Name: Holin Chen

Imputation of missing data: imputing symptom onset dates based on the COVID-19 case data in Georgia

Background

The disease transmission group at Emory was using the data of confirmed COVID-19 cases from February 1st , 2020 to July 13th , 2020 provided by the Georgia Department of Public Health to study the patterns of COVID-19 transmission dynamics in Georgia by estimating the time-varied reproduction numbers in all 159 counties in Georgia. A transmission probability model adapted by previous study (Teunis 2013 [1] and Wang 2020 [2]) was built to estimate the county-level reproduction numbers based on time intervals between symptom onset dates of any pair of cases living in a same county. However, among 118,497 confirmed cases, 48,893 (41.3%) cases were missing their symptom onset dates. To study the transmission dynamics in Georgia, we need to impute the missing symptom onset dates based on other given date information.

Method

We handled missing symptom onset dates by making use of the dates of first positive specimen collection or the dates of reporting/confirmation as predictors. As shown in figure 1, among the complete data, most averaged observed delay days from symptom onset date to first specimen collection date are gradually declining in a stable way from the beginning of the reported dates of February 28th, 2020 to the reported date of July 13th, 2020. Most cases in March had to wait for more than a week to be tested after developing symptoms while most cases after May only needed wait 2 or 3 days to be tested. This pattern was expected as the COVID-19 testing capacity was increased as time went on with more testing sites opened and better testing technology developed. However, the averaged observed delay days from symptom onset date to date of reporting are much more varied along with the reported dates, which pattern is less obvious than that of delay days for date of specimen collection above. This could be caused by the large variance of delay of getting the results after testing.

Based on this observation, we decided to prioritize using the dates of first specimen collection as predictors to impute the delay days from symptom onset dates to first specimen collection dates for data with dates of first specimen collection but without symptom onset dates, using the model built by all the complete data with observed symptom onset dates and observed dates of first specimen collection. Then we estimated the missing symptom onset dates by subtracting the delay days from the dates of first specimen collection. For data without both the symptom onset dates and dates of first specimen collection but with the dates of reporting, we used the dates of reporting as predictors to impute the delay days from symptom onset date to reporting dates, using the model built by all the complete data with observed symptom onset dates and dates of reporting, and subtracting the delay days from the dates of reporting to estimate the missing symptom onset dates.

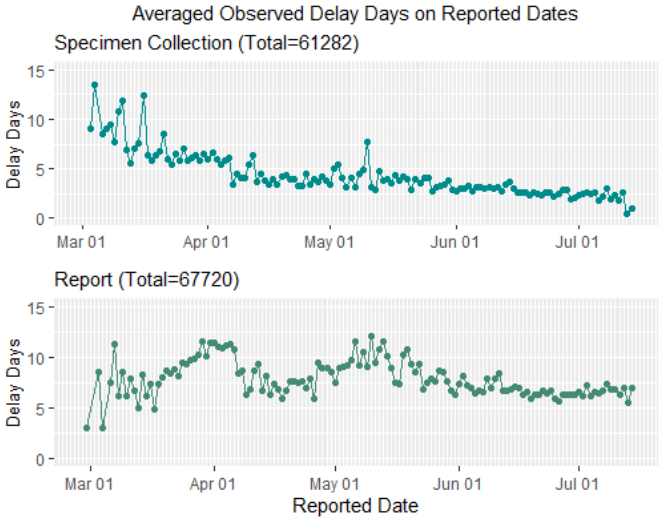

Figure1: the averaged observed delay days from symptom onset date to date of first specimen collection based on reported dates from February 28th, 2020 to July 13th, 2020, and the averaged observed delay days from symptom onset date to date of reporting based on reported dates from February 28th, 2020 to July 13th, 2020. The plots did not include data before February 28th, 2020 as the cases were extremely scarce during this time

We explored three different regression methods to impute the missing data of symptom onset dates, which were log-linear regression, Possion regression, and negative binomial regression, and we finally decided to use negative binomial regression method to impute the missing data as the imputed outputs align well with the observed data while the outputs imputed by the other two methods are not. The imputation processes in these three regression methods are elaborated as follow.

Log-linear regression

The distributions of the numbers of delay days for first specimen collection and the numbers of delay days for report are both right-skewed (Figure 2a and b). To set the numbers of delay days to be normally distributed, we log-transformed the delay days. We used the delay days in log scale as the outcomes and the observed dates of first specimen collection or observed dates of reporting as the predictors to build these two linear regression models. Let s_i be the date of first specimen collection, let r_i be the date of reporting and let o_i be the symptom onset date. d_{si} is the delay day from symptom onset date to first specimen collection date, which $d_{si} = s_i - o_i$. d_{ri} is the delay day from symptom onset date to reporting date, which $d_{ri} = r_i - o_i$. The log-linear regression models are $\log(d_{si}) = \beta_0 + \beta_1 s_i$ and $\log(d_{ri}) = \beta'_0 + \beta'_1 r_i$.

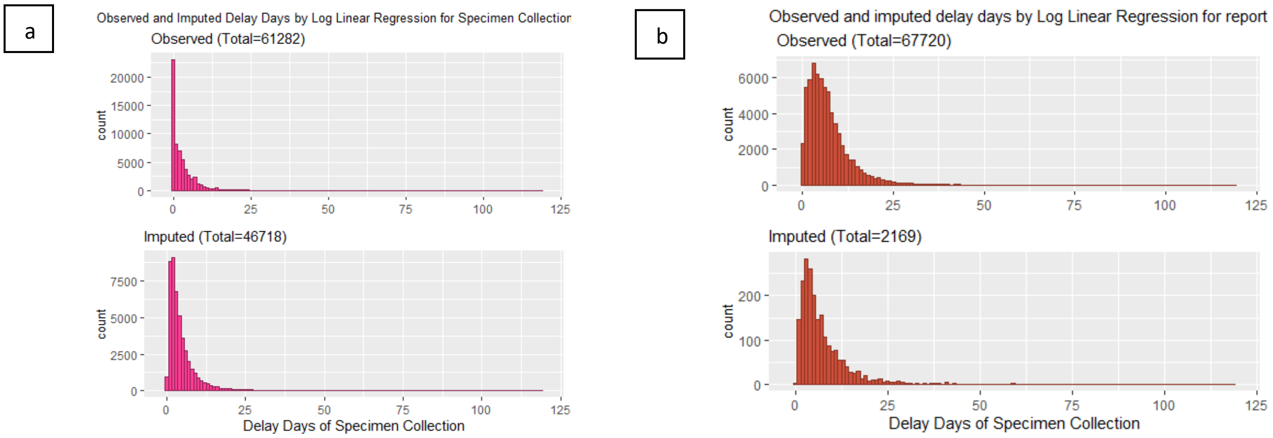

Figure 2, (a) the distribution of numbers of delay days from symptom onset dates to first specimen collection dates for observed data and imputed data by log linear regression. (b) the distribution of numbers of delay days from symptom onset dates to reporting dates for observed data and imputed data by log linear regression.

For each case with a reported s_i or r_i, we obtained o_i by predicting the value for the delay days in log scale from one of the two lineal regression model, and generated a random variable, X_i, as the exponential of a normally distributed random variable with parameter mean in predicting value for the delay days in log scale and variance given by the prediction error obtained from the regression model. Then the imputed symptom onset date is $\hat{o}_i = \hat{s}_i - [X_i]$ or $\hat{o}_i = \hat{r}_i - [X_i]$, where $[X_i]$ is the rounded value.

However, the imputed data is not fitted well to the observed data by using this method. As shown in the figure 2a and b, the distributions of numbers of imputed delay days are biased away from the that of observed delay days. Besides, it is hard for log transformation to deal with data that has 0 values as it will give back infinite values, so we have to set the 0 values as missing when building these models, while there are many real cases having 0 delay days without days of waiting for specimen collection or report in the original data, which made the imputed results can not reflect the characteristics of observed data.

Possion regression

Then we tried the Possion regression method to impute the missing symptom onset dates. We assumed that the distributions of the numbers of delay days for first specimen collection and the numbers of delay days for report are both belong to Possion distribution. We built the two Possion regression models, which formulas are $\log(d_{si}) = \beta_0 + \beta_1 s_i$ and $\log(d_{ri}) = \beta'_0 + \beta'_1 r_i$ assuming the d_{si} and d_{ri} are following Possion distribution. For each missing data, we generated a random variable as the estimated delay day from the Possion distribution which parameter λ is the exponential of the predicting value of delay days in log scale. The imputed outputs are as follows.

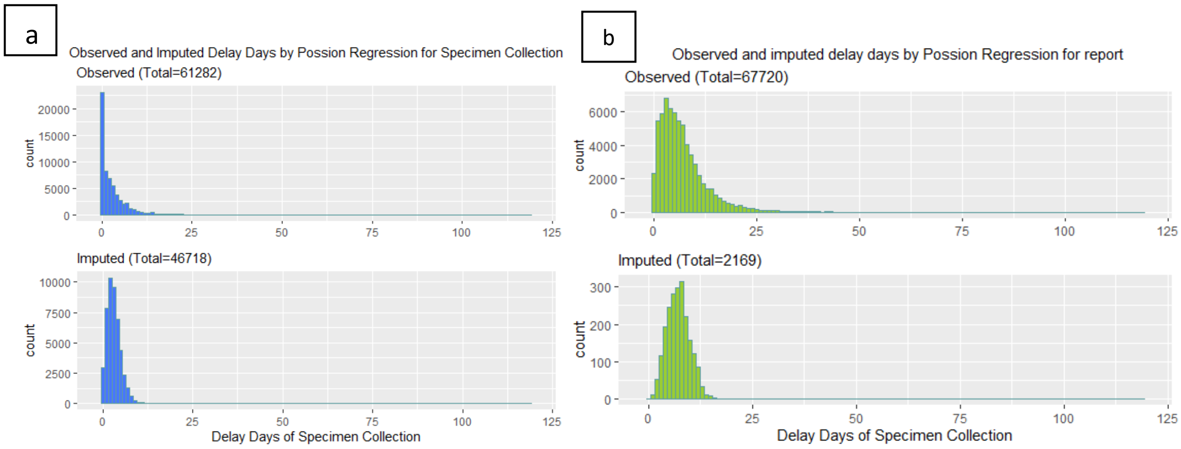

Figure 3a: the distribution of numbers of delay days from symptom onset dates to first specimen collection dates for observed data and imputed data by possion regression. Figure 3b: the distribution of numbers of delay days from symptom onset dates to reporting dates for observed data and imputed data by possion regression.

As shown in figure3, the distributions of numbers of delay days imputed by Possion regression were very different from that of distributions from observed data, so it may be not appropriate to use Possion regression to impute the missing symptom onset dates.

Negative binomial regression

Finally, we tried the negative binomial regression method. Assuming the distributions of the numbers of delay days for first specimen collection and the numbers of delay days for report are both belong to negative binomial distribution, we built the two negative binomial regression models, which formulas in linear form are also $\log(d_{si}) = \beta_0 + \beta_1 s_i$ and $\log(d_{ri}) = \beta'_0 + \beta'_1 r_i$ by using log as link in these regressions. Let $\hat{d}_{si}$ and $\hat{d}_{ri}$ be the predicting values from these two negative binomial regression models. For each missing data, we generated a random variable X_i as the estimated delay day from the negative binomial distribution which parameter mean $\mu = \exp(\log(\hat{d}_{si}))$ or $\mu = \exp(\log(\hat{d}_{ri}))$ and shape parameter θ obtained from the negative binomial regression model. Then the imputed symptom onset date is $\hat{o}_i = \hat{s}_i - X_i$ or $\hat{o}_i = \hat{r}_i - X_i$. The imputed outputs are as follows.

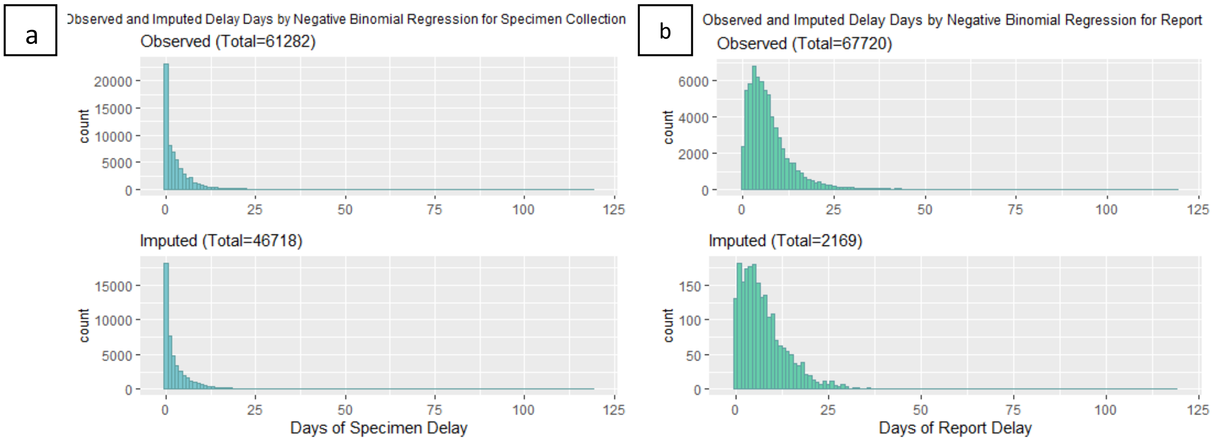

Figure 4a: the distribution of numbers of delay days from symptom onset dates to first specimen collection dates for observed data and imputed data by negative binomial regression. Figure 4b: the distribution of numbers of delay days from symptom onset dates to reporting dates for observed data and imputed data by negative binomial regression.

As shown in figure 4, the distributions of numbers of delay days imputed by negative binomial regression fit well with that of distributions from observed data, proving that the negative binomial regression method is a good choice to impute the missing symptom onset dates. The procedures of the imputation process are summarized by the flow chart as follows.

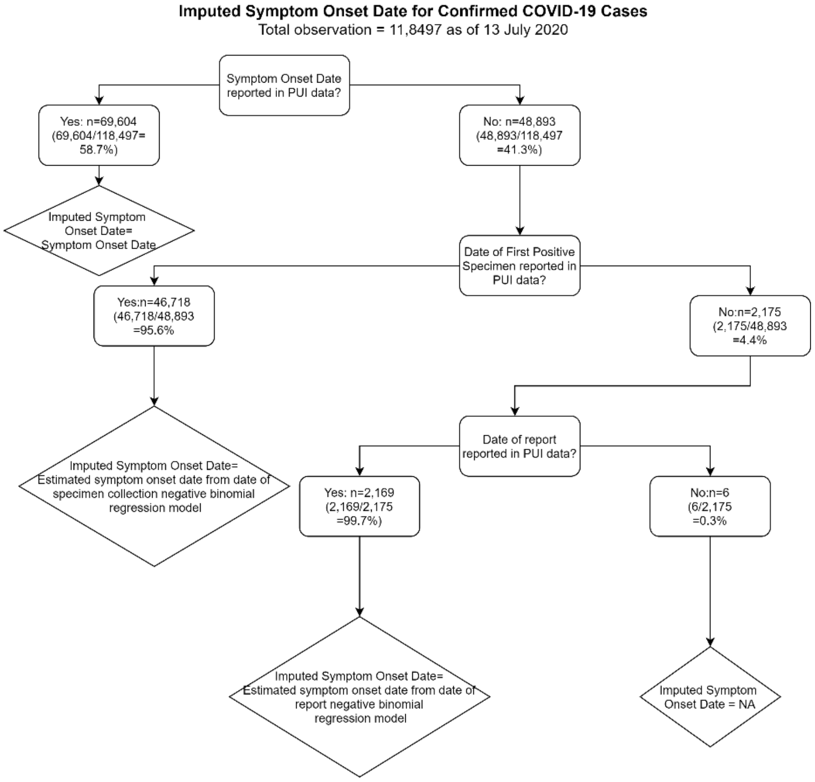
Imputed Symptom Onset Date for Confirmed COVID-19 Cases
Total observation = 11,8497 as of 13 July 2020
Symptom Onset Date reported in PUI data?
Yes: n=69,604 (69,604/118,497= 58.7%)
No: n=48,893 (48,893/118,497 =41.3%)
Imputed Symptom Onset Date= Symptom Onset Date
Date of First Positive Specimen reported in PUI data?
Yes:n=46,718 (46,718/48,893 =95.6%
No:n=2,175 (2,175/48,893 =4.4%
Imputed Symptom Onset Date= Estimated symptom onset date from date of specimen collection negative binomial regression model
Date of report reported in PUI data?
Yes: n=2,169 (2,169/2,175 =99.7%)
No:n=6 (6/2,175 =0.3%
Imputed Symptom Onset Date= Estimated symptom onset date from date of report negative binomial regression model
Imputed Symptom Onset Date = NA

Figure 5: the process of imputing symptom onset dates by negative binomial regression method

Result

The symptom onset dates of 48,893 cases were imputed. Only 6 cases without any of the date information could not be imputed and were dropped out from the analysis. To exam whether the summarized statistics of imputed outputs of delay days have the same pattern of varying by reported dates as the observed data, we generate the quantile plots of the numbers of delay days (figure 6a and b) and the plots of the averaged numbers of delay days in time series from March 1st to July 13th (figure 6c and d) in cases with observed data and cases with imputed data separately. Both the quantile plots, which show the error bars with the values of these two kinds of delay days in quantiles of 0.025, 0.5 and 0.975 based on each date, and the plots of averaged delay days have similar patterns between the observed and imputed ones in the delay days from dates of first specimen collection or dates of reporting. We can see that the imputed results based on dates of first specimen collection are closer to their corresponding observed data compared with that of the imputed results based on dates of reporting, which further proves that date of first specimen collection is a better predictor than date of reporting as it has less variation and a more obvious pattern along with the reported time.

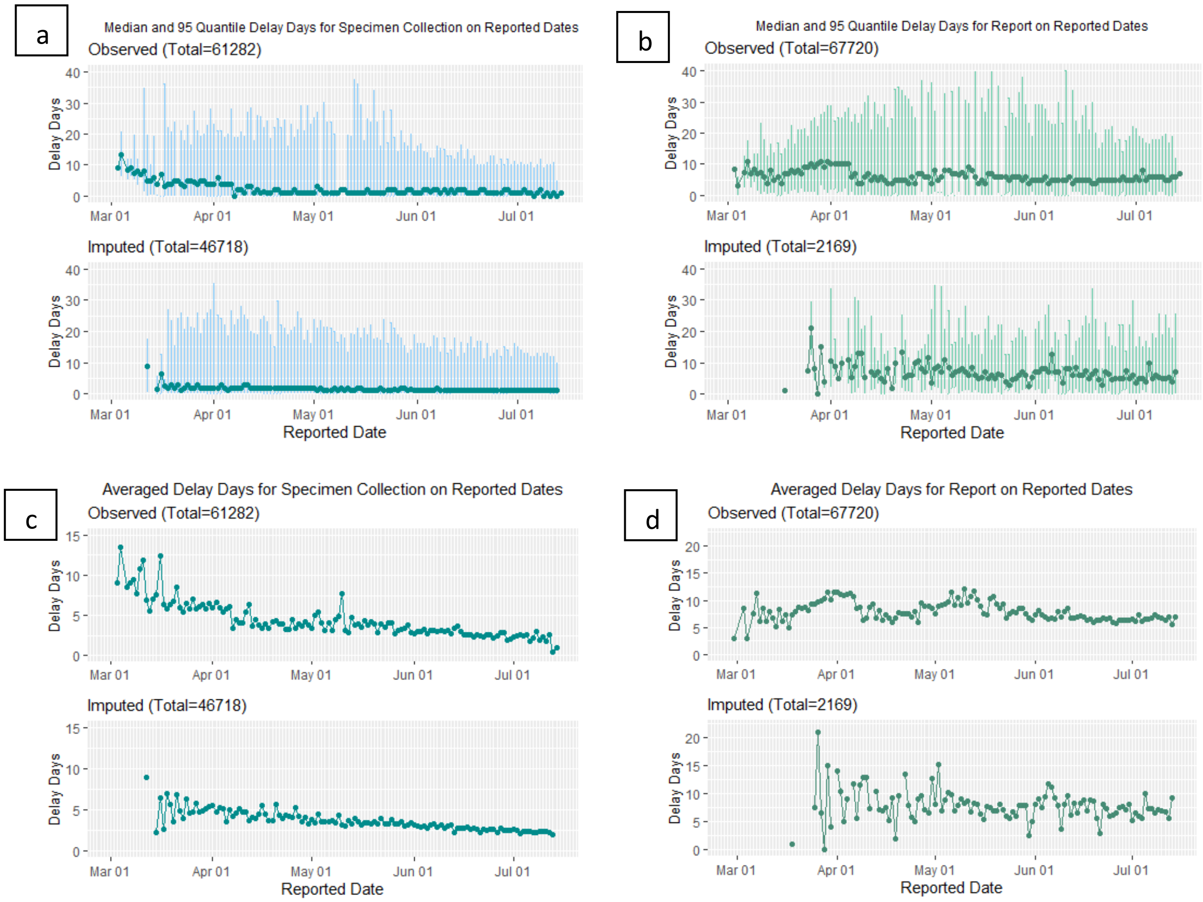

Figure 6a: the delay days from symptom onset date to date of first specimen collection in 0.025, 0.5, 0.975 quantile based on reported dates in observed and imputed data from February 28th, 2020 to July 13th, 2020. Figure 6b: the delay days from symptom onset date to date of reporting in 0.025, 0.5, 0.975 quantile based on reported dates in observed and imputed data from February 28th, 2020 to July 13th, 2020. Figure 6c: the averaged delay days from symptom onset date to date of first specimen collection based on reported dates in observed and imputed data from February 28th, 2020 to July 13th, 2020. Figure 6d: the averaged delay days from symptom onset date to date of reporting based on reported dates in observed and imputed data from February 28th, 2020 to July 13th, 2020. The plots did not include data before February 28th, 2020 as the cases were extremely scarce during this time.

Name: Holin Chen

Reference

1. Teunis P, Heijne JC, Sukhrie F, van Eijkeren J, Koopmans M, Kretzschmar M. Infectious disease transmission as a forensic problem: who infected whom? Journal of the Royal Society Interface 2013;**10**(81):20120955.

2. Wang Y, Teunis P. Strongly heterogeneous transmission of covid-19 in mainland china: Local and regional variation. Frontiers in Medicine 2020;**7**. doi:10.3389/fmed.2020.00329. URL http://dx.doi.org/10.3389/fmed.2020.00329.