Analysis of Daily COVID-19 Cases, Hospitalizations, and Deaths in NYC

Anja S, Bin Y, Safiya S, Ting-Hsuan C, Yijin W

2023-04-05

Introduction

Publicly available data for COVID-19 collected by governments, healthcare facilities, and academic institutions has facilitated research on the evolution of the pandemic. In this study, we analyzed COVID-19 trends in each of the five boroughs of New York City (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) using data collected by the New York Department of Health (NYDOH). The NYDOH dataset includes city-wide and borough-specific daily counts of COVID-19 cases, hospitalizations, and deaths since February 29, 2020 (the day of the first laboratory-confirmed case in New York City). A subset of this data up to December 11, 2020 was included in our analyses.

Figure 1 shows the citywide daily counts of COVID-19 cases, hospitalizations, and deaths from February 29 to December 11, 2020. We defined two pandemic waves in this time frame based on observations of daily cases: the first wave starts from February 29, 2020 and ends on September 14, 2020, in which a spike in cases was seen around April and dropped significantly until mid-September; the second wave starts on September 15 and is characterized by a resurgence of cases as winter approached.

The Richard's growth function is a S-shaped function that is commonly used to model the growth of a population. Let $N(t)$ be a population at time t, the function takes the form:

$$N(t) = \frac{a}{\{1 + d\exp\{-k(t - t_0)\}\}^{1/d}}$$

where (a, k, d, t_0) are shape parameters. The parameter a is the upper bound of the function; k is the growth rate, which controls the slope at an inflection (where the curve changes from convex to concave); t_0 is the time at an inflection; and d is another shape parameter that has no clear substantive meaning.

Let t_i be the number of days since the beginning of a pandemic wave. Let $(y_i, t_i)_{i=1,\ldots,n}$ be a sequence of observed daily cases/hospitalizations/deaths at time t_i. Let $Y_i = \sum_{k=1}^{i} y_k$ be the cumulative number of cases/hospitalizations/deaths by time t_i. We assumed that (Y_i, t_i) follows the following non-linear model:

$$Y_i = N(t_i, \boldsymbol{\theta}) + \epsilon_i$$

where $N(t_i, \boldsymbol{\theta})$ is the Richard's growth function with parameters $\boldsymbol{\theta} = (a, k, d, t_0)^T$ and ϵ_i is the random error with mean zero, i.e. $E[\epsilon_i] = 0$.

Our analysis was composed of three tasks. Task 1 developed a Newton-Raphson algorithm to fit a Richard's curve to each borough's cumulative cases. Task 2 then applied the same algorithm to each borough's cumulative hospitalizations and deaths. The fitted curves were compared across the five boroughs as well as the two pandemic waves. Lastly, Task 3 aimed to predict the trends (in cases, hospitalizations, and deaths) after December 11, 2020 for each borough in order to provide suggestions regarding the distribution of vaccination, which was authorized by the FDA on December 11.

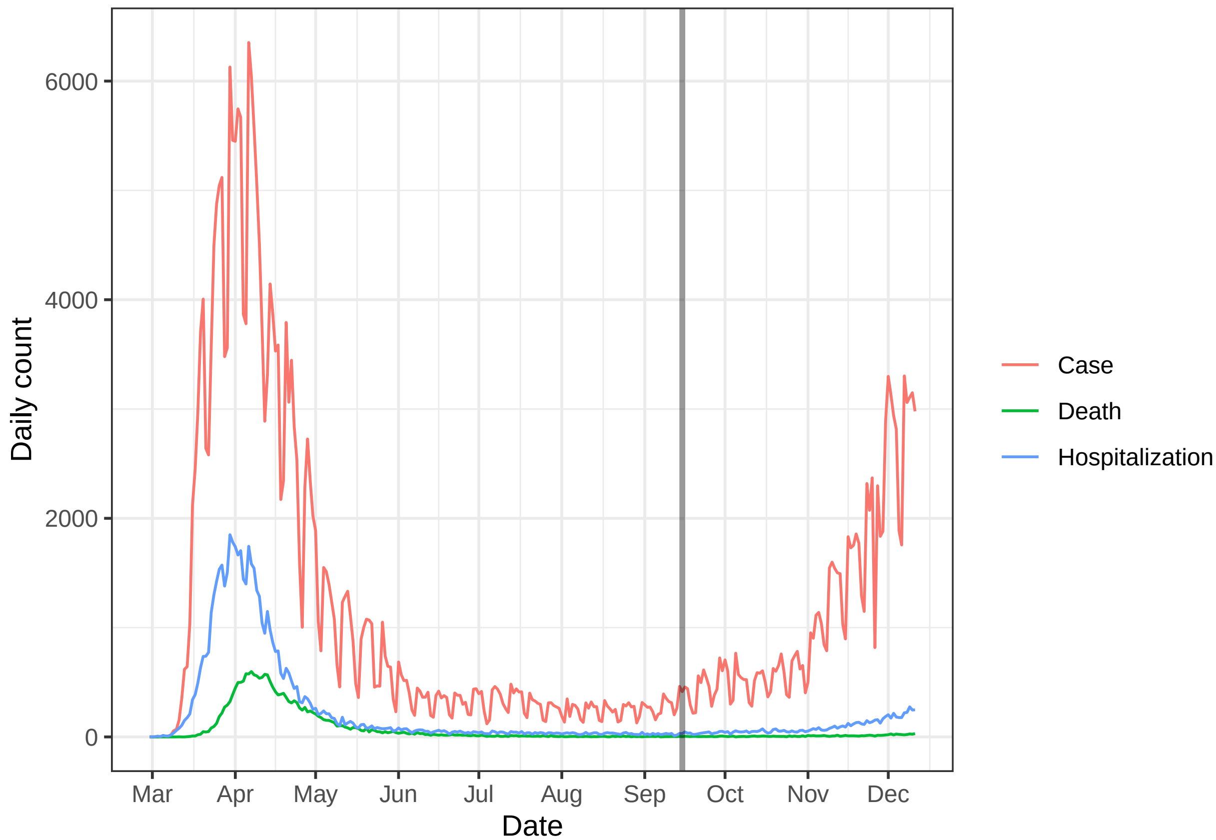


Figure 1: Daily counts of COVID-19 cases, hospitalizations, and deaths in New York City

Task 1

Our goal for Task 1.1 was to develop a Newton-Raphson algorithm to fit Richard curves to each NYC borough's cumulative number of cases for a pandemic wave. We used a simple gradient descent algorithm that aims to minimize the sum of squared errors (SSE), $h(\boldsymbol{\theta})$.

Since $E[Y_i] = N(t_i, \boldsymbol{\theta})$, then $h(\boldsymbol{\theta})$ is defined as:

$$\begin{aligned} h(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (Y_i - N(t_i, \boldsymbol{\theta}))^2 \\ &= \sum_{i=1}^{n} \left[Y_i - a\left\{1 + d \exp\left\{-k(t_i - t_0)\right\}\right\}^{-1/d} \right]^2 \end{aligned}$$

The gradient of $h(\boldsymbol{\theta})$, $\nabla h(\boldsymbol{\theta})$, is calculated as:

$$\begin{aligned} \nabla h(\boldsymbol{\theta}) &= -2 \sum_{i=1}^{n} [Y_i - N(t_i, \boldsymbol{\theta})] \cdot \nabla N(t_i, \boldsymbol{\theta}) \\ &= -2 \sum_{i=1}^{n} [Y_i - N(t_i, \boldsymbol{\theta})] \cdot \left(\frac{\partial N(t_i, \boldsymbol{\theta})}{\partial a}, \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial k}, \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial d}, \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial t_0} \right)^T \end{aligned}$$

where

$$\begin{aligned} \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial a} &= \frac{1}{(1 + de^{-k(t-t_0)})^{1/d}} \\ \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial k} &= \frac{a(t-t_0)e^{-k(t-t_0)}}{(1 + de^{-k(t-t_0)})^{1+1/d}} \\ \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial d} &= -\frac{a(e^{-k(t-t_0)}d - \ln(1 + e^{-k(t-t_0)}d)(1 + e^{-k(t-t_0)}d))}{d^2(1 + de^{-k(t-t_0)})^{1+1/d}} \\ \frac{\partial N(t_i, \boldsymbol{\theta})}{\partial t_0} &= -\frac{ake^{-k(t-t_0)}}{(1 + de^{-k(t-t_0)})^{1+1/d}} \end{aligned}$$

Our gradient descent algorithm can be summarized as follows:

1. Set $\boldsymbol{\theta}_0$, the starting values for $\boldsymbol{\theta}$ (see later in this section for more details on how to pick these).
2. Update $\boldsymbol{\theta}$ based on $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1} - \lambda \cdot I_{4\times4} \nabla h(\boldsymbol{\theta}_{j-1})$, where λ is a user-chosen learning rate (setting λ is to a small value, such as 1^{-10}, works well for this data).
3. If $h(\boldsymbol{\theta}_j) \geq h(\boldsymbol{\theta}_{j-1})$, then decrease the learning rate further and recalculate $h(\boldsymbol{\theta}_j)$, replacing λ with $\frac{\lambda}{10}$. Continue repeating this step until $h(\boldsymbol{\theta}_j) < h(\boldsymbol{\theta}_{j-1})$, i.e. until there is a decrease in SSE from the previous iteration.
4. Continue repeating steps 3 and 4 until convergence is reached, i.e. when the absolute difference between $h(\boldsymbol{\theta}_j)$ and $h(\boldsymbol{\theta}_{j-1})$ is smaller than a very small tolerance level.

Although our algorithm is not the most efficient algorithm, it has two main advantages. One, it is simple to compute and does not rely on the calculation of the Hessian, which is complicated. Two, using the symmetric and positive definite matrix $I_{4\times4}$ as a replacement for the Hessian guarantees that we have a descent direction. This means that we will be able to find some $\lambda \in (0, 1)$ that ensures that the updated $\boldsymbol{\theta}_j$ has a smaller SSE than the previous iteration's $\boldsymbol{\theta}_{j-1}$.

In step 1 of the algorithm, the user is required to choose the starting values, $\boldsymbol{\theta}_0$. We found that poor choices of $\boldsymbol{\theta}_0$ resulted in non-convergence or incorrect convergence issues. Choices of $\boldsymbol{\theta}_0$ based on the observed cumulative cases for the borough and the epidemiological interpretations of the Richard growth parameters resulted in fast convergence and good final estimates of Y_i. Our guidelines for how to tailor the starting values for each borough and wave are as follows:

- Since a is the upper bound of the Richard growth function, then let its starting value be the maximum number of cumulative cases in the pandemic wave. For the first wave, this is easily calculated as $\max(Y_i)$. For the second wave, since we only have observed data for the first half of the wave (i.e. up until about the inflection point), we must estimate what the maximum number of cumulative cases will be. We recommend using about $2 \times \max(Y_i)$ under the assumption that half of the maximum cumulative cases occurs at the inflection point. We made this assumption because we saw this trend across the boroughs in the first wave and assumed that the second wave would follow a similar pattern.
- Since t_0 is the time of inflection, then let its starting value be the time t_i where the inflection point occurs. This can be chosen as the point where the curve goes from convex to concave based on a plot of Y_i against t_i. Note that for the second wave, since we only have observed data for the first half of the wave, this point will occur towards the end of the observed data.
- Since k is the growth rate that controls the slope at the inflection point, then let its starting value be calculated as the slope at the inflection point t_0 standardized by the cumulative cases at the inflection point: $\frac{(Y_{i,t_0+m} - Y_{i,t_0-m})/[(t_0+m)-(t_0-m)]}{Y_{i,t_0}}$, where Y_{i,t_0-m} and Y_{i,t_0+m} are the cumulative cases corresponding to m days before and after the inflection point, respectively, Y_{i,t_0} is the cumulative cases corresponding to the inflection point, and $(t_0 + m) - (t_0 - m)$ is a small range around the inflection point. This can easily be computed by using two points around the inflection point to fit a linear regression, regressing Y_i on t_i.
- Since d has no epidemiological interpretation, picking its starting value is tricky. We found that values between 0 and 0.5 worked well. However, we advise investigators to try a few different values for d and pick what works best.

Figure XXX gives an example of how a starting curve for Wave 1 data on NYC would look, compared to the final fitted curve. We can see that the starting curve follows the general trend of the observed cumulative cases, but the final curve obtained from running our algorithm fits the observed data much better. We also note that from about $t_i = 100$ to $t_i = 200$ of the first wave, the observed cumulative cases follow a linear trend, so we would expect that our final Richard growth curve will never be able to fit the observed data very well in this portion of the wave since the Richard growth function assumes a sigmoidal shape.

The algorithm described in this section was also used in Task 1.2 to fit Richard curves to each NYC borough's cumulative number of hospitalizations and deaths for each pandemic wave.

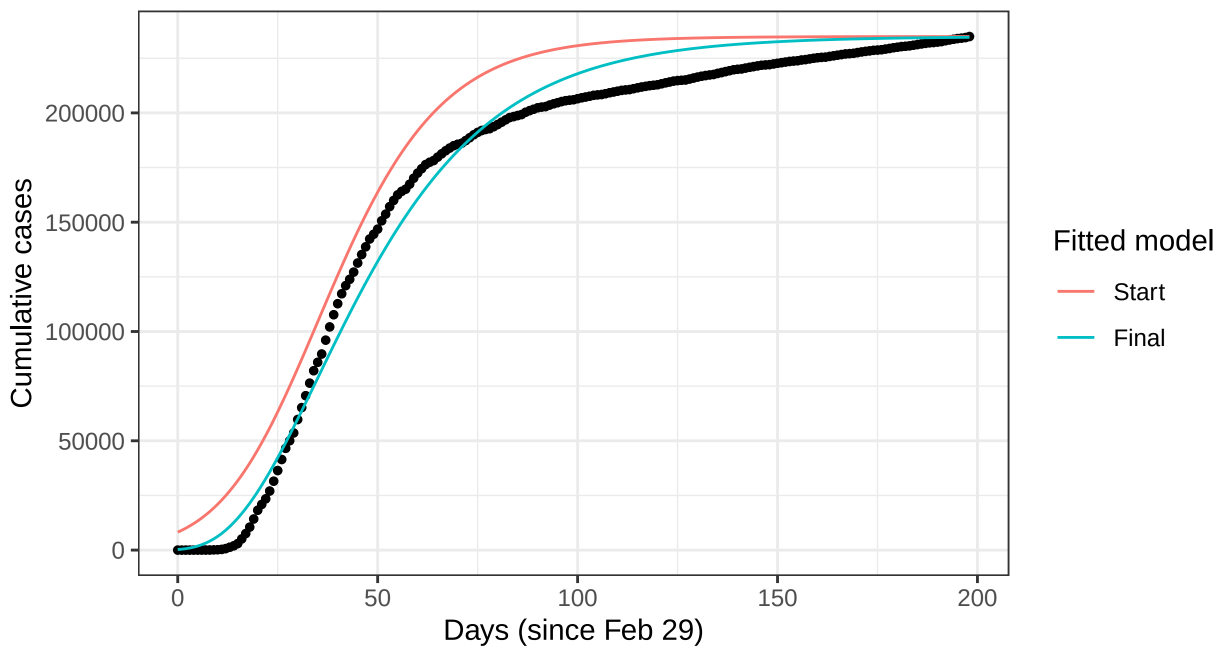
Task 2

Cumulative Hospitalizations

We fitted the Richard's growth curve to cumulative hospitalizations in all boroughs in NYC. Simmilar to cumulative cases, we split hospitalizations into two waves: 1) February 29, 2020 to September 14, 2020; 2) September 14, 2020 to December 11, 2020.

When fitting the first wave of hospitalizations, we used similar assumptions to pick starting values:

- a is $\max(Hospitalizations)$ for all boroughs and NYC, assuming there is not much increase in hospitalization resources.



	a	k	d	t0
start	234933	0.062	0.5	35
final	234933	0.04	−0.169	34.976

Figure 2: Example for Wave 1 NYC data comparing starting values and starting curve to the final values and fitted curve obtained from our algorithm.

- Different inflection date t_0 for different locations, as we observed slopes at different dates seem to fit the overall wave better. There are minor differences among inflection dates. Brooklyn and Manhattan have March 31 as starting value for inflection date; Staten Island has April 1; Queens, Bronx, and NYC have April 3, 2020.
- k is the standardized slope for cumulative hospitalization data on 4 days before and after inflection date.
- d is 0.5

Figure ? are plots for fitted curve for the first hospitalization wave. Figure ? are the parameters used to fit the first wave.

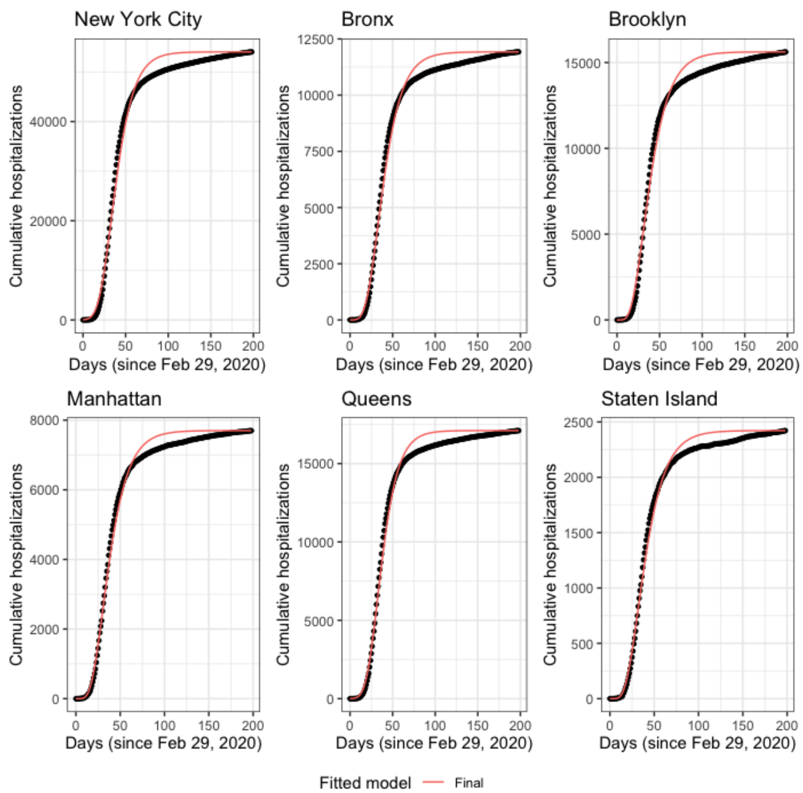


Figure 3: The first hospitalization wave - fitted curves

a	k	d	t0	location
54041.997	0.0708072	0.0997570	33.59340	New York City
15621.999	0.0593658	-0.1582541	31.05414	Brooklyn
11923.966	0.0648936	-0.0554206	32.94398	Bronx
7699.978	0.0634434	-0.1427894	30.67906	Manhattan
17104.972	0.0771895	0.0816411	32.68012	Queens
2421.995	0.0596605	-0.1477803	32.04848	Staten Island

Figure 4: The first hospitalization wave - parameters

The fit is very close to the observations in all locations. The fitted curve is overestimating hospitalizatons around the dates that the curve starts to level out. However, since healthcare resources were extremely limited during the outbreak of the pandemic, it is safe to overestimate than underestimate hospitalization counts. The final fitted t_0 for all locations are indeed a couple of days apart, corresponding to our initial belief.

Observed hospitalization in the second wave seems to be the first half of a regular Richard's growth curve. We do see a gradual increase and then more rapid increase in hospitalization during this period of time. The changes in slopes are tricky to model this time. For the second wave of hospitalization, the initial values are decided as the following:

- a is $2.2 \times \max(hospitalization)$ for all boroughs and NYC, assuming that half of the maximum hospitalization occurs at the inflection point.
- December 9, 2020 as t_0 for all locations.
- k is the standardized slope for cumulative hospitalization data on 2 days before and after inflection date due to rapid change in hospitalizations around our assumed inflection date.
- d is 0.25 for Staten Island and 0.5 for all other locations.

Figure ? are plots for fitted curve for the second hospitalization wave. Figure ? are the parameters used to fit the second wave.

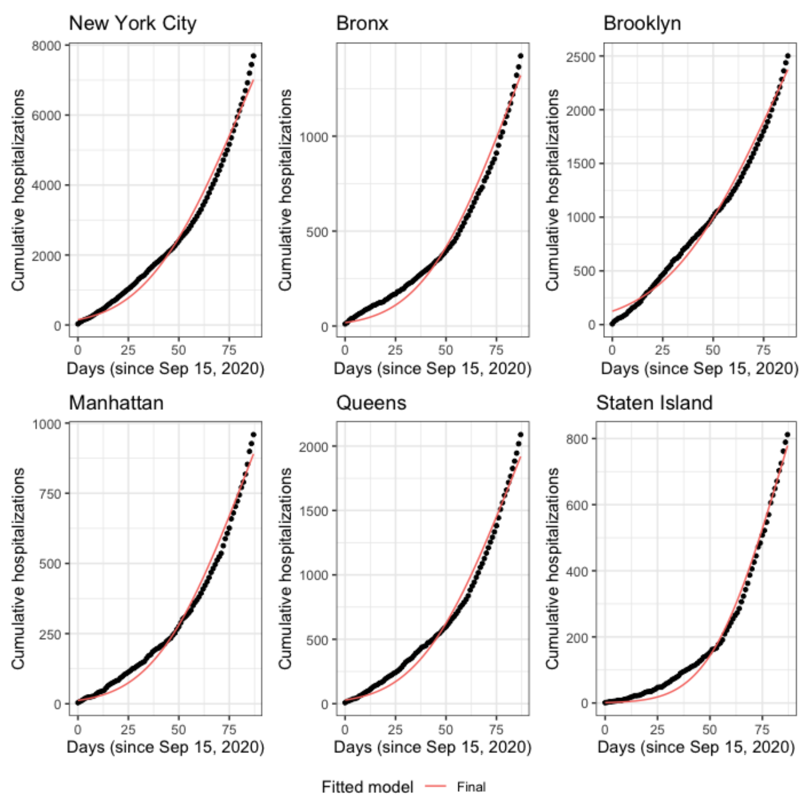


Figure 5: The second hospitalization wave - fitted curves

The fitted curves stil look very close to the observation. We tend to underestimate a little bit when the observed hospitalization is mildly increasing, and overestimate when the speed of the observed increase is growing.

a	k	d	t0	location
16922.400	0.0236201	0.1807922	85.03988	New York City
5502.200	0.0229339	0.2981760	85.01396	Brooklyn
3130.600	0.0265017	0.2164909	85.03464	Bronx
2109.799	0.0265173	0.2180125	85.03329	Manhattan
4595.800	0.0254216	0.1943075	85.05150	Queens
1786.400	0.0365780	0.2671106	85.00761	Staten Island

Figure 6: The second hospitalization wave - parameters

Cumulative Deaths

Task 3

Cumulative Cases

Cumulative Hospitalizations

We used the parameters in the fit for second wave of cumulative hospitalizations to make long-term predictions and hope to draw insight on vaccine distribution from the predicted curve.

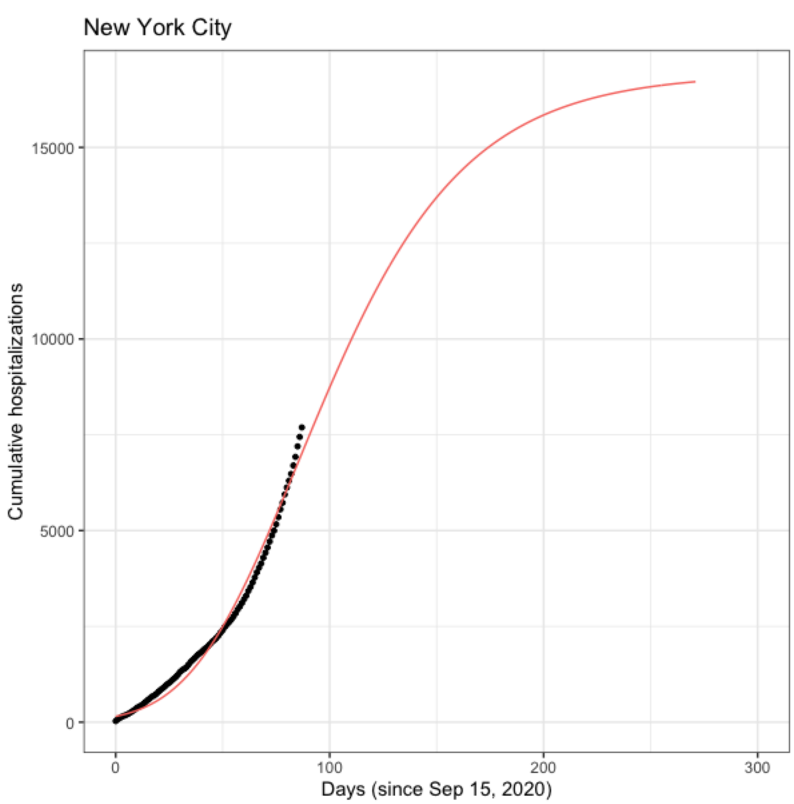


Figure 7: Predicted hospitalizations for NYC

To make our predictions more comparable, we plotted predicted cumulative hospitalizations per 100,000 capita for each borough.

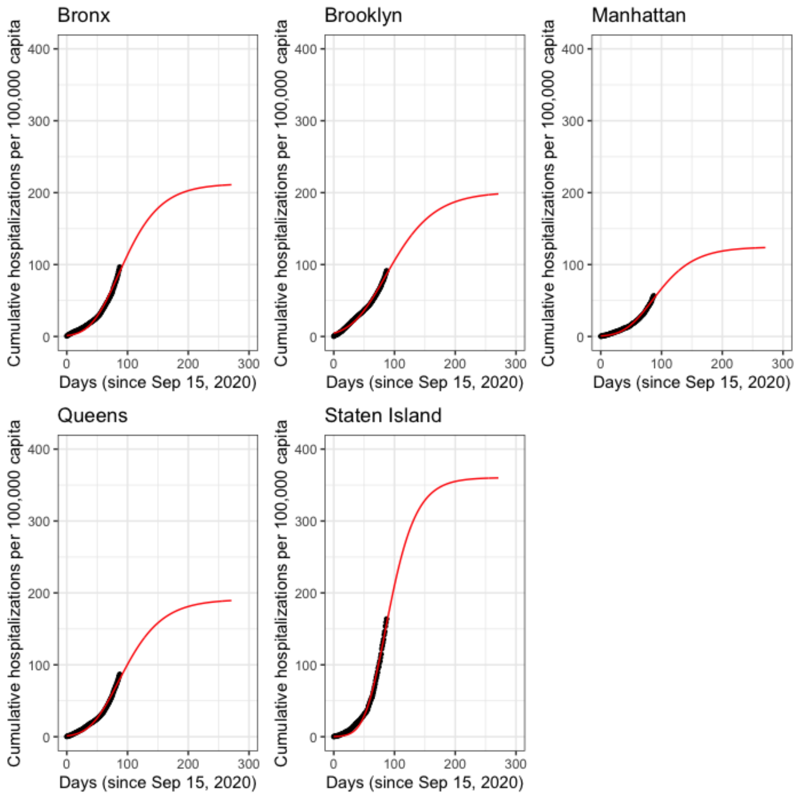


Figure 8: Predicted hospitalizations for each borough

Unfortunately, Staten Island has the highest predicted cumulative hospitalizations per 100,000 capita and its predicted growth rate is the highest.

Bronx and Manhattan both have the second highest predicted growth rate. But Bronx has the higher predicted hospitalizations per 100,000 capita.

Even though growth rate and predicted hospitalization per 100,000 capita are different for each borough, the predicted inflection date is almost the same for all boroughs. $t_0 = 85$ corresponds to December 9, 2020. It is also very similar to the predicted inflection date of cumulative cases. This could be a result of rapid disease progression, as people who were contracted need to be hospitalized immediately.

COVID-19 vaccines first became available in the US on December 14, 2020 according to U.S. Department of Health & Human Services. Therefore, while prioritizing hot spots such as Staten Island and Bronx, the government should also prepare enough inventory to meet needs city-wide to provide maximum protection against COVID-19 due to its rapid disease progression.

Cumulative Deaths

Conclusions