

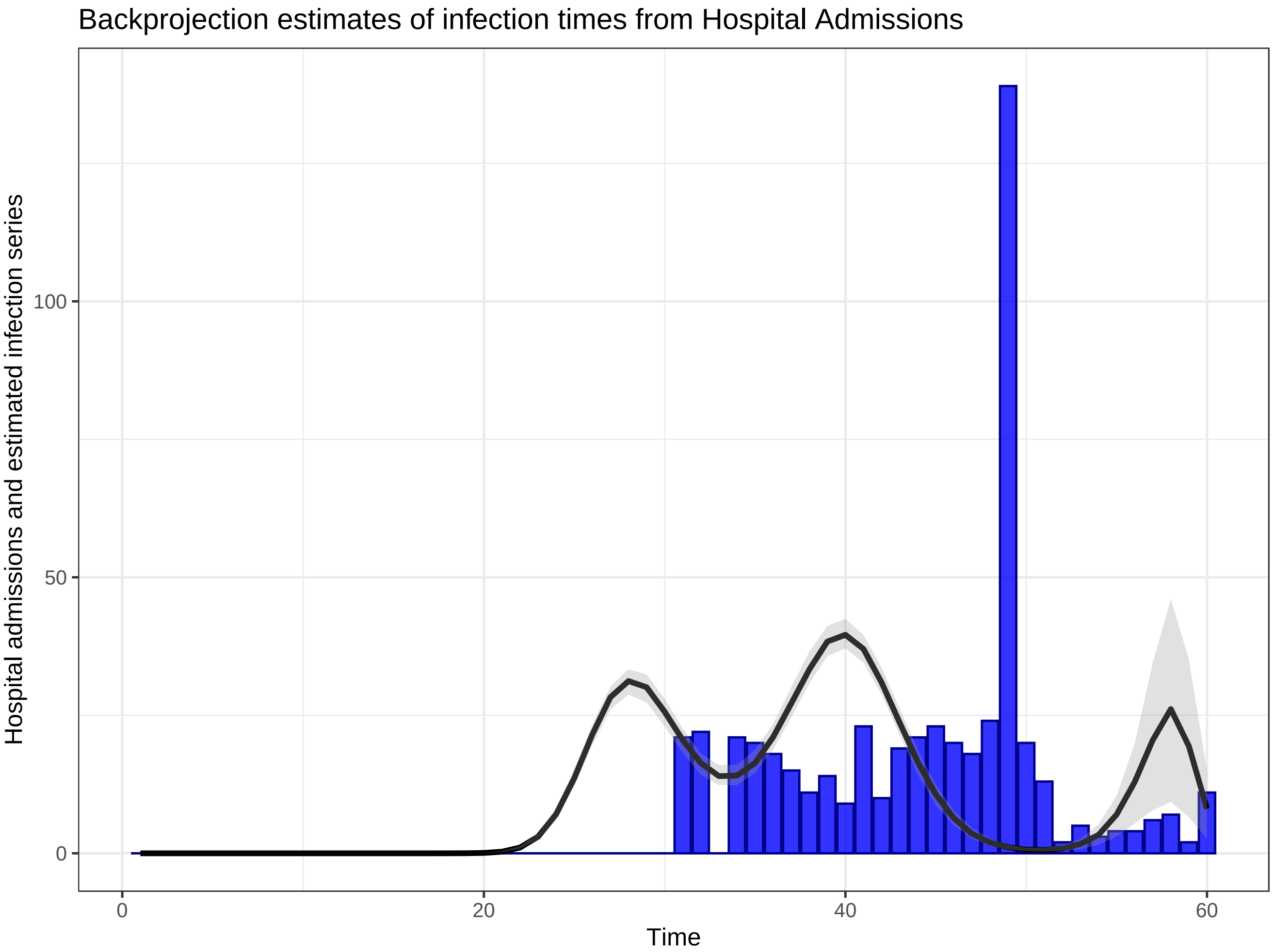

Cases
75
50
25
0
pct
0.4
0.3
0.2
0.1
0.0
pos
90
60
30
0
tests
4000
3000
2000
1000
0
underreporting_estimate
0.7
0.6
0.5
0.4
0.3
Value
Mar
Apr
May
Jun
Date


400
300
200
100
Cases_p100k
Mar
Apr
May
Jun
Date
Adjust
adj_cases
total_cases

Back projection from hospitalizations

TODO: Generate case estimates based on proportion of cases that beome hospitalized

TODO: Methods for age, race, ethnicity adjustment if given line list data



References

1. Gostic KM, McGough L, Baskerville E, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. medRxiv. 2020; 2020.06.18.20134858. doi:10.1101/2020.06.18.20134858

2. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology. 2013;178: 1505–1512. doi:10.1093/aje/kwt133