A design for hospital-based coronavirus tracking

Len Covello, Andrew Gelman, Yajuan Si, and Siquan Wang

09/19/2020

In the presence of an epidemic, it is important for hospitals and health maintenance organizations to have indications of changes in rates of exposure among patients and the general population. We present here an approach to collecting and analyzing data on viral exposure among patients in a hospital system and performing statistical adjustment to estimate viral incidence and trends in the community. We are currently applying this monitoring and analysis protocol in a small hospital system in Indiana, and this could serve as a model for other medical groups to perform local data collection and analysis, ultimately with an eye to linking to national data collection efforts.

The Problem

Knowledge of incidence and trends of viral behavior in communities is important and, as yet, poorly characterized. As we transitioned in our area of interest (Munster, Indiana, a suburb of Chicago) from a high clinical burden of disease in an economically closed environment into a decreasing burden with increased social and business commerce, it was apparent that we lacked a means of deciding how our actions were influencing virus behavior. At the onset of the disease in the U.S., we had the luxury of internalizing the experiences of Italy, Spain, and eastern Asia in modeling what the illness metrics might be without needing to be too clever. As there was a strong political motivation here to increase human interaction well ahead of virus mitigation experienced elsewhere, we were to be a vanguard of reopening and needed to develop our own predictors.

At the time of conceiving this project, the country lacked uniform testing quality and capabilities. There was no acceptable proxy for random sampling of any community. Limited availability of testing meant only highly symptomatic individuals would be tested, skewing any interpretation of prevalence. Suspicions of asymptomatic incidence and spread of disease were developing and remain poorly characterized yet.

Our community needed a way of reliably assessing viral incidence and trends so as to prepare for hospital burdens effectively. There was a need for a random sample of the community or proxy thereof.

Proxy Sample

We were fortunate in that the anesthesia professionals in our hospital system were sufficiently concerned about asymptomatic viral shedding to want to preoperatively test all patients. All elective surgical patients are presumptive asymptomatic, as anyone exhibiting symptoms would have surgery cancelled or deferred. This population presented a potentially valuable resource. There is a broad age, racial/ethnic, and economic diversity to this group, and its only overt correlation to disease status is that it is selected for a lack of symptoms. Such a group should behave more like the population as a whole with respect to the disease than the symptomatic population. Naturally, we would have preferred a sample chosen without reference to symptom status, but we feel that the current population, if not ideal, is the best that can be done under routine clinical circumstances. Further, if we were to extend testing beyond acute sampling for viral RNA to also include assay for IgG antibody, we would have an even better random proxy. That is, this group is not

random with respect to acute disease, but should be quite nearly random with respect to a history of disease one month or more ago, for which the population IgG status wishes to identify. We were able to institute routine IgG testing for any preoperative patient who was otherwise requiring preoperative blood tests.

**Admittedly, this group trends somewhat older and sicker than the group at large (undergoing the RNA test), but still seems a fairly representative sample for our community at large.

Measurement

We anticipated some statistical challenges. Sensitivities and specificities vary greatly among available testing regimens. We are using Abbott tests for which these parameters are reasonably stable and well analyzed. The PCR RNA test shares sensitivities similar to all the better tests out there. They all report a 70% sensitivity, but that is standardized by detection of RNA in strongly symptomatic patients. There is a fair amount of data to suggest that asymptomatic and presymptomatic patients are both harder to detect than the highly symptomatic ones and that the date of infection and symptom onset have a large effect on the sensitivity. These effects would need to be acknowledged and, to the degree possible, accounted for in the model. Specificity is likely to be theoretically 100%; the only false positives should be cross contaminated or switched samples. However, even very small false positive rates can result in large uncertainties about exposure rates if underlying prevalence is low enough. Matching the RNA sequence of interest at the detection threshold should be virtually impossible mathematically with the viruses involved. The IgG test also has high sensitivity and specificity, but we would need these statistics respected in the model, unlike in some other reported data sets.

Statistical analysis

In addition to adjusting for measurement error, we need to normalize for variation of this population demographics compared to the actual community. We expected that we would be helped by having a fairly representative group to begin with and hoped that poststratification to the target population would help enhance the accuracy of our conclusions.

We are interested in rates of coronavirus exposure, both among hospital patients and in the larger community. We recognize that patients, even asymptomatic patients, are not representative of the general population; nonetheless adjusting for demographics and geography is a start. Multilevel regression and poststratification (MRP) is a standard method of adjustment used in survey research that is particularly effective when sample sizes are small in some demographic or geographic slices of the data (Andrew Gelman and Little 1997; Si et al. 2020). We use the Bayesian approach of A Gelman and Carpenter (2020) to apply MRP to testing data with unknown sensitivity and specificity, here using the following adjustment variables: sex, age (0-17, 18-34, 35-64, and 65+), race (white, black and others), and county (Lake and Potter).

We poststratify to two different populations: patients in the hospital and residents of Lake/Potter County, Indiana. For the hospital, we use the following database to represent the population of inpatients from three hospitals in the Community Health Systems (Community Hospital, St. Catherine Hospital and St. Mary Medical Center). For the community, we use the Americian Community Survey data from the three counties of interest.

In addition, we are interested in changes over time. Indeed, even if our demographic and geographic adjustment is suspect (given systematic differences between sample and populations), we still can learn from time trends, and here the adjustment is particularly important, given the way the mix of patients has changed during the past few months.

We use the following model of changes in the multilevel regression parameters over time.

$$\pi_i = \text{logit}^{-1}(\beta_1 + \beta_2 * male_i + \alpha^{\text{age}}_{age[i]} + \alpha^{\text{race}}_{race[i]} + \alpha^{\text{time}}_{time[i]} + \alpha^{\text{county}}_{county[i]} + \alpha^{\text{age * sex}}_{age*sex[i]}), \quad (1)$$

where π_i is the viral incidence of individual i, *male* with value 1 indicates men and 0 for women, $age[i], race[i]$, and $county[i]$ represent age, race and county categories, with a two-way interaction term $age * sex[i]$, $time[i]$ indices the time in weeks when the test result is observed for individual i; and the α parameters are vectors of varying intercepts. We assign hierarchical priors to the varying intercepts

$$\alpha^{name} \sim \text{normal}(0, \sigma^{name}),\ \sigma^{name} \sim \text{normal}_+(0, 2.5), \text{ for } name \in age, race, county, time. \quad (2)$$

We perform all computations in Stan and R [Python], and all data and code are available at [Github site].

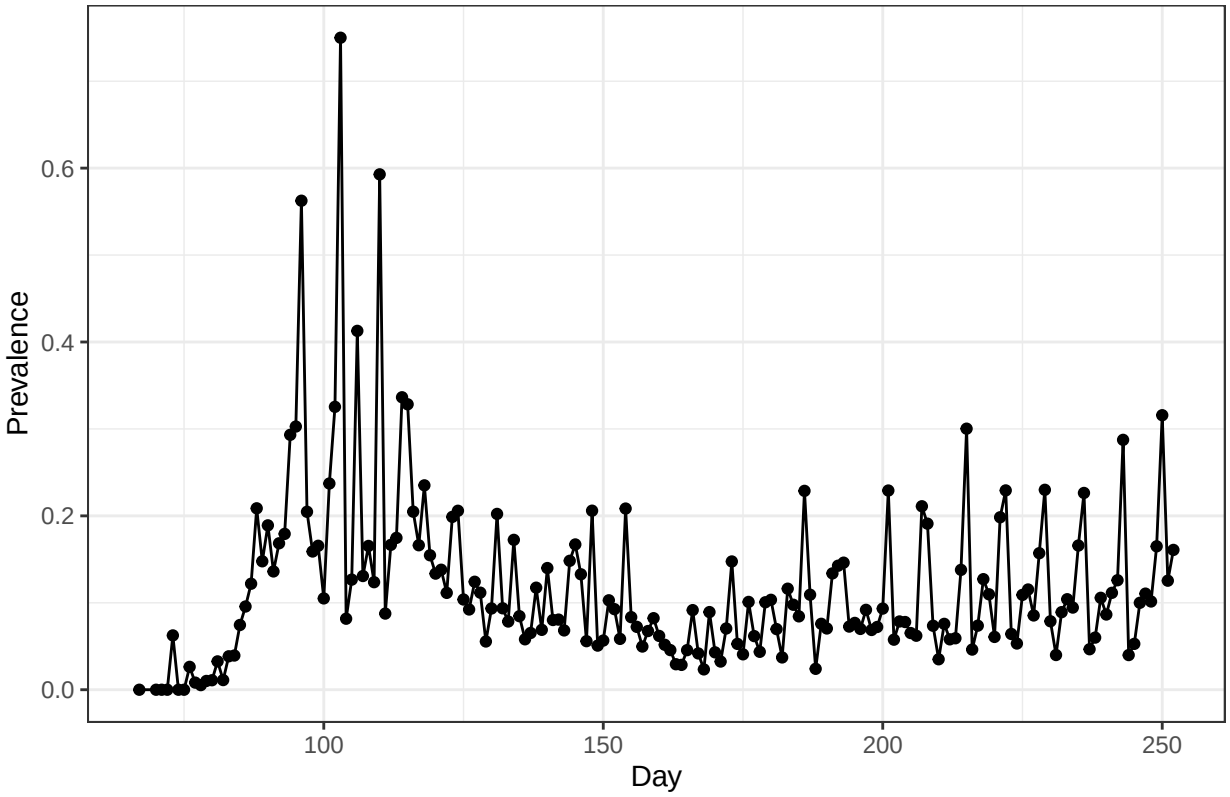
Pre-study conjectures

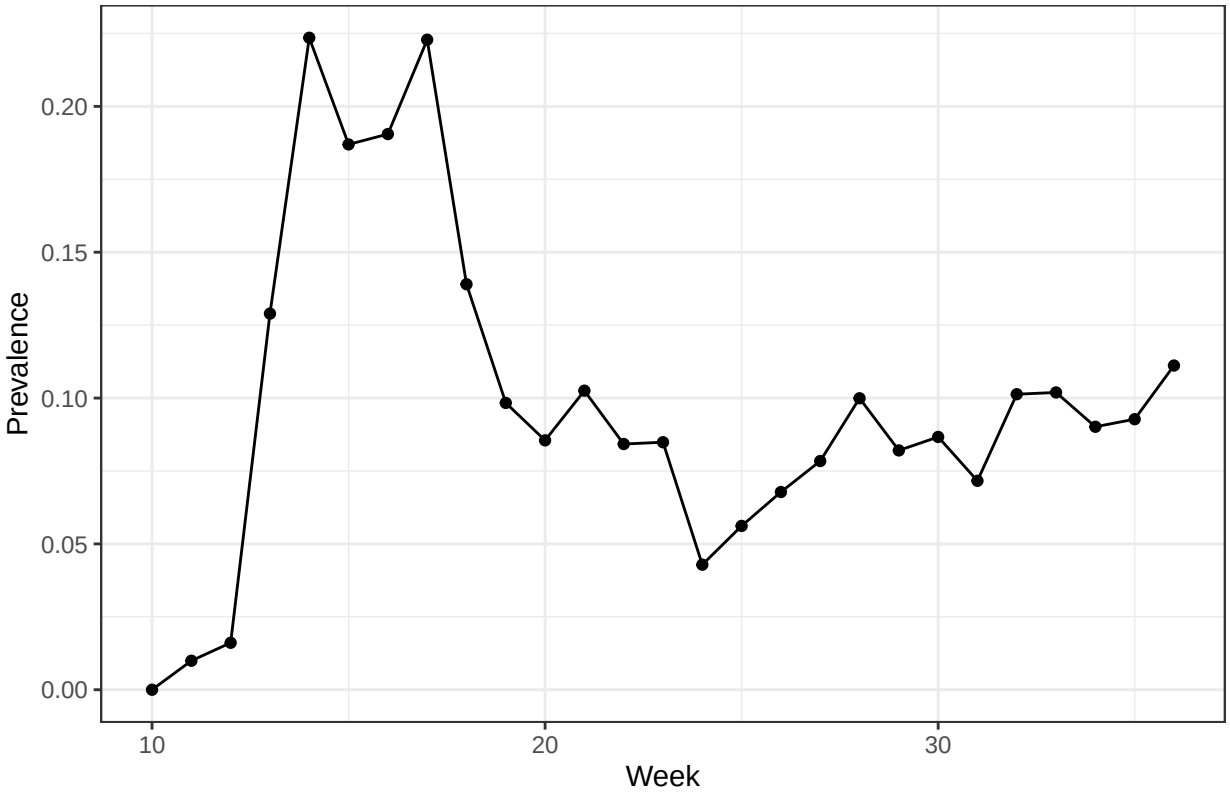
We began the data collection with a few hypotheses or speculations. First, that the ratio between asymptomatic and symptomatic patients would be relatively constant. Second, that changes in asymptomatic measurable RNA would precede parallel changes in symptomatic measurable RNA by a couple of days, given the known temporal relationship of viral shedding to the onset of clinical disease. Third, that trends in our asymptomatic RNA positives would therefore parallel and predict the behavior of the virus within the community as a whole.

In summary, we anticipated that appropriate modeling of the RNA dataset would allow us to measure changes in acute viral infectious incidence early in the trend so as to stay ahead of the disease, or at least in concert with any changes. Further, we hoped that accumulation of IgG data and its intelligent analysis would help us to understand increasing immunity in the community from viral exposure and perhaps, evidence for a loss of immunity with time, should the IgG positivity flatten or decrease.

Analytic Results

Comparasion with State Website Data (focus on District 1)





Data Import and Manipulation

We should be careful regarding missing data. Here we check whether there is any missing data in age, gender, race and PCR test result. Also the data quality might be a concern, for example, in symptotic patients data, patient with id 31545242 had Speciman Date & Time on 6/9/2020, but IGG SPECIMEN_TAKEN_TIME on 4/30/2020, so there might be some incorrespondence but I did not adjust for it here. Also there might be some duplicated samples, for example, in symptotic patients data, patient with id 31545242 appeared twice.

Exploratory Data Analysis

Table 1: Summary of test results and sociodemographics

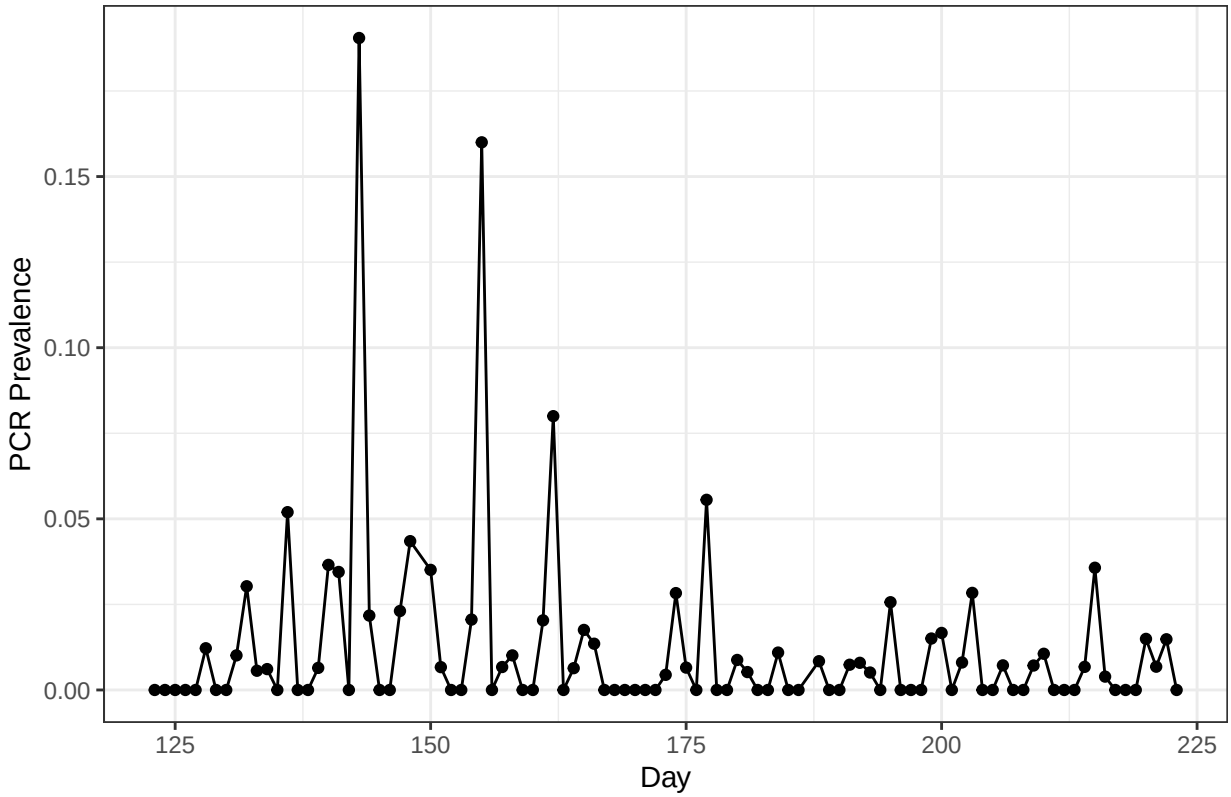
	Asymptomatic PCR	Symptomatic PCR	Asymptomatic IgG	Symptomatic IgG	Hospital	Community
Size	11192	2356	1314	160	35838	654890
Incidence(%)	1	17.1	6	41.9	NA	NA
Female(%)	58.3	57.6	58.8	65	57.3	51.3
Male(%)	41.7	42.4	41.2	35	42.7	48.7
Age0-17(%)	3	9.4	0.9	1.9	8.7	23.5
Age18-34(%)	9.5	19.7	8.4	11.9	12.1	21.2
Age35-64(%)	44.3	44.1	48.7	52.5	30	39.6
Age65+(%)	43.3	26.9	41.9	33.8	49.1	15.6
White(%)	73	61.6	71.8	61.3	65.4	69.2
Black(%)	13.7	18.5	15.4	18.1	18.7	18.8
Other(%)	13.3	19.8	12.7	20.6	15.9	12.1
Lake(%)	84.3	85.8	90.9	85	88.4	74.3
Potter(%)	15.7	14.2	9.1	15	11.6	25.7

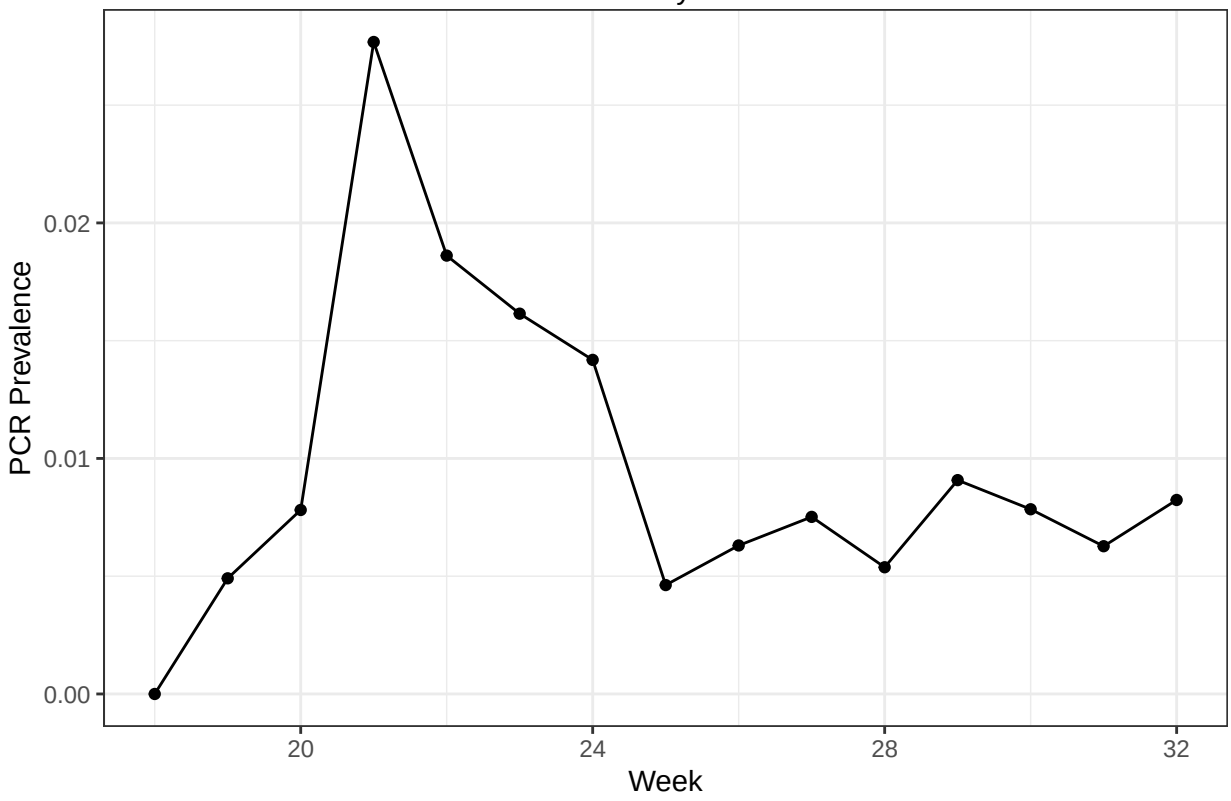
Time Trend, PCR Test and IGG Test

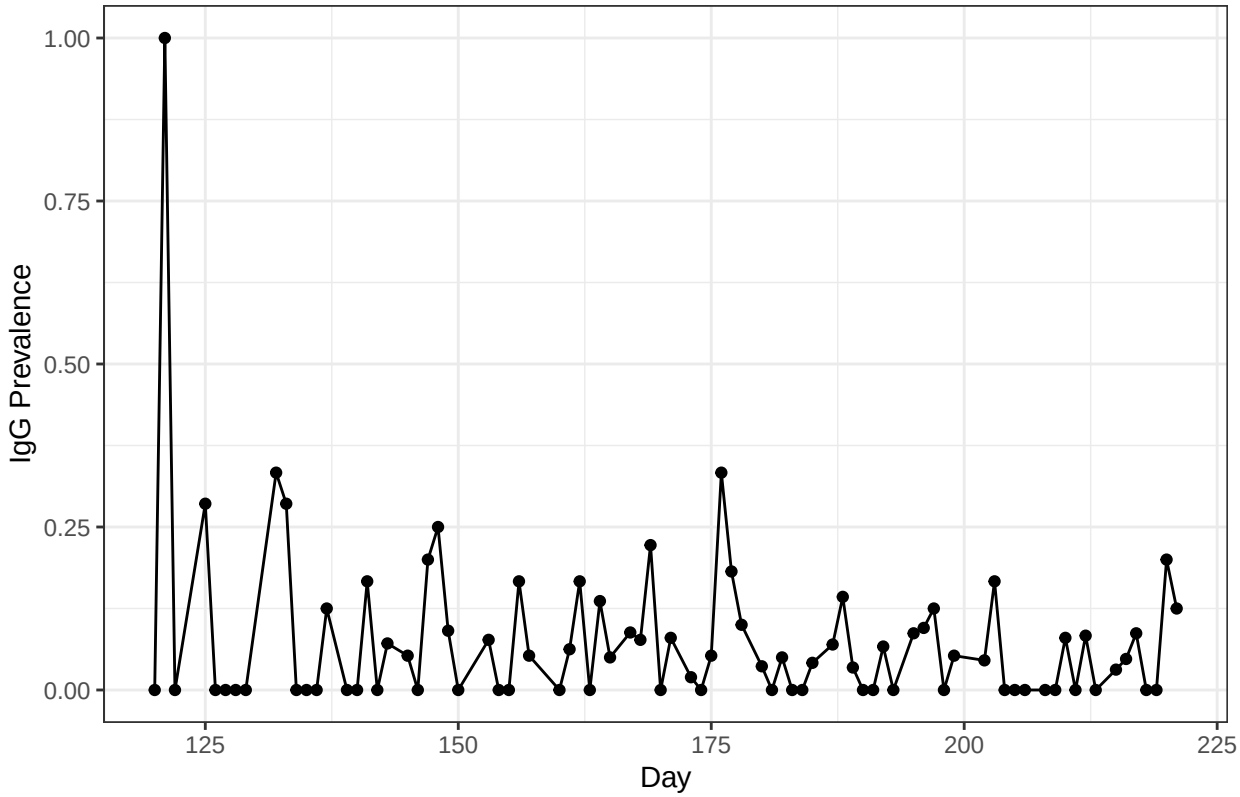
We are also interested in the time trend of the prevalence across the data we have. Specifically, we are interested in how the prevalence changes over time and whether the ratio between asymptomatic and symptomatic patients would be relatively constant across time. Besides, all the patients recored had the PCR test results but only a few of them have IGG results, so we are also interested in the correspondence between these two test results.

There are few patients patients data measured far before it or without an date so we dimply drop them off. To make the timing framework same as in asmptotic patients group and symptotic patients group, we unified the starting time as 5/2/2020 for both PCR and IGG tests, in both asymptotic and sumptotic patients.

We first present the prevalence over time for asymptomatic patients: for PCR test results we observed a peak in Week 3, then it dropped but recently started to rise again, while the IGG test we observed a peak at the begining, but this might because of the limited number of tests conducted and the positive rate of IGG test is decreasing over time. Notice that the reason why in Week 13 (the most recent week) the prevalence is 0 is because that many of the patients did not have their test results released yet.


PCR Prevalence
0.15
0.10
0.05
0.00
125
150
175
200
225
Day


PCR Prevalence
0.02
0.01
0.00
20
24
28
32
Week


1.00
0.75
0.50
0.25
0.00
IgG Prevalence
125
150
175
200
225
Day

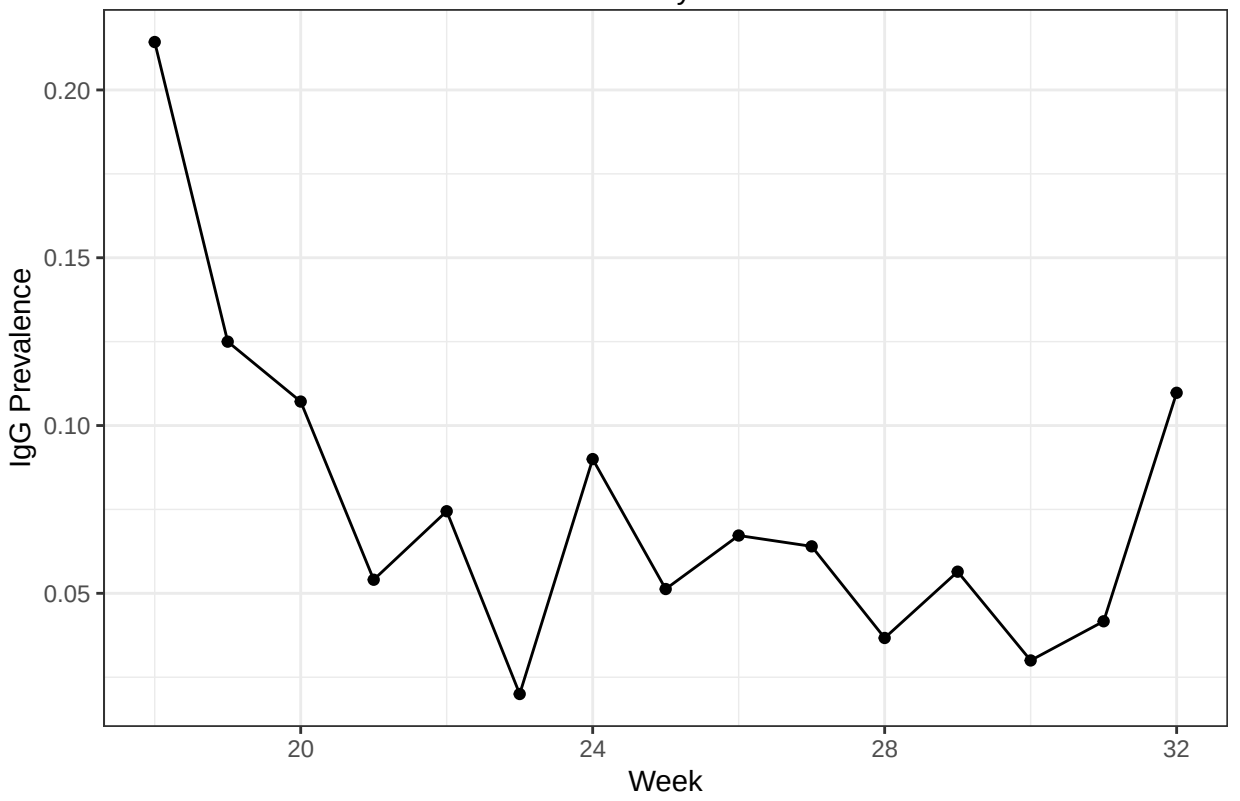
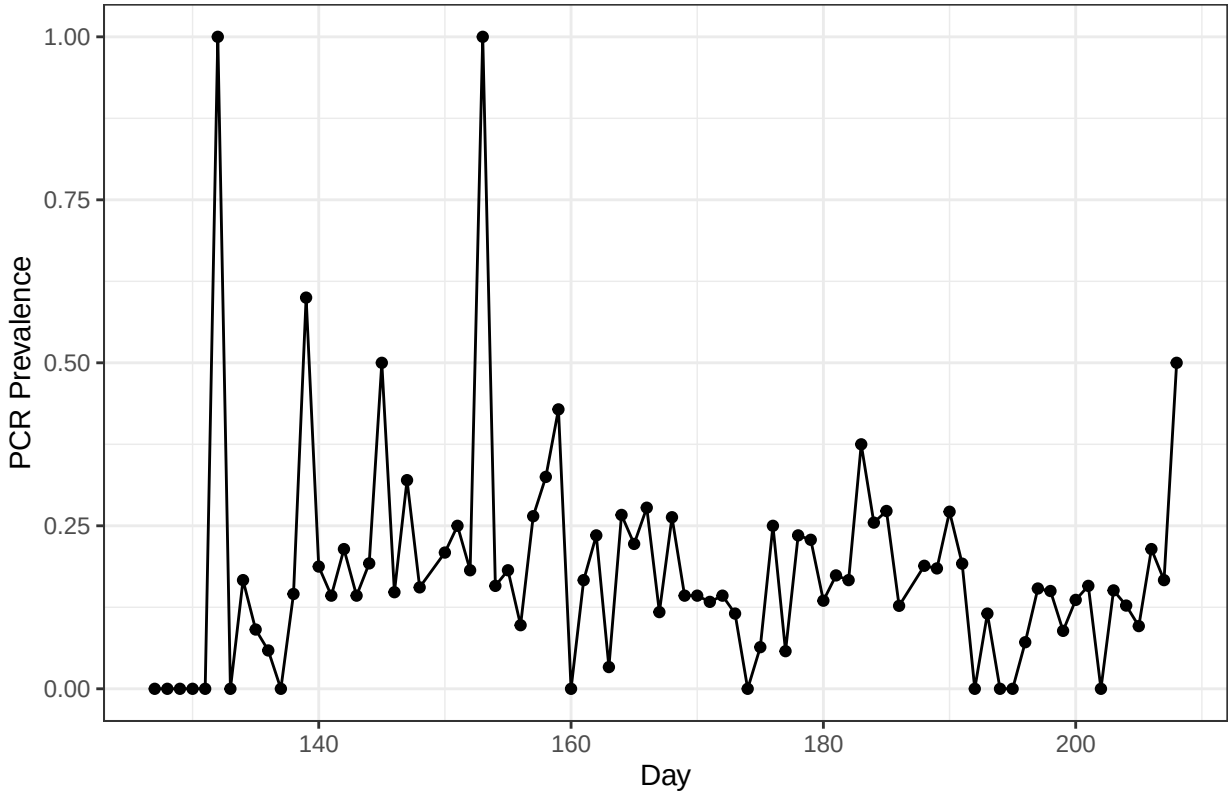
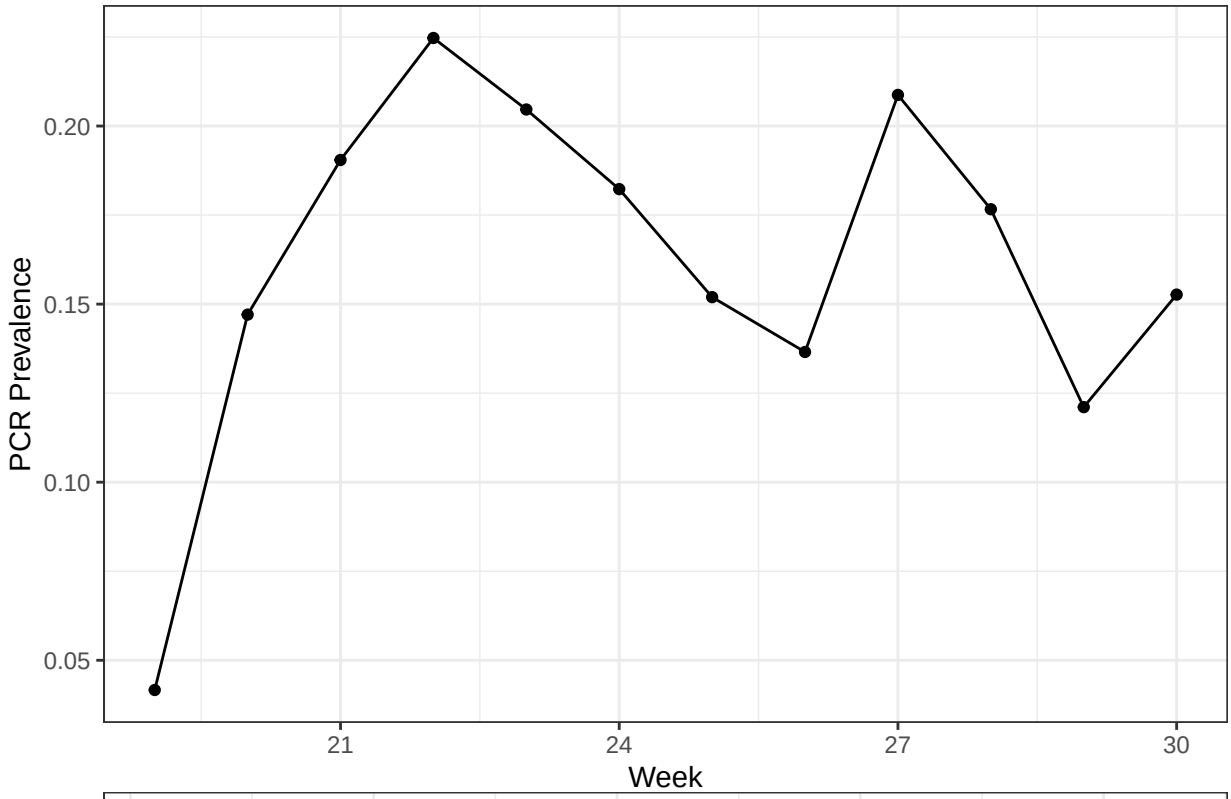

0.20
0.15
0.10
0.05
IgG Prevalence
20
24
28
32
Week

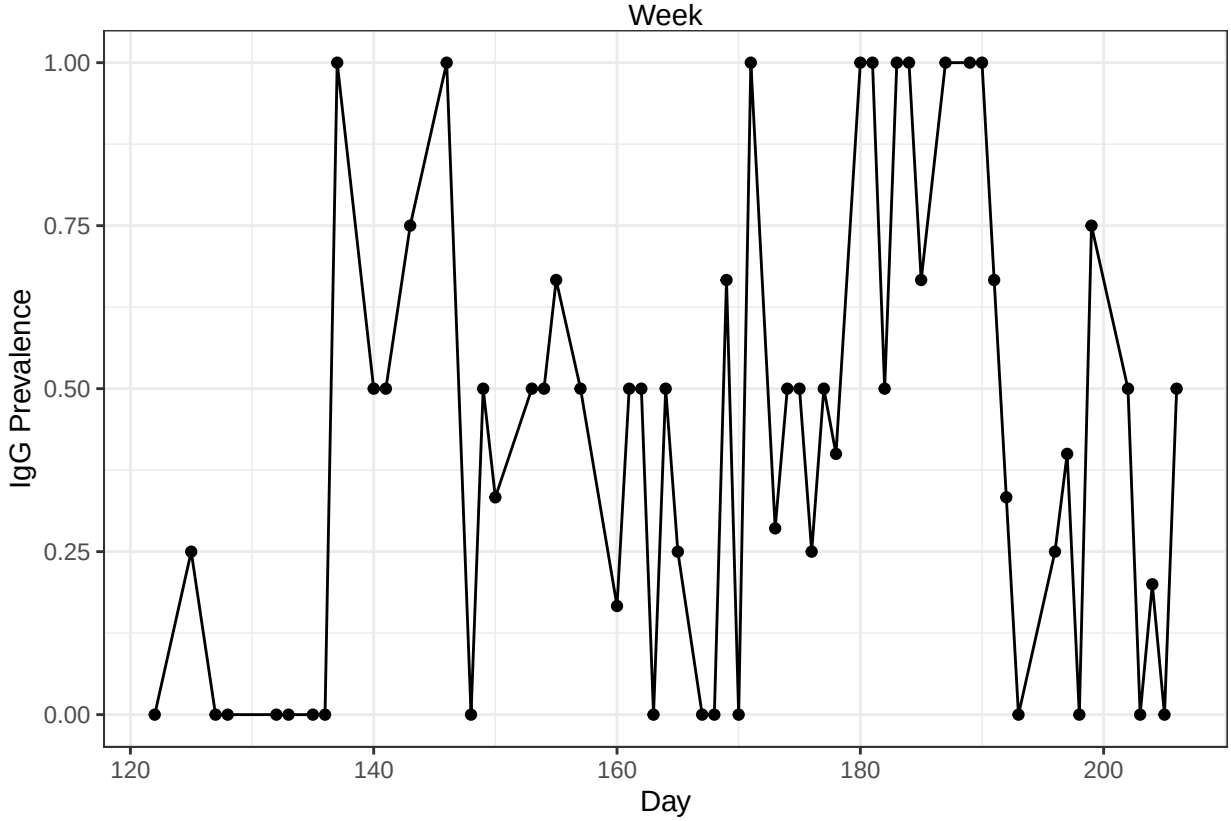
Table 2: When two test results are different for asymptomatic patients

same_result	different_result	same_positive_result	same_nagative_result
1245	69	20	1225

Then we presented the results for symptomatic patients: The start time is also 5/2/2020 for both PCR and IGG tests, and for PCR test results we observed a peak in Week 9 and it started to rise again recently while the IGG test prevalence reached its peak also in Week 9 but started to decrease recently.




PCR Prevalence
0.20
0.15
0.10
0.05
21
24
27
30
Week


IgG Prevalence
1.00
0.75
0.50
0.25
0.00
120
140
160
180
200
Day

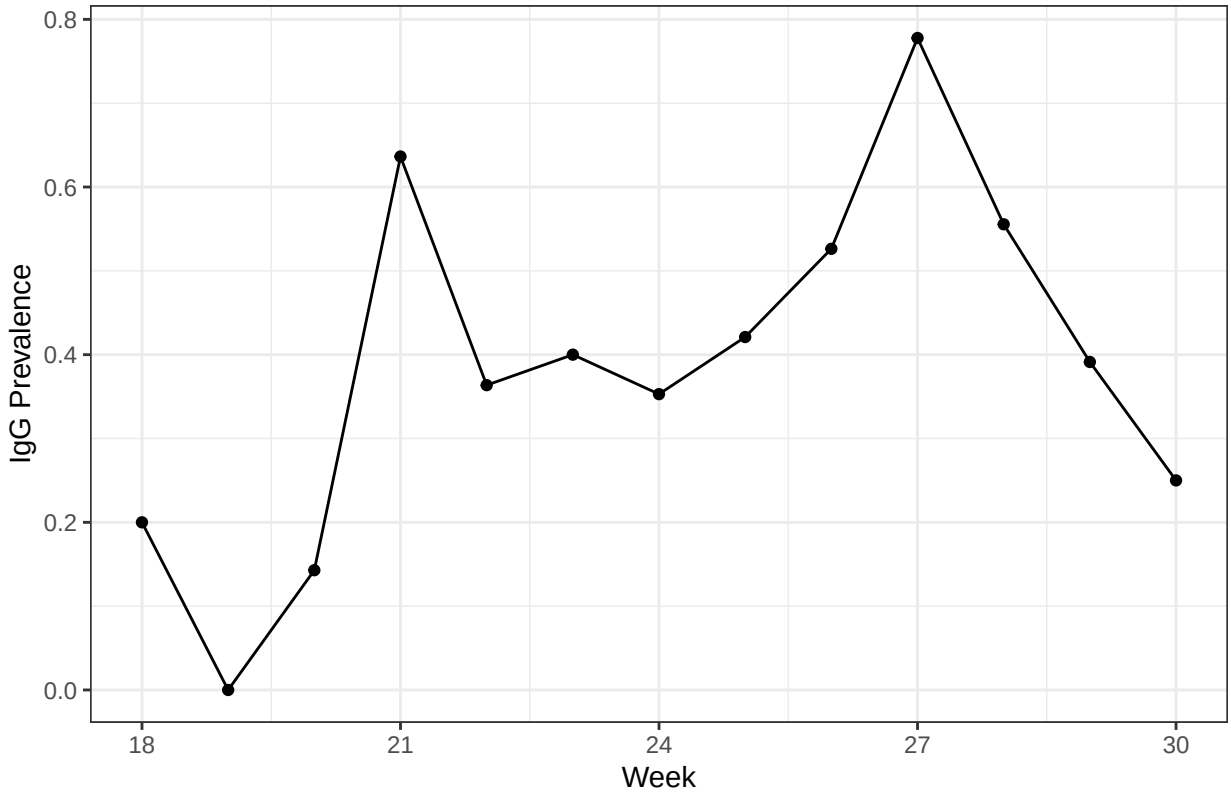
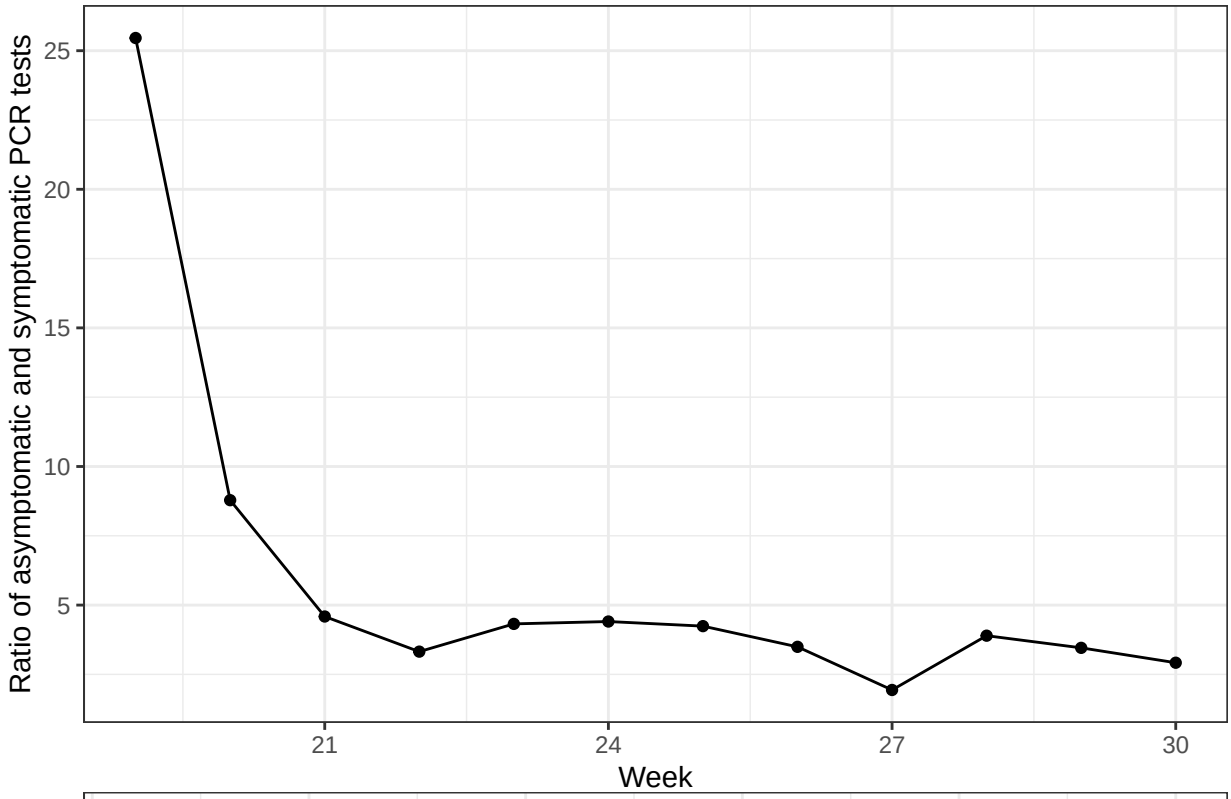
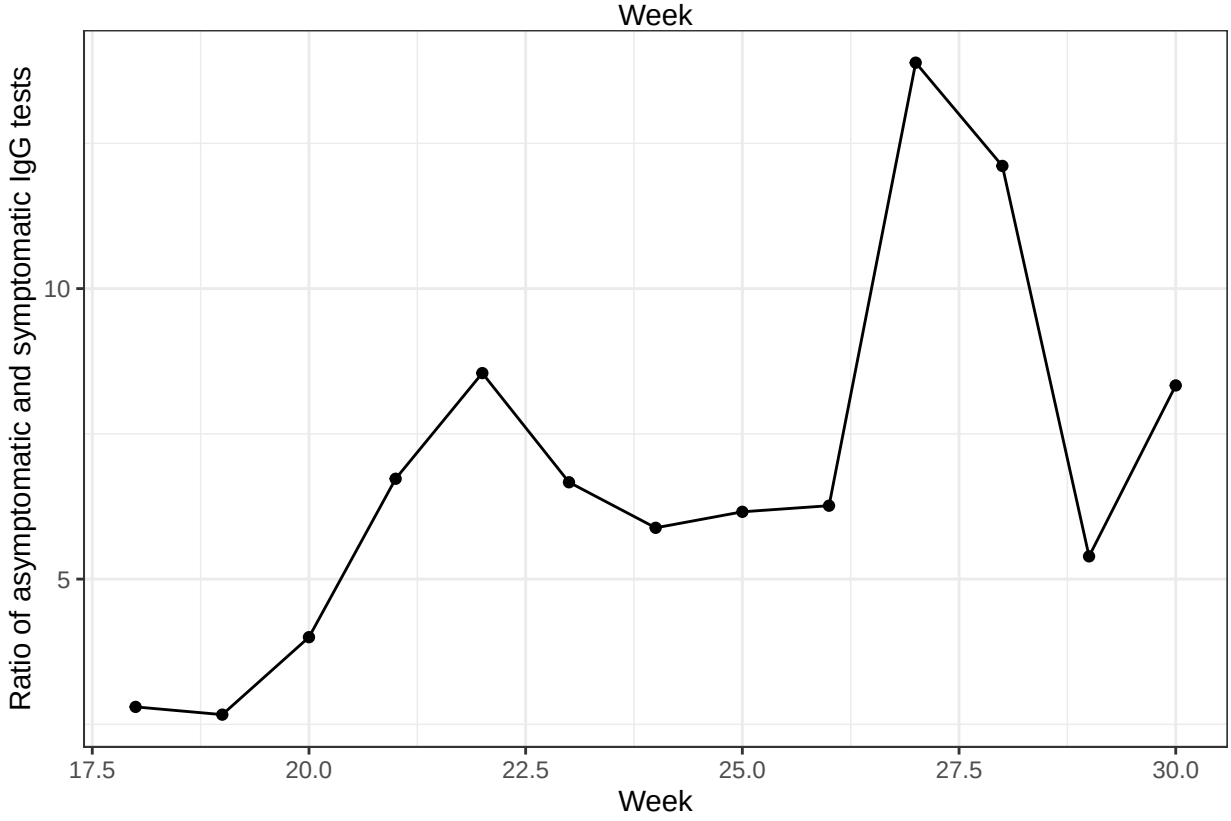


Table 3: When two test results are different for symptomatic patients

same_result	different_result	same_positive_result	same_nagative_result
119	41	38	81

Finally we would like to examine the hypothesis that whether the ratio between asymptomatic and symptomatic patients would be relatively constant across time. Since the available starting date of the asymptomatic and symptomatic patients dataset are not the same, we unified the time framekwork and set the common starting time as 5/2/2020 and we should avoid looking at the tails. From the graph we conclude that the ratio between asymptomatic and symptomatic patients would be relatively constant across time.

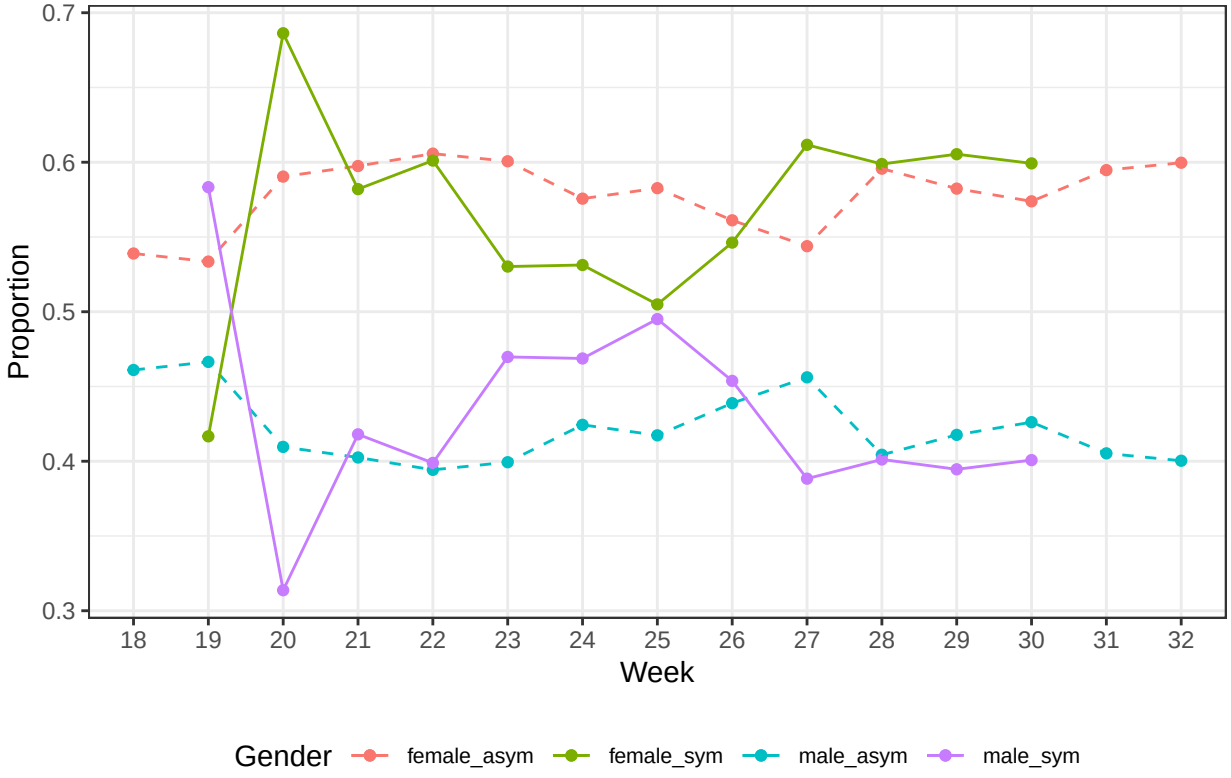

Ratio of asymptomatic and symptomatic PCR tests
25
20
15
10
5
21
24
27
30
Week


Ratio of asymptomatic and symptomatic IgG tests
10
5
17.5
20.0
22.5
25.0
27.5
30.0
Week

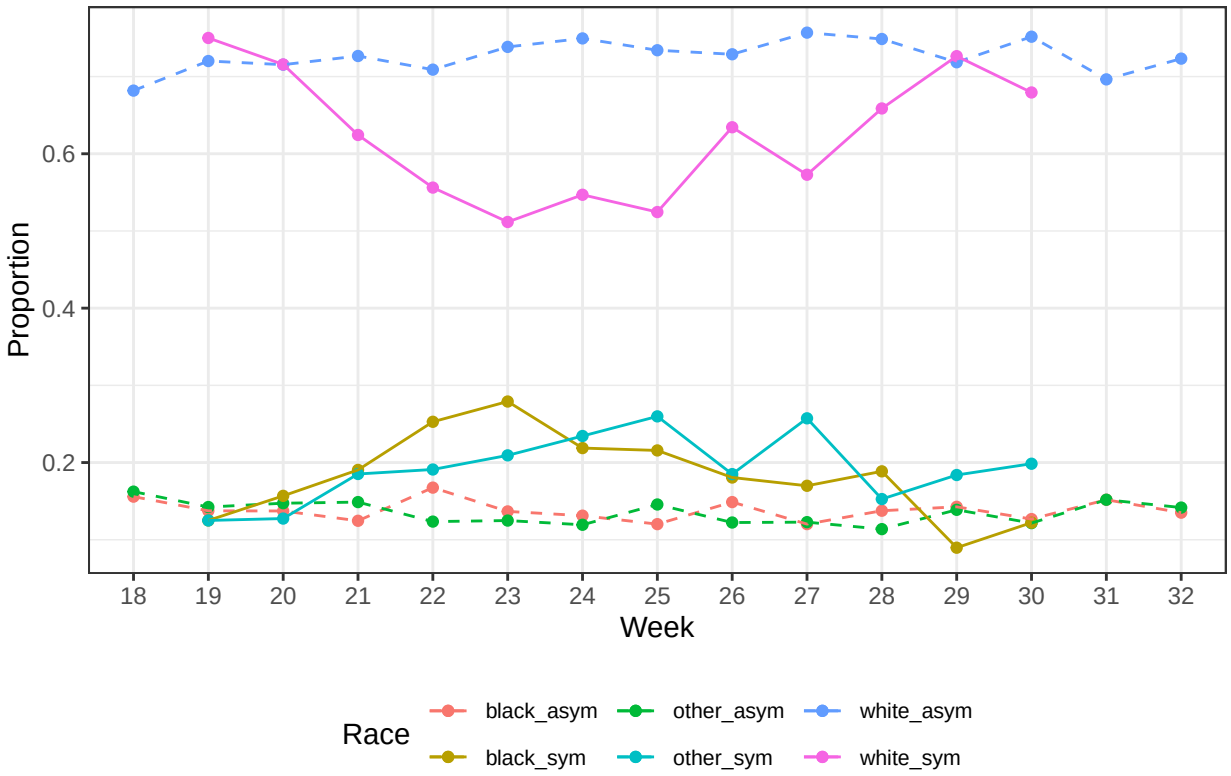
Plot of demographics of sample changing over time

We groupd our data by PCR test result date, again the start date for both asymptotic and symptotic patients are 5/2/2020.

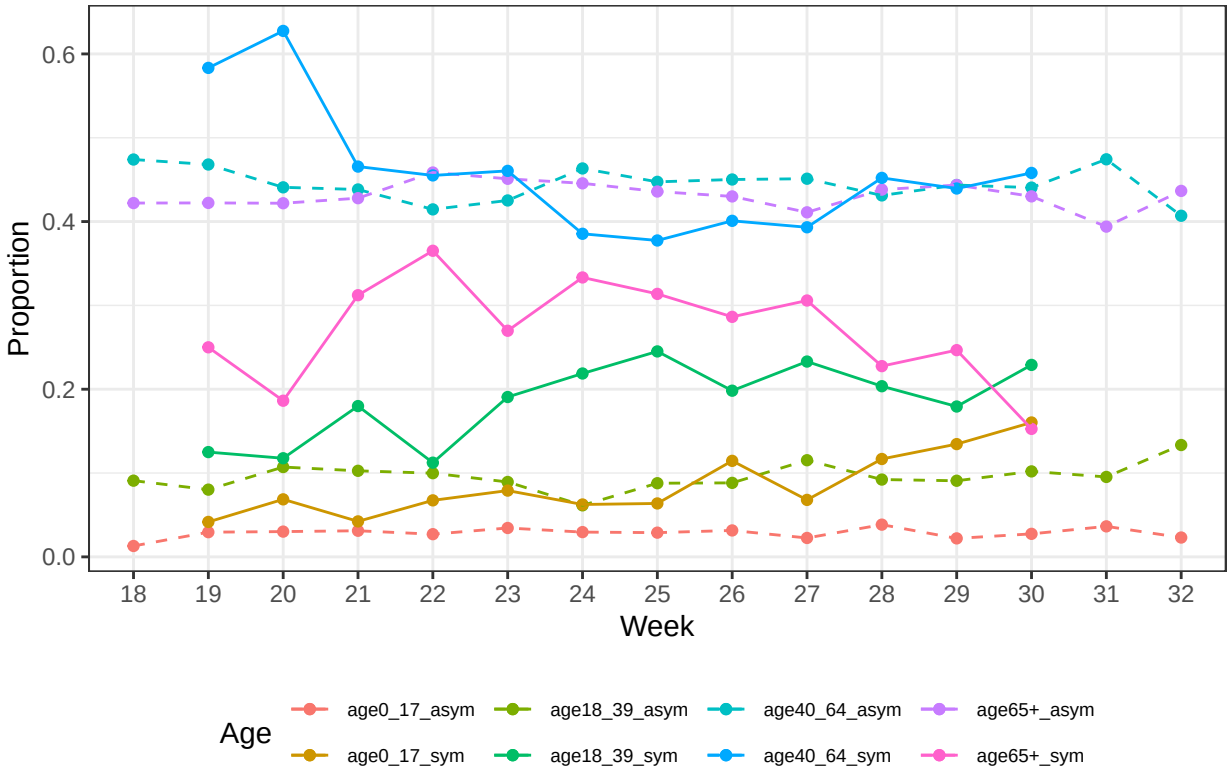
Relative sex percentages over time



Relative race pct over time



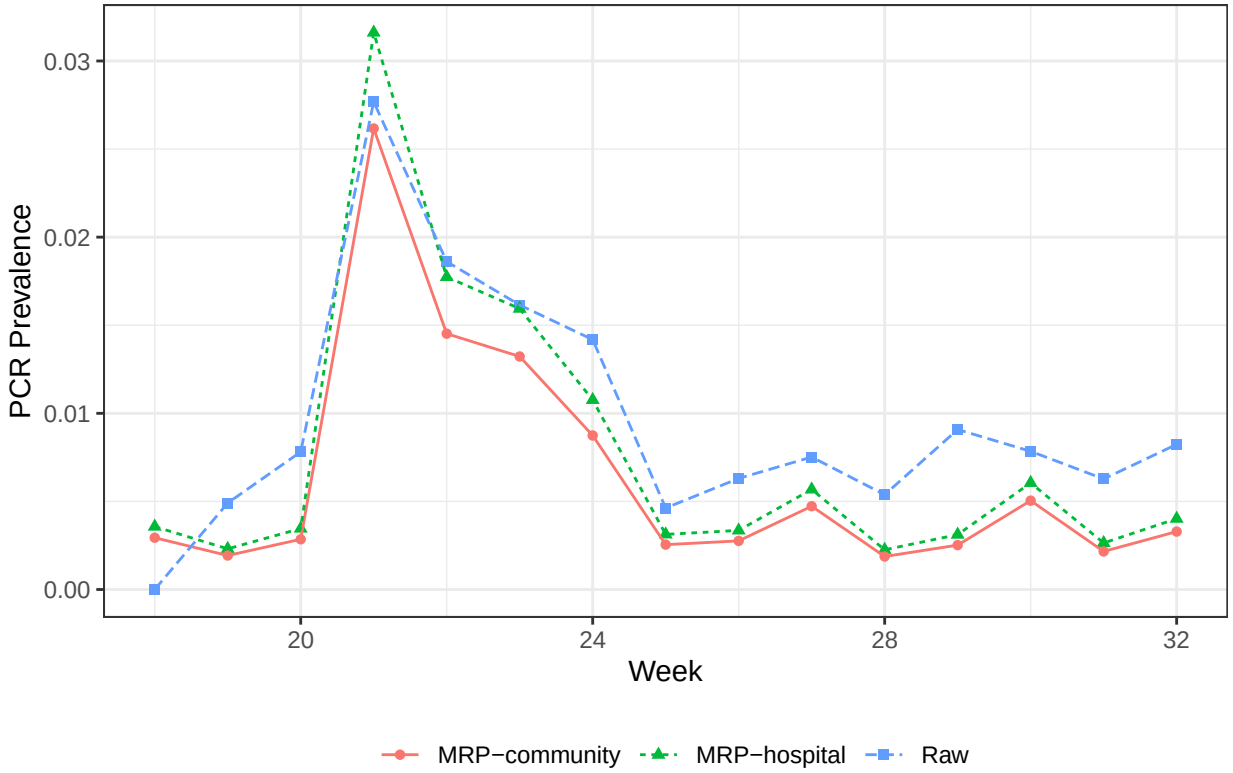
Relative age pct over time

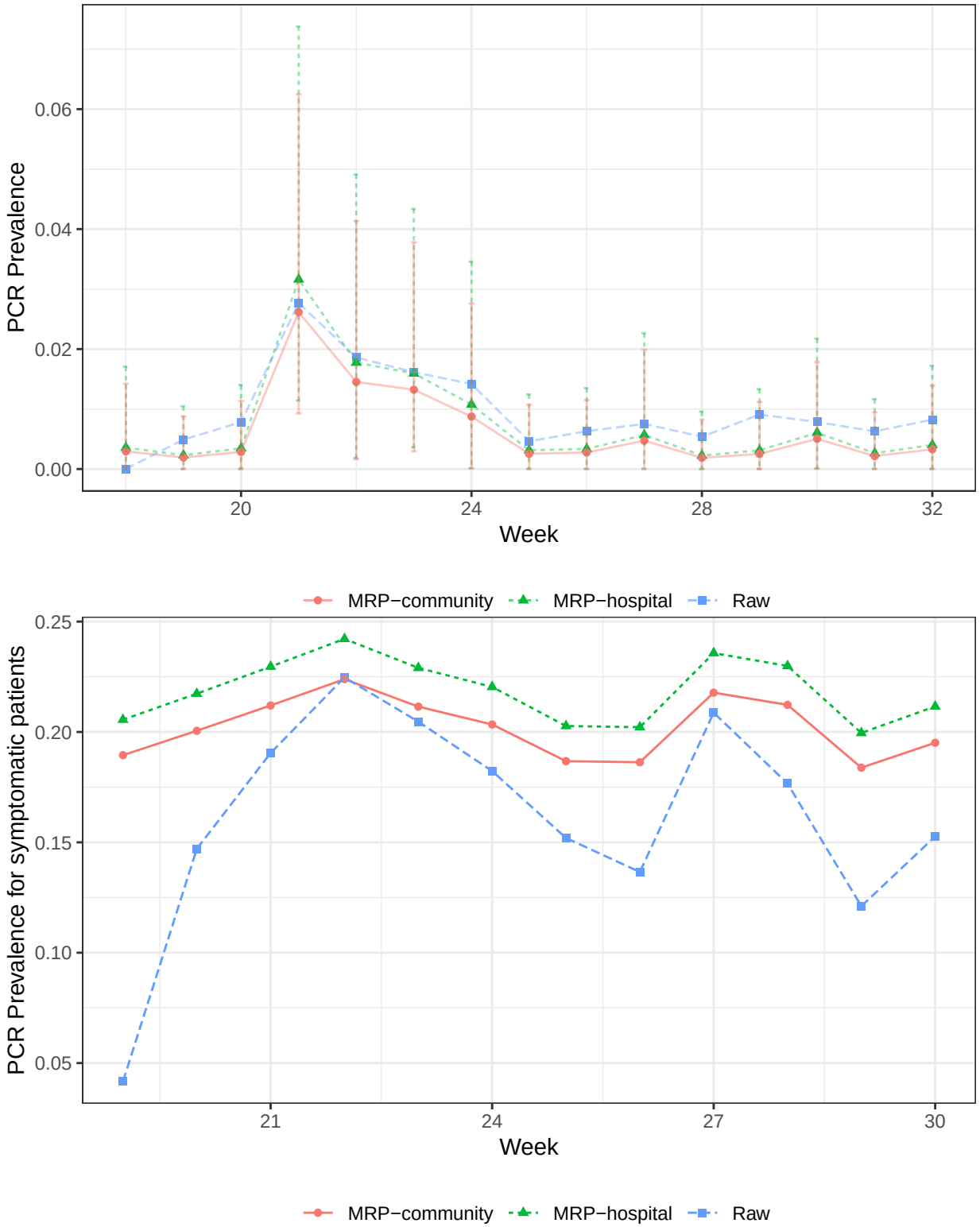


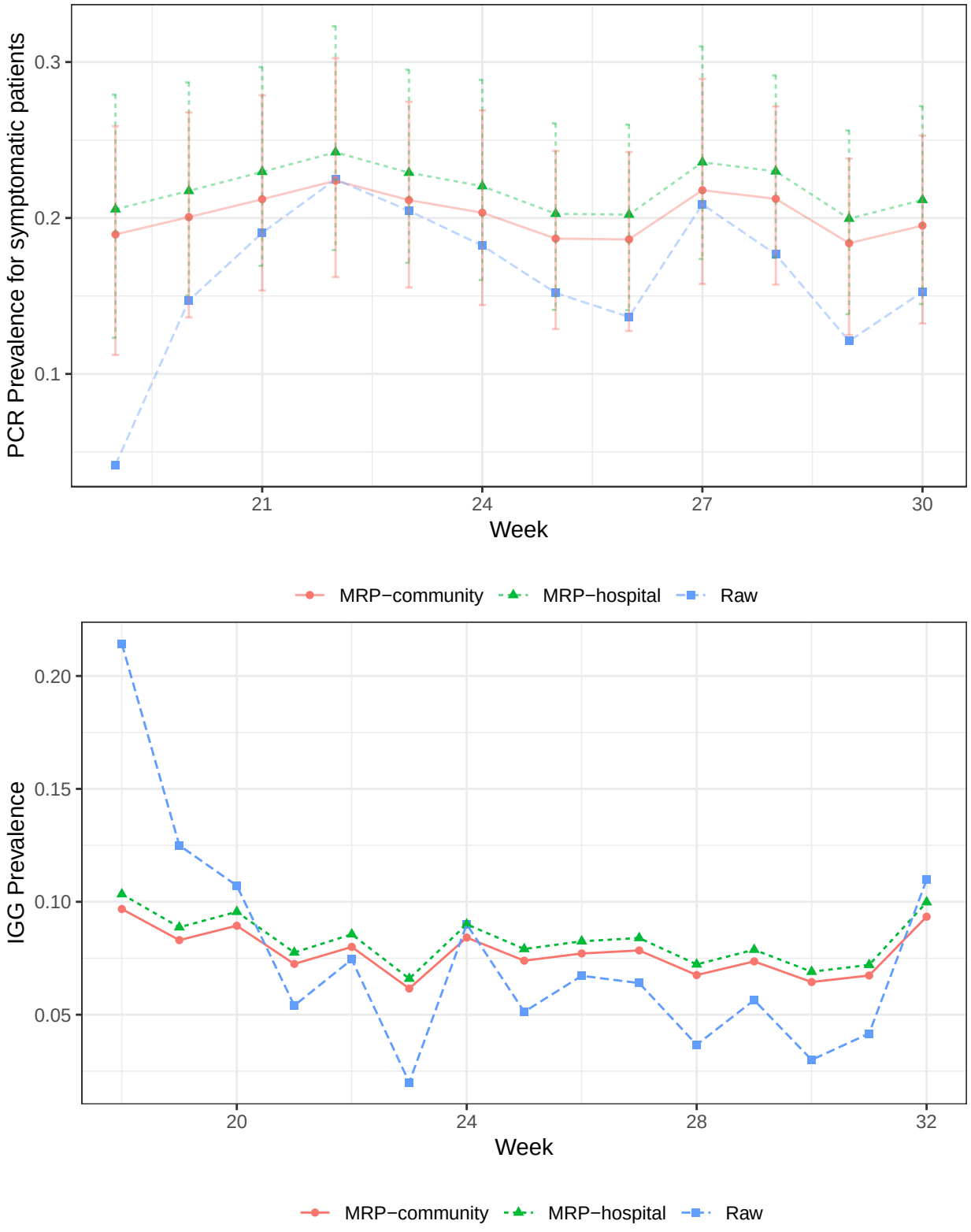
MRP

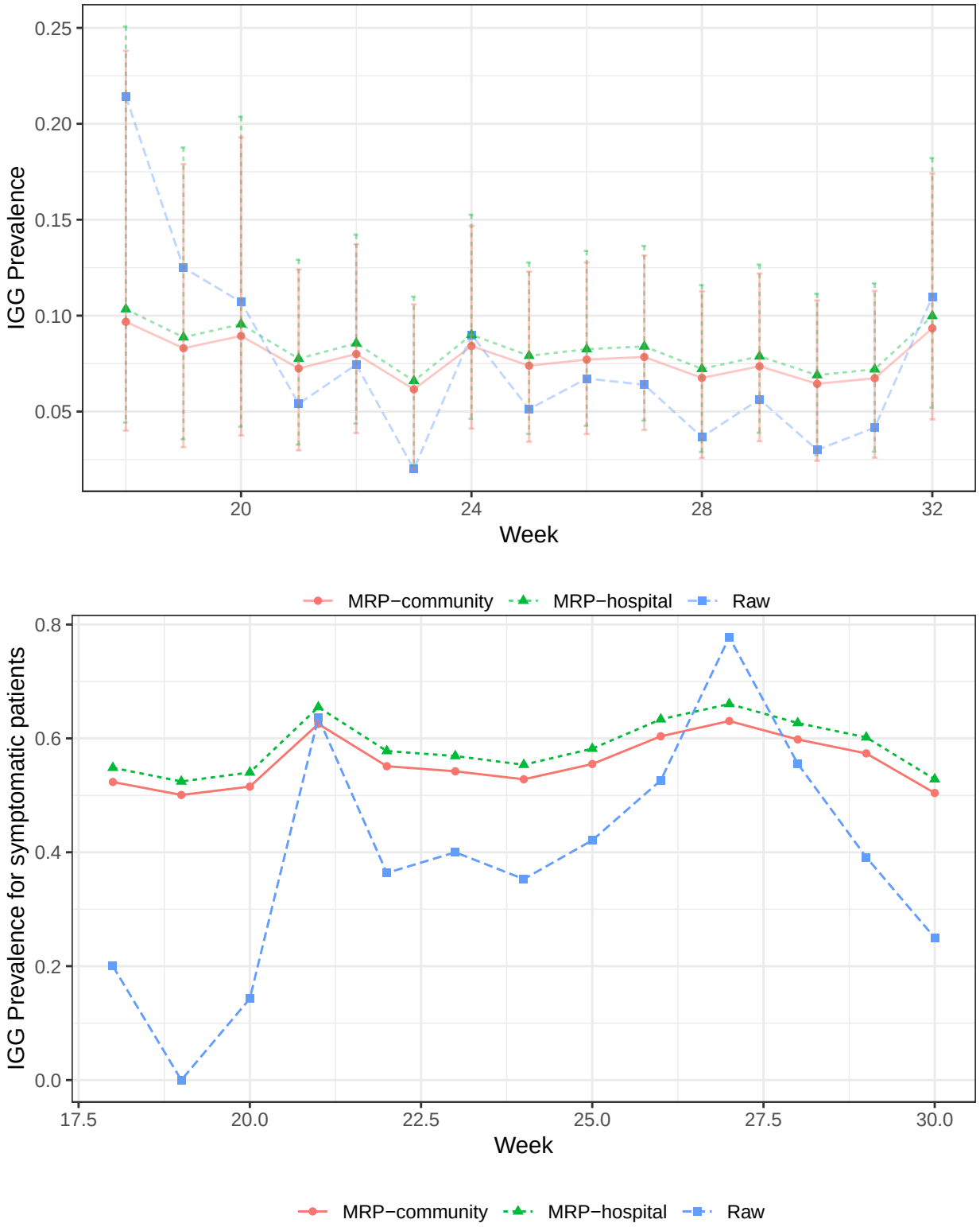
Besdies the patients data, we also collected the general population demographic data in the area, which plays key roles in our MRP modeling process. Note that since there are some discorrespondence in the zip code variable (there are 209 zip codes in the asymptotic patients data, 101 zip codes in the symptotic patients data but only 44 zip codes avaialable in the population data), so we currently did not put zip code as a vriable in our MRP modeling process. The model covariates include sex, age, rance, week and the two-way interaction between sex and age.

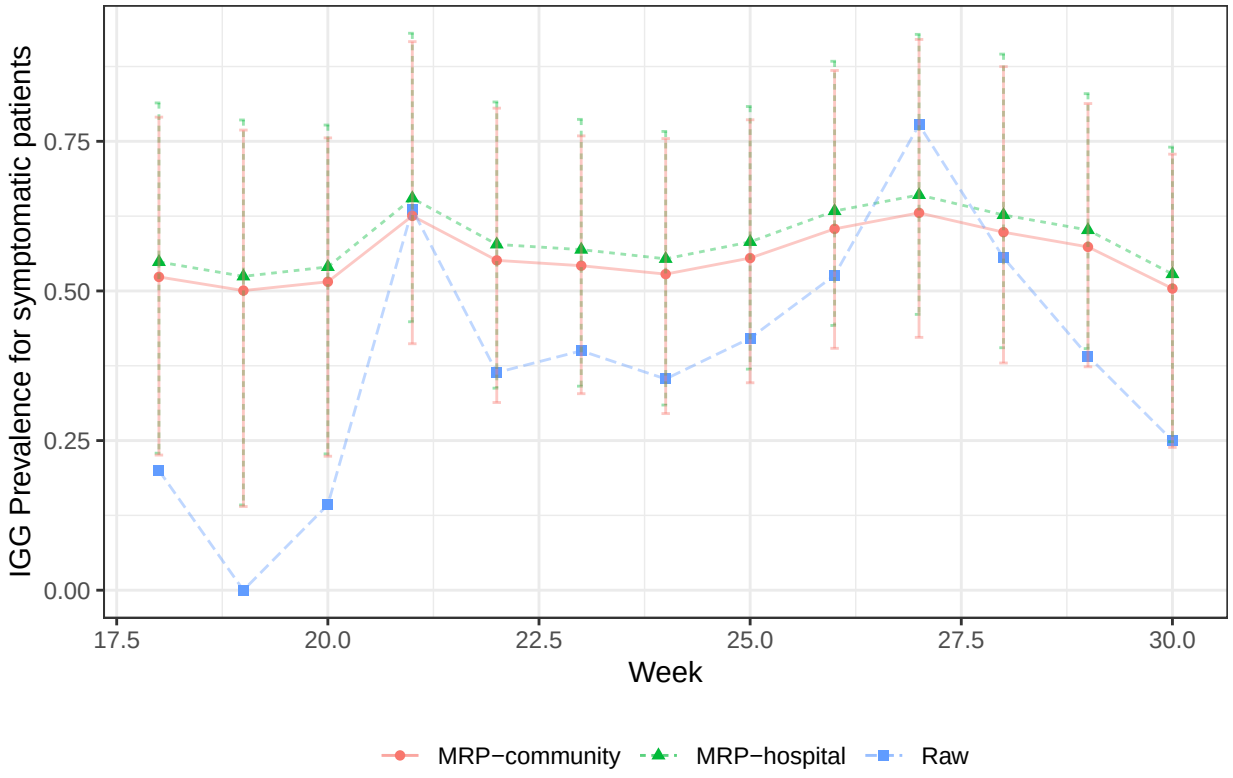
Here we performed 2 chians each with 10000 iterations (with 5000 iterations for warm-up) From the MRP results, we could see that the asymptotic patients and symptotic patients have quite different estimated prevalence.



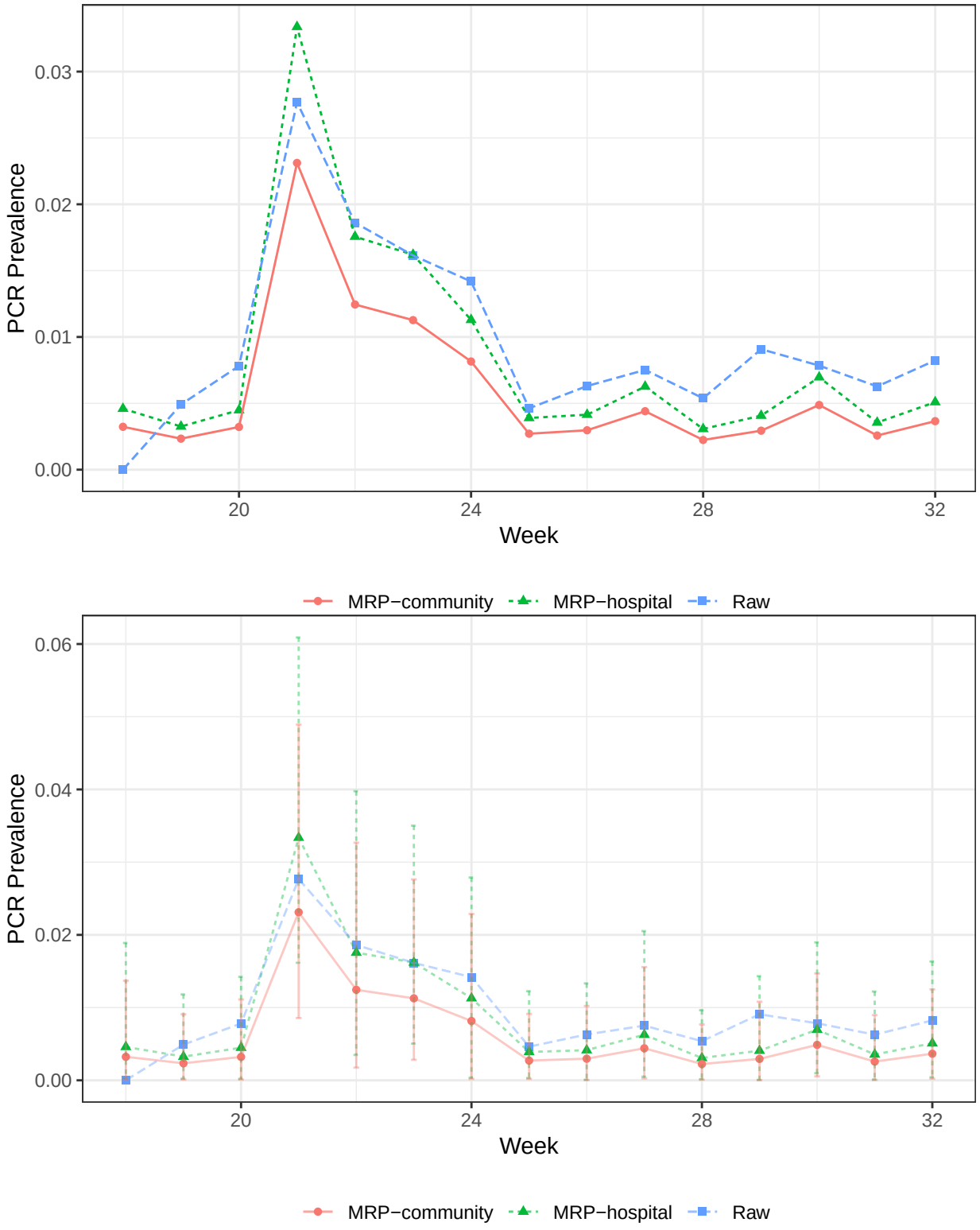

PCR Prevalence
0.06
0.04
0.02
0.00
20
24
28
32
Week
MRP−community
MRP−hospital
Raw
PCR Prevalence for symptomatic patients
0.25
0.20
0.15
0.10
0.05
21
24
27
30
Week
MRP−community
MRP−hospital
Raw


PCR Prevalence for symptomatic patients
0.3
0.2
0.1
21
24
27
30
Week
MRP−community
MRP−hospital
Raw
IGG Prevalence
0.20
0.15
0.10
0.05
20
24
28
32
Week
MRP−community
MRP−hospital
Raw


0.25
0.20
0.15
0.10
0.05
IGG Prevalence
20
24
28
32
Week
MRP−community
MRP−hospital
Raw
0.8
0.6
0.4
0.2
0.0
IGG Prevalence for symptomatic patients
17.5
20.0
22.5
25.0
27.5
30.0
Week
MRP−community
MRP−hospital
Raw


IGG Prevalence for symptomatic patients
0.00
0.25
0.50
0.75
17.5
20.0
22.5
25.0
27.5
30.0
Week
MRP−community
MRP−hospital
Raw

Model time using GP



Updated observations

We all have noted the PCR RNA positivity is quite low and stable. We have tried to develop some perspective on these findings. The incidence appears to have remained relatively stable at about 0.5% positive since the testing started around the first week of May, so now with 7000 or so patients over a 2-month interval. Our peak incidence of severe disease was around April 10-15 and we probably reached our lowest flow of patients into the hospital about 2-3 weeks thereafter; that is, right around the time we started testing. It seems reasonable, if not actually demonstrated, that the viral incidence we are seeing constitutes a reasonable approximation of the asymptomatic incidence when viral counts are at or near the near-term low. The symptomatic data you now have available may add some color to that interpretation, but is obviously subject to selection biases that may have changed over the course of time, as physicians have been more likely to test more mildly symptomatic people as time has progressed. The naïve expectation would be that symptomatic positive rate would be expected to decrease somewhat due to that effect. Or increase as physicians became better able to diagnose? Not absolutely sure on this one.

As has been discussed, the very low rate of positivity presents other difficulties in statistical analysis. We have a working assumption of an analytical false positive of 0%, but surely there may be contamination issues, so we are in a range where zero incidence presumably lies within some reasonable uncertainty interval. It may be helpful to consider other ways to decide if these are true positives, if that issue becomes important. We could consider using a different assay on that patient sample, if still possible, or using IgG verification, an imperfect other possibility. We will also need to account for the other end of the problem: namely, that the sensitivity is insufficient to pick up more than 70% of true positives and that the asymptomatic nature and timing of sample may decrease that sensitivity substantially. This issue will need to be incorporated into any interpretation.

It will remain interesting to see if there is any clear upward trend on the PCR RNA data as the disease increases in the community. The metric may not end up being predictive for any next wave, but that wave may provide verification of the conjecture or shoot it full of holes. Either way is interesting in its own way.

We may find it useful to use the data to come up with an estimate of the percentage of asymptomatic incidence with respect to symptomatic incidence. Most talking head chatter currently is suggesting that asymptomatic incidence is quite high. Our early numbers at least question that conjecture. It strikes me that there may be a path toward integrating the two datasets to come up with a reasonable estimate of that incidence, though there are clearly some apples-to-oranges problems with structuring that analysis. If that nut can be cracked, we may end up encouraging others to use their own experiences under this model to reach their own estimates and if there is some concurrence, the insights might be revealing. In any event, any reasonable analysis that can help narrow down how common asymptomatic viral shedding may be would be very helpful to the scientific conversation today.

A general protocol for monitoring and analysis

The monitoring and analysis plan we have set up can be implemented at other hospitals around the country and the world. The biggest immediate challenge for hospital staffs would be the statistical analysis, but this we have already programmed. The larger issue is that much can be learned by combining information from different hospital systems, as this would alleviate the small-sample problem that we have encountered. The best way forward might be for individual hospitals and medical groups to gather and analyze their data as we propose in this article, with all the (de-identified) data shared in a common public repository, so then it will be possible for researchers to learn more by analyzing trends as they develop in the pooled dataset. This could be similar to other national data pooling efforts such as performed by Delphi Research Group (2020) in the United States and Rossman et al. (2020) in Israel.

Gelman, A, and B Carpenter. 2020. "Bayesian Analysis of Tests with Unknown Specificity and Sensitivity." *Journal of the Royal Statistical Society Series C (Applied Statistics)* forthcoming.

Gelman, Andrew, and Thomas C. Little. 1997. "Poststratifcation into Many Cateogiries Using Hierarchical

Logistic Regression." *Survey Methodology* 23: 127–35.

Si, Yajuan, Rob Trangucci, Jonah Sol Gabry, and Andrew Gelman. 2020. "Bayesian Hierarchical Weighting Adjustment and Survey Inference." Survey Methodology (accepted); https://arxiv.org/abs/1707.08220.