The Usage of Biased Smapling Causing Underestimation of Infected Population and Weak Policies

Tianxiao Ma

22 December, 2020

Abstract

Current infected population from COVID-19 could be substantially underestimated by the usage of reproduction number and confirmed cases as "reliable" estimates. The unstoppable trend of growing cases indictaes weak public policies were made as a result of using biased sampling as reference statistics. Based on Monte Carlo simulation model conducted in the report with uncertain reproduction number and people's willingness to get tested as random variables, the actual infected population might be 2 to 3 times more than confirmed cases in Ontario. Weak public policies were made due to the highly distorted data. Policymakers shouldn't be overconfident about the collected data due to the rapidly changing virus and non-representative data. Therefore, policy makers should consider introducing a free nationwide testing program and increase the effectiveness of contact tracing to get a more accurate COVID-19 data. Strict policies should also be implemented in the near future to flatten the curve.

Keywords: Biased sampling; COVID-19; Simulation; Reproduction number; Ontario; Monte Carlo simulation

Introduction

COVID-19 has attacked the whole world and threatened public health since Dec, 2020. Given the unique challenges that COVID-19 poses in the realm of being a pandemic that must be navigated with political and economic concerns, it has represented a breaking point for many social issues. Some of the most pivotal have been discussions of inequality, government efficacy, and racism, all catalyzed and exacerbated by this pandemic. No one perspective offers the whole story for this devastating disease, and while the virus poses a unique and troubling circumstance from a political point of view.

Governments around the world published policies such as large-scale lockdowns taking place across the country, and the extensive range of concerns from a medical side such as the average crowd density in locales, and the overall risk assessment of a country. The slogan "Flatten the curve" has been blended into society as a part of people's daily life. However, is the curve really representative? Is there something wrong with those policies? If people followed the public policies, what's the secret behind the rising confirmed cases? Does everyone follow government guidances? From an empirical perspective, the delay in data and biased sampling provided in front of governments as reference statistics misleads policymakers to build weak policies. As the butterfly effect shows, decisions people make have ripple effects beyond expectation – the slightest mishap on our part snowballing into an unprecedented catastrophe.

To minimize the overall loss caused by COVID, governments execute rational policies such as restrictions on gatherings. These public policies vary across countries, states and regions. Policy making highly relies on explicit theory and/or empirical evidence. Typical evidence that policymakers pay attention to is the COVID curve. COVID-19 spreads from person-to-person through droplet and contact transmission [WHO,2020]. One crucial step to analyze a disease is to investigate and record in detail how it is distributed in the population to see if the distribution follows a certain pattern.

At the beginning of the epidemic, COVID tests were only offered to the most symptomatic patients and front line workers due to the lack of testing resources. Bias occured if only testing this fraction of population, actual infected population is highly underestimated while overestimating the severity rate. The highly distorted data results in extremely unreliable analysis and corresponding predictions related to COVID. Policy makers usually do not have a scientific background, what they need to do when making a decision is to look through the abstract part of a report that a data analyst created, without questioning the data source of the report.

Many governors underestimated the pandemic at the beginning, delayed in pandemic response, and harmed the world as a whole as the prevalence growth is exponential in nature. Some policymakers rush to implement policies that are more flexible in order to boost the economy when the number of confirmed cases drops, without thinking about the accuracy of the source data and the exponential prevalence growth.

Most reports of confirmed cases rely on polymerase chain reaction–based testing of symptomatic patients[Spychalski et al.,2020]. Estimates of COVID cases based on PCR could give undercoverage biased sampling for missing asymptomatic patients and patients not been tested due to the scarce testing ability. In April, a community seroprevalence study in Los Angeles County shows the prevalence of antibodies to COVID was 4.65%[Sood et al., 2020], which is substantially greater than the 0.1% confirmed infections in the county on April 10 [Bendavid et al., 2020]. The observational study indicates the existence of information bias leads to a substantial underestimation of COVID infections. Besides, the study indicates using the reproduction number(r) and contact tracing methods to stop the spread is challenging.

Besides the problematic reproduction number, the curve is highly influenced by people's willingness to get tested. It's common to see people have different opinions and choices about one thing with various perspectives. There's a portion of individuals who refuse to get tested because of religion, financial, physical and other concerns. As we can see, the rising anti-science comments on social media and anti-mask protests in reality is a huge slap to medical workers and public policies. It's extremely hard to force some people to do COVID tests before a nationwide testing programs launches, even those people may have shown symptoms.

The report aims to show how distorted data underestimates the COVID severity by conducting two Monte Carlo simulation models under different scenarios. Report structure as followed: Data, the section contains empirical data reproduced from various public sources as evidence to support the model, and data needed to

build the model. Following model section contains the model settings, The first model aims to show the actual daily infected cases with consideration of Ontario residents' uncertain willingness to get tested using empirical data adapted from Public Health Ontario. The second model aims to show the estimated actual infected population results from one hypothetical outbreak with overstating assumptions under different settings of reproduction number.Result section includes resulting graphs and statistics derived from model section. Implications of the model results, weakness and future opportunities for future studies, and suggestions for policymakers are discussed in the discussion section. The report is built with R[R Core Team,2020], corresponding code and supplement dataset can be found at appendix 1.

Data

Dataset used for the report contains variables collected from various public sources. Most data are used as evidence that support assumptions made in the model, such as using such ranges of values of R0/ using exponential growth function for cumulative confirmed cases. (See Table 1 for more details)

Table 1: Variables used in the model/as evidence adapted from various sources

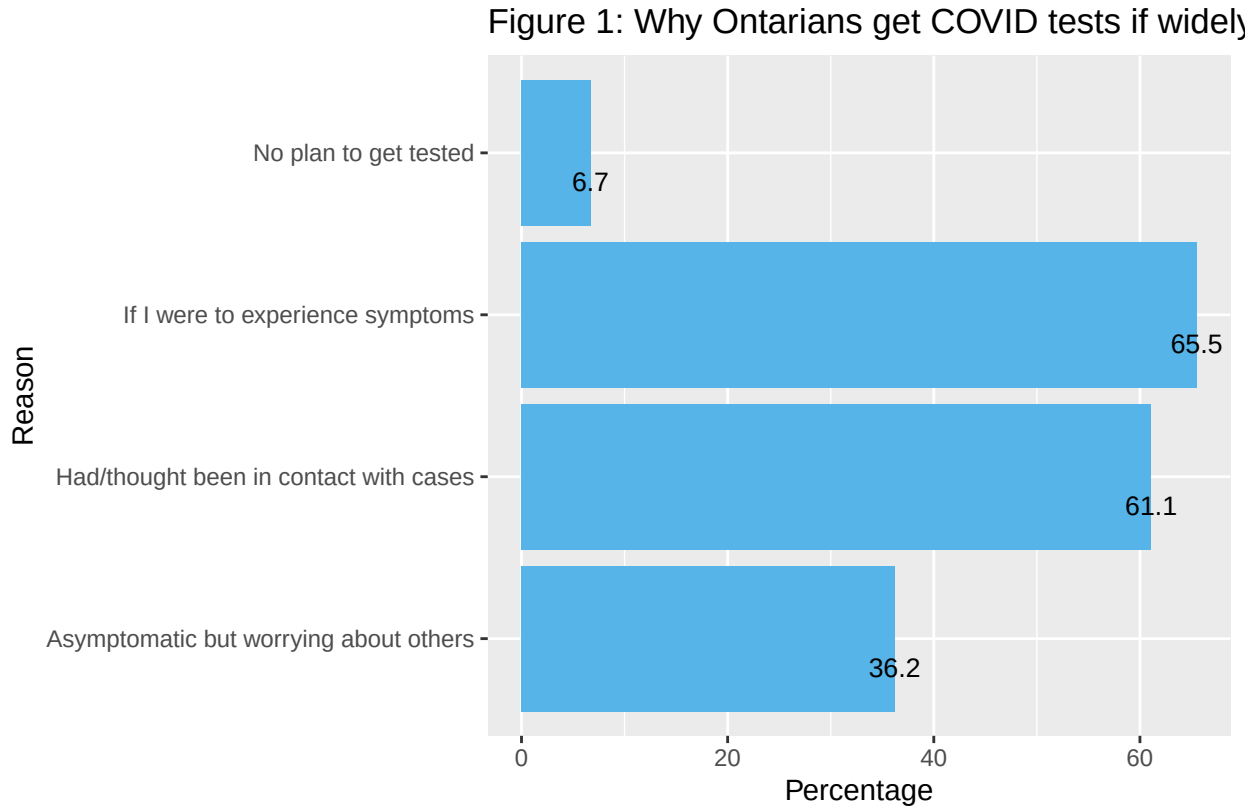
Variables	Description
date	Date that the case was reported to the local public health unit(PHU)
total cases	Cumulative COVID-19 cases
new cases	Daily reported cases released by Public Health Ontario
r	Reproduction number: an estimate of the average infections that a COVID patient results
location	location
C(t)	the cumulative number of confirmed cases at time t in days
C0	initial number of confirmed cases at the time when the count starts
w0	citizen's willingness to get tested(if symptoms developed)
s0	the percentage of symptomatic infected population, equals to 0.7 in the model

A critical factor influencing the accuracy of recorded confirmed cases is citizens' willingness to get tested. Results from the Canadian Perspectives Survey Series (CPSS) shows 65.5% (60.5% - 70.2%, with 95% CI) Ontario residents would get tested for the COVID-19 virus if they were to experience symptoms. That is, 45.85% of the infected population would get tested. 61.1% (56.0%- 65.9%, with 95% CI) Ontario residents would get tested if they had been, or thought they had been, in contact with people who had symptoms or who had tested positive for COVID-19. Just over one-third of Canadians reported that they would get tested if they were not experiencing symptoms but had concerns about infecting others (35.7%).Results from this study are based on the CPSS question: "If testing were widely available to all Canadians, why would you go to get tested for the COVID-19 virus?" Respondents were able to select more than one reason for getting tested for COVID-19. In the model section, responses based on the CPSS HR_Q05: "If testing were widely available to all Canadians, why would you go to get tested for the COVID-19 virus?" Respondents were able to select more than one reason for getting tested for COVID-19.

This study uses data from the 3rd wave of Statistics Canada's Canadian Perspectives Survey Series (CPSS), collected between June 15 and June 21, 2020. The CPSS is a panel of Canadians that has agreed to complete a number of short online surveys. 7242 respondents participated in the third wave of the CPSS, which focused on respondents' perspectives on resuming economic and social activities during COVID-19. The target population is all Canadians. The survey frame is a list of email addresses of Canadians who participated in the Labour Force Survey before. Non-responses and potential duplicate records are dropped. In the survey, CPSS used a weighted subsample of the Labour Force Survey sample to make the data representative of the general Canadian population. The LFS sample is drawn from an area frame and is based on a stratified, multi-stage design that uses probability sampling. Demographic variables such as age and living areas were drawn from the Labour Force Survey data then applied weights to make the population representative.

The target response rate for each survey in the series is approximately 60%, however, the sign-up rate of participation is only 23%. Response errors may occur may result in a systematic bias when the response provided differs from the real value. Additionally, coverage errors arise in the survey as the CPSS is completed online, those without email addresses or internet connections are not covered. People over 65 years old are the most affected.

Figure 1 shows Ontario residents' responses collected from Canadian Perspectives Survey Series 3. Responses are based on the CPSS HR_Q05: "If testing were widely available to all Canadians, why would you go to get tested for the COVID-19 virus?" Respondents were able to select more than one reason for getting tested for COVID-19. More than half of the respondents won't go to have a COVID-19 test even they developed some levels of symptoms.



Source: Statistics Canada, Canadian Perspectives Survey Series 3 (June 2020).

Another critical measure of infection rate is the reproduction number(r), the average number of secondary infections caused by an average infectious individual during the entire infectious period in a susceptible population[Vanessa Bates Ramirez, 2020]. Generally speaking, if r is greater than 1, each infected patient causes greater than one new infection, thus the growth rate of disease is growing.

The World Health Organization estimates the r for coronavirus to be between 1.4 – 2.5. That means every person who contracts the virus could potentially spread it to up to 2.5 more people. However, current methods do not correct for important selection and under-ascertainment biases.

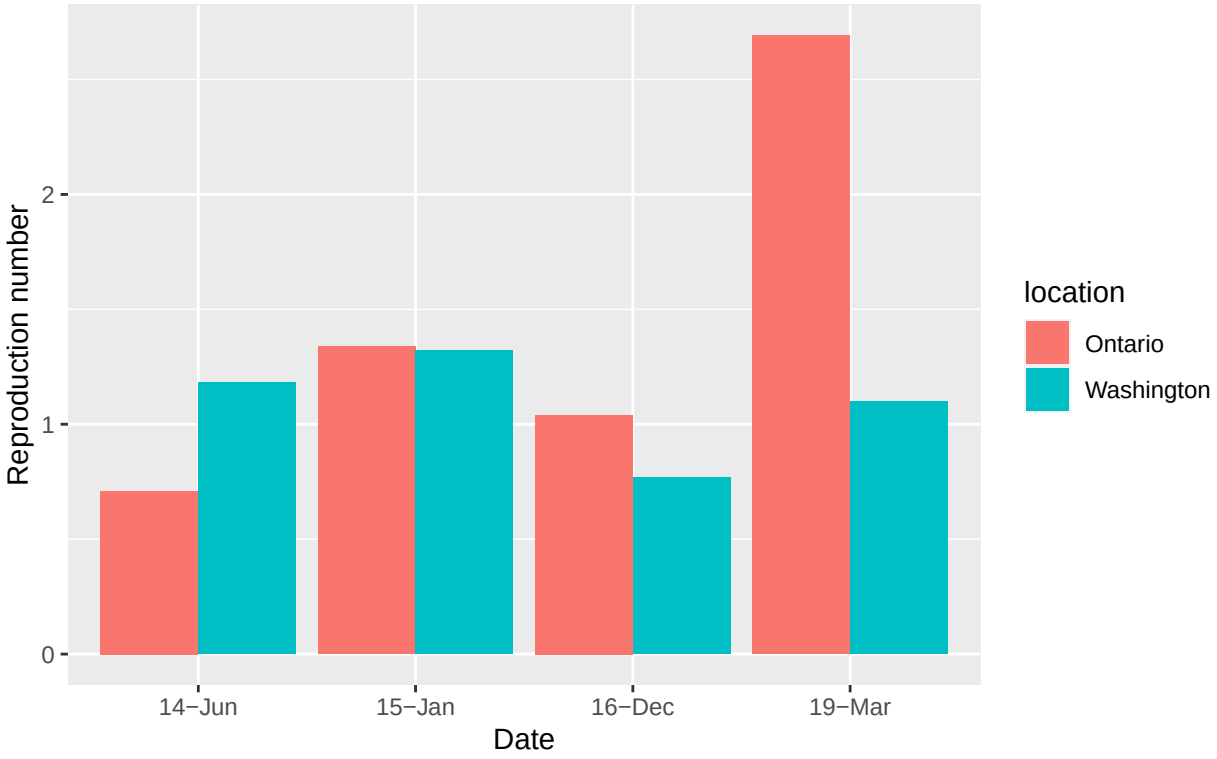
Figure 2 showed the varying reproduction number in Ontario, Canada [CCM, 2020] and Washington, U.S.[The COVID Tracking Project, 2020] at four time stages. However, the reported r is highly biased since the exponential growth indicates an infected person is able to infect more than 1 person. Reproduction number highly depends on other variables, such as time, country, strengths of government policy, even the person who estimates it. It's hard to identify reproduction numbers as an effective measure because it's based on biased available public data especially in North America. However, Asian countries have a relatively complete national Testing system, the R0 value used in the report is retrieved from a research conducted by NCCID, they used symptom onset dates of 41 cases from December 1, 2019 to January 1, 2020 and estimated R0 with a quantile range of 1.7 – 7.6 on December 15 in the initial stages of the Wuhan outbreak [Affan S. et al.,2020].

One important measure of testing rate among the population is the asymptomatic rate, According to the study conducted by South Korean researchers published in the Journal of the American Medical, asymptomatic carriers compose around 33.3% of the total infected population.[Joseph Workman, 2020]. In recent papers, Nishiura et al. [2020] estimated a 30.8% rate of asymptomatic COVID-19 cases. CDC gave an estimate of 25% asymptomatic rate. As a result, the hypothetical value for asymptomatic rate used in the paper is 30%, and corresponding symptomatic rate is 70%, with reference to current asymptomatic rate studies conducted by researchers across the world.

Figure 3 is a series of pie charts illustrates different asymptomatic rates estimated by different studies. The

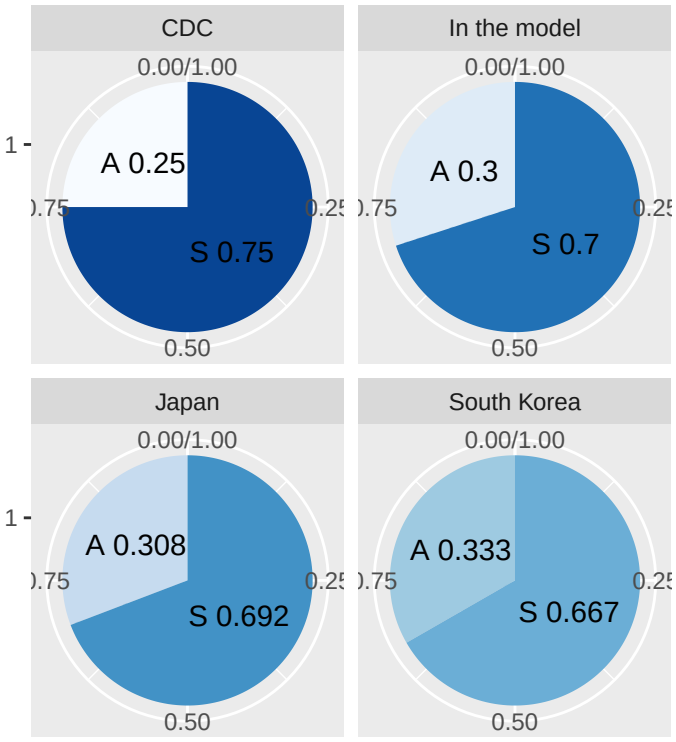
hypothetical value for asymptomatic rate used in the paper is 30%, and corresponding symptomatic rate is 70%, with reference to current asymptomatic rate studies conducted by researchers across the world.

Figure 2: How reproduction number varies from time and location



Source: Washington: Kevin et al.(2020), Ontario: CCM(2020)

Figure 3: Asymptomatic rates estimated from different studies



ce: South Korea: Joseph Workman(2020), Japan: Nishiura et al. (2020), CDC: CDC(2020)

The empirical data of cumulative cases in Canada and Italy is distributed from European Centre for Disease Prevention and Control(ECDC). ECDC's Epidemic Intelligence team has collected the number of COVID-19 cases and deaths on a daily basis, based on reports from health authorities worldwide. The data screening is followed by ECDC's standard epidemic intelligence process. Epidemic intelligence is a method for systematic collection and collation of information on communicable diseases of concern to the EU from a variety of sources. The aim is to ensure a timely response, based on an adequate risk assessment with recommendations on appropriate public health measures.

Figure 4 displays the total number of confirmed COVID-19 cases in Canada and Italy from January 26 until April 10,recorded every 3 days.Total confirmed cases in both countries following an exponential growth pattern. This is an evidence showing exponential growth function can be used to estimate cumulative cases.The variables used in the report are derived from the original dataset by setting the time range from January 26 to April 10, with focus on total cases in Canada and Italy. The corresponding csv can be find at Appendix 1.

Figure 4: Cumulative cases in Canada and Italy from Jan 26 to Apr 1

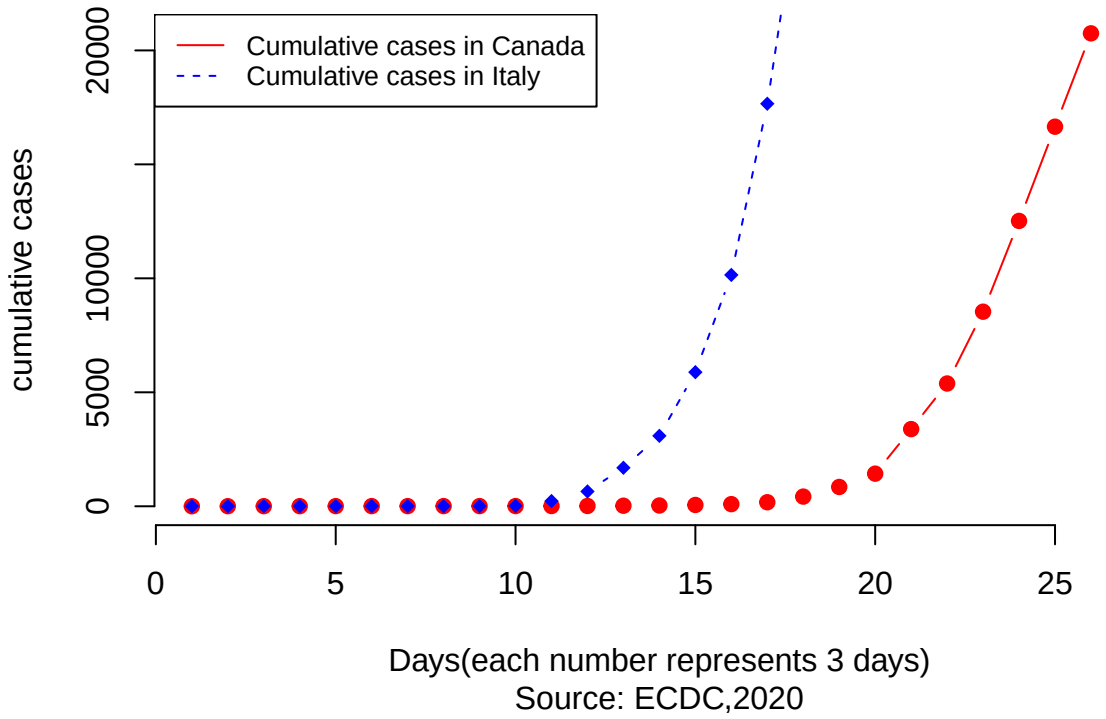
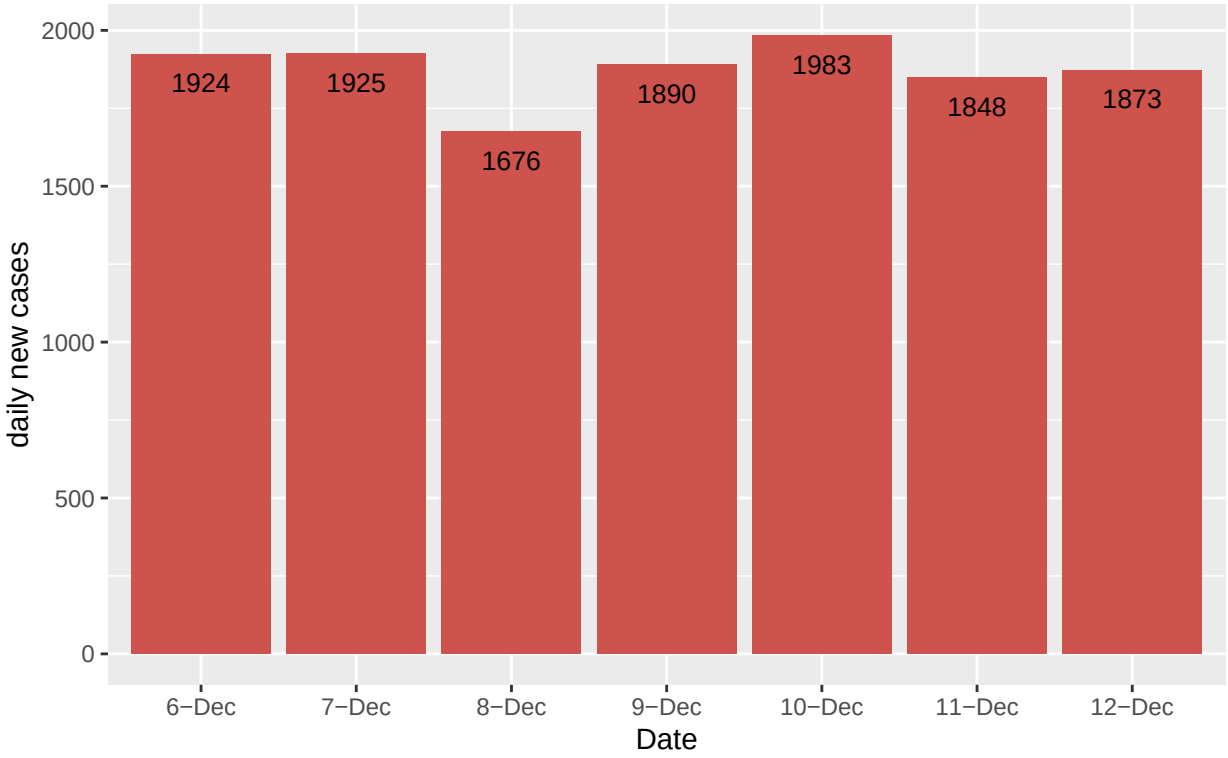


Figure 5 shows data of daily newly added cases in Ontario, it was adapted from Public Health Ontario Weekly epidemiologic summary: COVID-19 in Ontario – focus on December 6, 2020 to December 12, 2020. We can observe the daily confirmed cases are all above 1650, with minimum value, 1676, on Dec 08th, and the highest value 1983 on Dec 10th. for the weekly summary was based on the Ontario Ministry of Health (Ministry) integrated Public Health Information System (iPHIS) database for Toronto Public Health as of December 15, 2020 at 3 p.m. Information was uploaded through local systems and CCM plus (a dynamic disease reporting systems).

All of Ontario's Public Health Units report COVID-19 case data to CCM for their areas, with the exception of Toronto Public Health which enters data into its own Toronto Public Health Coronavirus Rapid Entry System (CORES). Case data in CORES is added to CCM by Public Health Ontario each day. Each case data The target population is all people who live in Ontario with a positive COVID-19 testing result. The data represents cases reported to public health units and recorded in CCM plus that meet laboratory standards. The frame is a list of infected people recorded in CCM plus. The survey involves key demographic information such as gender, age and living area. Each patient has to answer some questions related to contact tracing, potential transmission way, symptom severity etc.. Data are collected for all units of the target population, therefore no sampling is done. All the non-response or missing responses are excluded to be recorded.

The target population is all people who live in Ontario with a positive COVID-19 testing result. The data represents cases reported to public health units and recorded in CCM plus that meet laboratory standards. The frame is a list of infected people recorded in CCM plus. The survey involves key demographic information such as gender, age and living area. Each patient has to answer some questions related to contact tracing, potential transmission way, symptom severity etc.. Data are collected for all units of the target population, therefore no sampling is done. All the non-response or missing responses are excluded to be recorded.

Figure 5: Ontario daily new cases from Dec 06 to Dec 12, 2020



Source: Public Health Ontario(2020)

Model

Monte Carlo Model

The actual cumulative number of confirmed cases dynamics was characterised by a few key parameters in building the simulation model. The key parameters and assumptions were: (1) the growth factor parameter 'r': both reproduction number(r) and effective reproduction number (Re) used by Ontario Laboratories Information System are referred to 'r' in the model; (2) the actual number of people being infected by each of the infection active cases is determined by an exponential growth function (3)The initial number of infectious people, the observation period (in days),it could be infected by citizens' willingness to get tested once shows symptoms and the symptomatic rate. The purpose for the model is to understand the bias between actual infected population and current confirmed cases in the pandemic. As a result, the reproduction number, and people's willingness to get tested are two key variables that contribute to the bias. Thus, I derived a reproduction number(r) in the initial stages of the Wuhan outbreak on December 15 issued by NCCID in the data section, and used the quantile range 1.7 – 7.6 R0 in the model to create randomness. Similarly, the range used for each resident with symptoms is 60.5% - 70.2%[CPSS,2020].

At an early stage of the outbreak, new cases are directly proportional to the existing infected cases at the beginning of an infection chain, as a result, the cumulative cases grows exponentially. It can be expressed as an exponential growth function(Please see data section for more details of why using such a function)

Following is the exponential growth function:

$$C(r,t) = C_0 r^t$$

, C(r,t) represents the cumulative number of confirmed cases at time t in days with reproduction number r. Where r is the number of people infected by each sick person, the growth factor, or the reproduction number. C0 is the initial number of confirmed cases at the time when the count starts. The exponential model provides an upper bound of future situations by assuming the outbreak continues to grow and follows the same pattern in the past. However, the initial number of confirmed cases could be biased due to the citizens' willingness to get tested and asymptomatic patients. We introduce a new term, w0, to indicate citizen's willingness to get tested once they show symptoms. and s0 to indicate the ratio of symptomatic infected population. Assume only patients developed some levels of symptoms with willingness to get tested went to test, the cumulative cases follows the exponential growth model.

$$C(r,t)_r = (C_0/(w_0 * s_0))r^t$$

, C(r,t)_r represents the cumulative number of confirmed cases at time t, where r is the reproduction number, which can be interpreted as growth rate here. C0 is the initial number of confirmed cases at the time when the count starts, s0 indicates the percentage of symptomatic infected population, and w0 shows citizen's willingness to get tested once showing some levels of symptoms.

In the following parts, two models were developed under different scenarios. The first model aims to show the actual daily infected cases with consideration of Ontario residents' uncertain willingness to get tested using empirical data adapted from Public Health Ontario.
The second model aims to show the estimated actual infected population results from one hypothetical outbreak with overstates settings in 3 days with uncertain growth factor, r.

#1 Consider the actual infected cases with consideration of Ontario residents' willingness to get tested

Monte Carlo simulation processes involve the random number generation which means the simulation analysis results would be subject to random variation due to different starting points defined intrinsically by a selected random seed value[Casella et al., 2020]. Therefore, when we would compare the actual infected population on different dates using the simulation model, the same random seed needed to be specified for each daily newly added case for a valid comparison. To count for the uncertainty in a Monte Carlo simulation study, a simulation approach was followed to obtain the estimated median and quantile values of the number of the total infection active cases on the day of observation. Each simulation was run for 10000 times and the

median actual infected population was considered as the most likely estimation, and the uncertainty level was characterised by the values range from the 1st to 3rd quantiles. As Monte Carlo simulation used for model 1generates a random sample of size 10 for each uncertain input variable of a model, then repeat 10000 times. It selects each point independently from the probability distribution for that input variable.

Empirical data shows 1924,1925 and 1676 confirmed cases on Dec 06 in Ontario[Public Health Ontario,2020], where t=0, w0 = 65.5%[CPSS,2020] and s0=70% [Joseph Workman, 2020]. Assume only symptomatic infected population(70% of total infected population) who are willing to get tested(65.5% of the 70% total infected population) went to test: The resulting actual infected population would be C(r,t)_r = (1924/(0.655* 0.7)) C(r,t)_r = 4196.29226 More than 4196 actual active infected cases rather than 1924 reported newly added cases on Dec 6th. The actual active infected cases for Dec 7th and 8th are 4199 and 3655 accordingly. A Monte Carlo simulation model is used to create some randomness, so we may not get exactly 65.5% of symptomatic residents going to test every time. Rather than assuming exactly s = 65.5% of Ontario residents would get tested for the COVID-19 virus if they were to experience symptoms, each resident with symptoms has a 60.5% - 70.2% chance of getting tested. s follows a uniform distribution from 60.5% to 70.2% as a variable in the model. The chance is derived from the Canadian Perspectives Survey Series (Please see Data section for more details). At this time, the term rˆt in the model equals to 1 since t=0 when counting on a daily basis.The model is shown below:

$$C(s)_r = (C_0/(w * s))$$

,

Model 2: With overstate assumption: Resulting infection cases occur in just 1 day of confirmed cases.

Assume the growth rate is ranging from 1.7-7.6. The model aims to see the cumulative infected population under uncertain R0. Assume 1000 hypothetical social gatherings were held at 5:00 P.M on Dec 6th with 10 attendees in each gathering, The observation period was 3 days. The COVID-19 spread process started from one person who was able to infect 1.7-7.6 people. The number of people being infected was modelled by an exponential growth function C(r,t) = C_0rˆt with parameter r, total number of people being infected could be different in different time stages of the observation period with varying r.

the model becomes

$$C(r, t) = C_0 r^t$$

, cumulative number of confirmed cases at time t is directly related to the growth factor,r. Using the seed of my interest, I randomly assigned 10000 r values following an uniform distribution ranging from 1.7 to 7.6 for each day. I simulated 10000 times for each day with randomized r following an uniform distribution ranging from 1.7 to 7.6, to see the resulting infected population on that day. I expected the average r to converge to 4.65 as the sample size approaches to infinite. Then we can use the exponential function to compute the value of C(t) for each value of t from 1 to 3, the resulting values for each simulation are estimated actual infected population in each trial at that time stage.

Assume that a minimum of one day is needed for a newly infected person to start transmitting the COVID-19 to the next generation of infected people. The infection times are independent of each other so that more than one person could be infected from the same infection active person on the same day. The simulation model was implemented in R and the r code can be found through github.

Convergence Diagnostics For Monte Carlo and Alternative Models

As Monte Carlo simulation relies on randomness, it can be inefficient. Some points might cluster closely while other intervals may involve no samples if the sample size is not big enough. Monte Carlo simulation and Latin Hypercube sampling are both unbiased estimation techniques: computed statistics approach their

theoretical values as the sample size increases[Lonnie Chrisman, 2014]. MC converges to theoretical values as the sample size increases. Even though we can increase result efficiency by increasing sample size, it's hard to define the sampling error in the computed results at a given sample size. We can plot running means of variables to see the stationarity of the simulation and check statistical summaries of the output to diagnose convergence.[Martin Haugh, 2017] Latin Hypercube sampling is an alternative model to make sample points spread more evenly across all possible ranges.[McKay et al., 2010]. Using the LHS model, there's no correlation between input variables since it partitions each input distribution into N intervals with equal probability, and selects only one sample from each interval[Lonnie Chrisman, 2014]

Results

The Result of the simulation model included three pieces of information: From model 1: 1. Sampling distribution after 10000 trials. Figure 6 to 8 showed the sampling distribution for each day with consideration of Ontario residents' uncertain willingness to get tested on December 6th, 7th and 8th respectively. The Sample statistic, which is the sample mean is shown as a red line. All three histograms are normally distributed with thin tails. Based on the Central Limit Theorem, the sample mean was considered as the most likely estimation. The estimated infected population in Ontario is 4207,4210, and 3664.

2. Estimated number of actual infected cases vs. actual reported daily newly confirmed cases Figure 9 Compared the sample mean with the observed infected cases, data was adapted from Public Health Ontario.The actual infected population from each day is 2 to 3 times more than observed infected cases.

3. Estimated infected population resulting from a social gathering with one initial infected person in three days. Figure 10 showed the estimated infected population in day 3 could exceed 100, with initially 1 infected case in day 0, With a random growth factor following the uniform distribution from min 1.7 to max 7.6.

Figure 6: Estimated actual infected Ontario population on Dec 06th

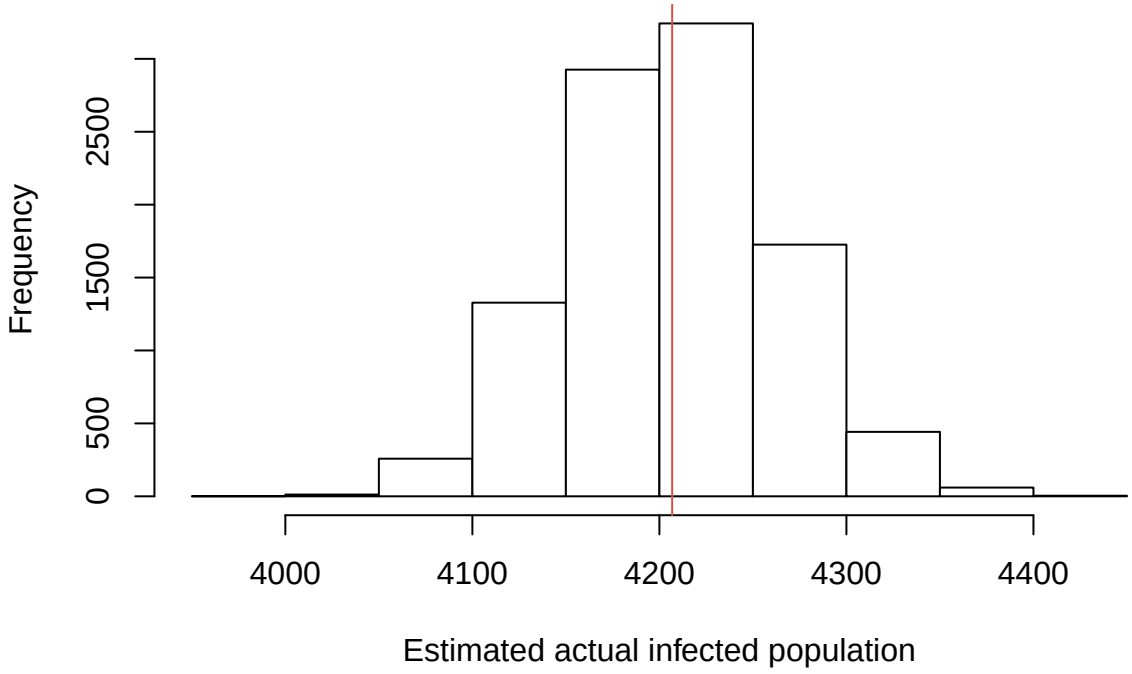


Figure 7:Estimated actual infected Ontario population on Dec 07th

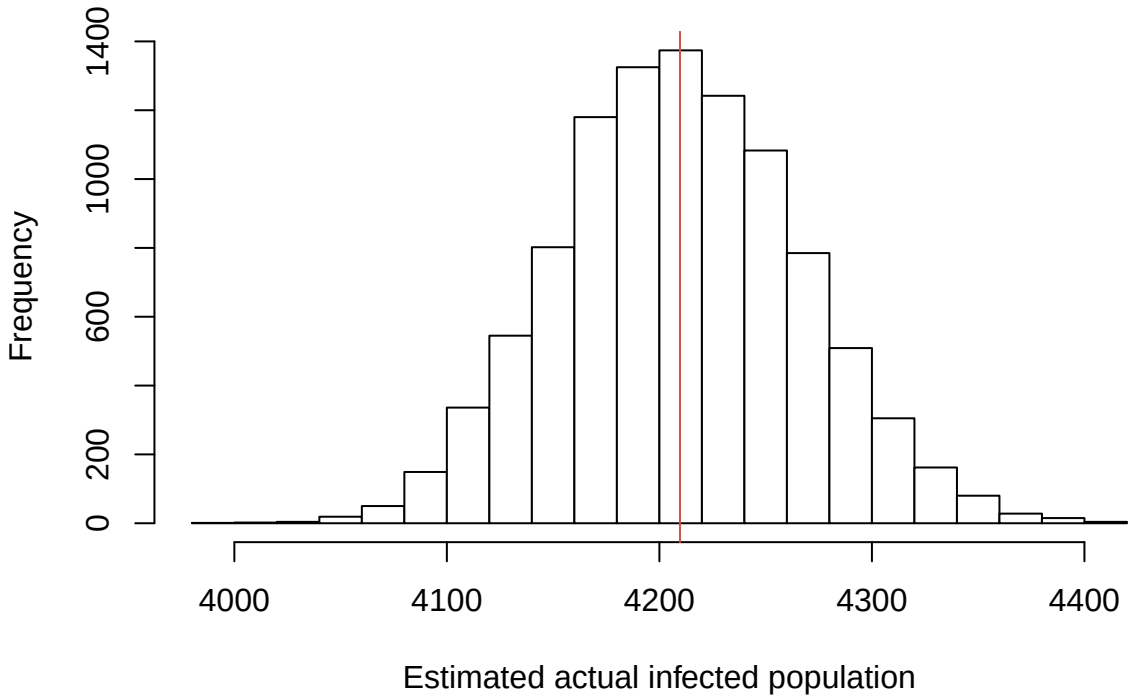


Figure 8: Estimated actual infected Ontario population on Dec 08th

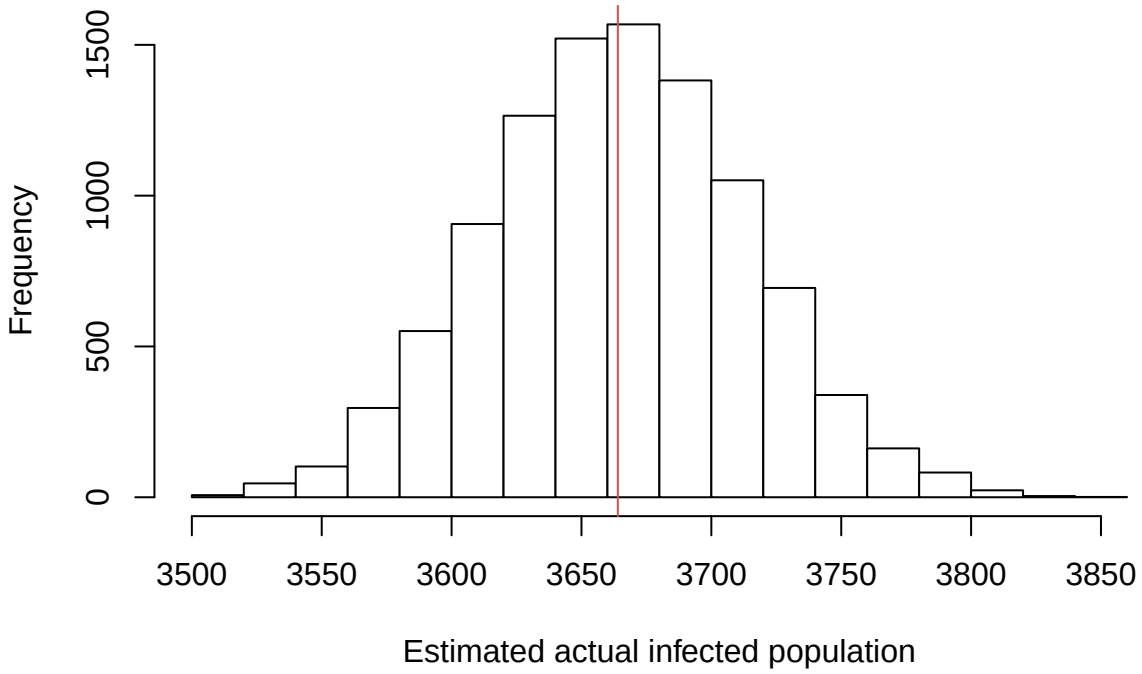
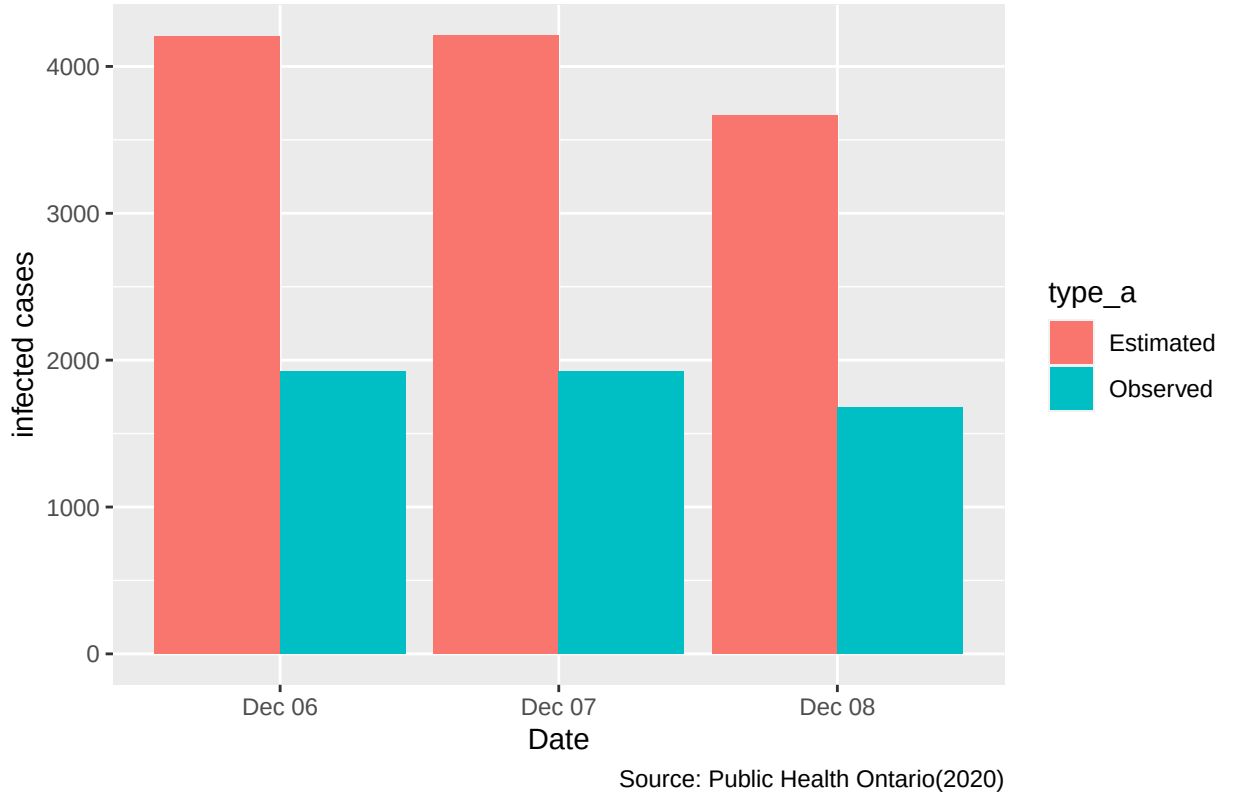
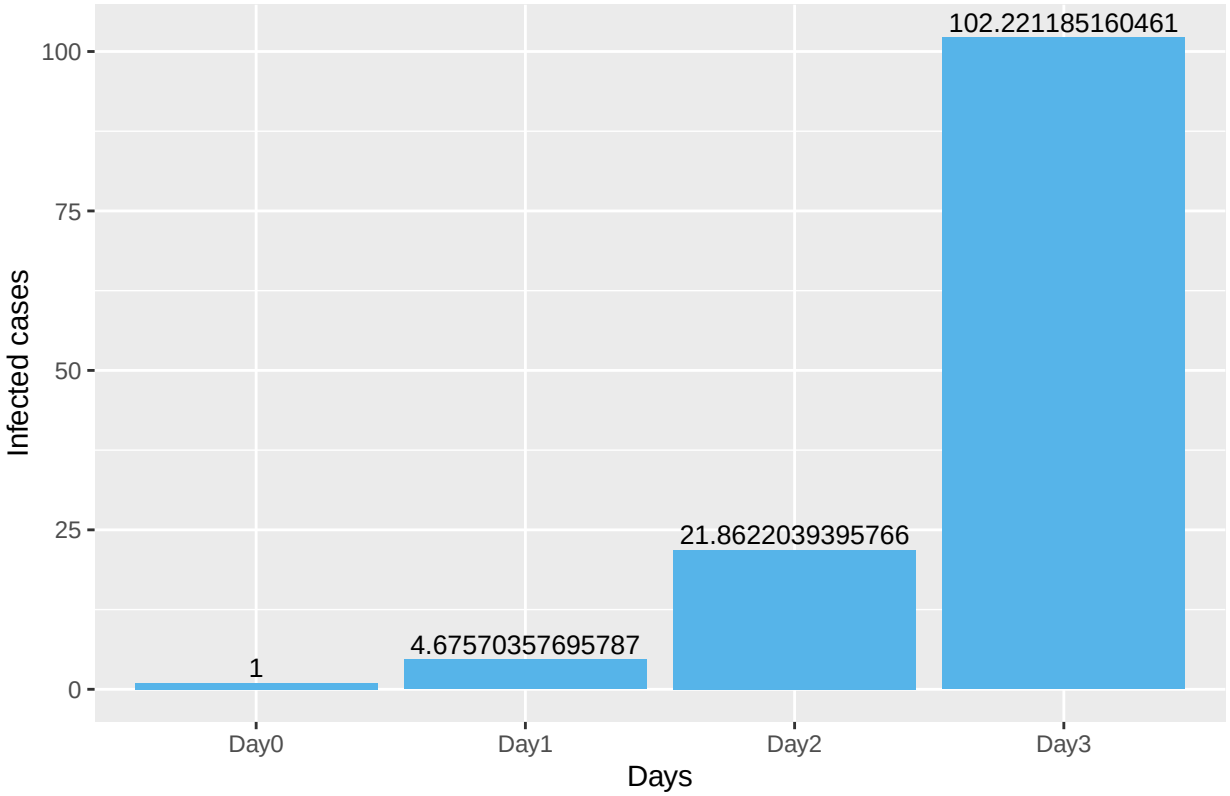


Figure 9: Compare observed confirmed cases and estimated actual infecte



Source: Public Health Ontario(2020)

Figure 10: Estimated infected population in three days



Discussion

Implication

Exponential proliferation of COVID, at least in the beginning phase, is undeniably dramatic. Any changes policymakers can make at this relatively early stage can flat the curve to a great content. The urgent need to understand the spread of COVID-19 across the world and the effective ways to control the curves was given to priority. One crucial step to analyze COVID is to investigate and record in detail how it is distributed in the population to see if the distribution follows a certain pattern. However, any graphs/statistics made based on the samples derived from observational studies and surveys could be highly unreliable due to biased sampling. The sampling designs used are often biased in that they do not reflect the true underlying populations[Daniel et al., 2020]. The pattern of increasing cases indicates weak prevention measures done by governments.

Both models assume the growth trend of cumulative infected cases growing exponentially pattern, but exponential Growth function may not necessarily be the perfect representation of the epidemic in the long run. Conditions in model 2 can be considered as strict and a little bit unrealistic compared to the observed cases in the real world. To optimal the estimation, the model applies to the beginning of an outbreak occurring in Ontario and nobody else in the community has had/is attacked by COVID, and only one day is needed for a newly infected person to start transmitting the COVID-19 to the next generation of infected people. The infection times are independent of each other so that more than one person could be infected from the same infection active person on the same day.

As the model result shows, the normally distributed estimated actual infected population of each time stage with small variance is due to large sample size, and the median actual infected population was considered as the most likely estimation in model 1. In model 2, once the virus begins to transmit in a community, it grows exponentially and could result in the infected population to double, triple, even hundreds of times larger in a very short time. According to the model 2, at the beginning of the infection chain, one infected person participating in a small social gathering, even a 5-people family dinner, might contribute to hundreds of infections in several days. Additionally, people with strong symptoms are more likely to be tested than those with no symptoms, and according to CPSS survey responses, only a portion of symptomatic patients would think about getting a COVID test. Sampling bias and voluntary response bias occurs when the actual infected population could be 2 and more times than current observed confirmed cases, those confirmed cases are non-representative.

The most effective way to fight with COVID is government nonpharmaceutical interventions before the vaccine is proven to be safe and effective. At the time of publication, the vaccine just issued out and remains uncertain about potential side-effects, and there's no widespread population immunity known. Current public policies can be viewed as "WEAK". Delays in response and overconfidence about the current situation distributes to low levels of national preparedness [Rabail et al., 2020] and fast loosening restrictions. Many data analysts use biased reproduction numbers as an estimate of the growth trend of COVID will result in unreliable predictions because of the limited resource they can use to make precise analysis. Nelly Yatich, an epidemiologist in Nairobi, Kenya said "It's extremely difficult at the beginning of an epidemic to get an accurate R0". However, the bias is understood by epidemiologists but not the policy makers. The highly distorted data results in extremely unreliable analysis and corresponding weak public policies related to COVID.When policymakers see statistics and predictions provided by data analysts, they should at least be concerned about the variables that data analysts used. They also should notice there are some citizens who didn't go to the test due to physical/mental/financial concerns, so the statistics may be not accurate at all.

Limitations and next steps

The model assumed the growth trend of cumulative infected cases growing exponentially pattern, but exponential Growth function may not necessarily be the perfect representation of the epidemic in the long run. It's only suitable when applied to an early stage of an outbreak, the growth will stop as healed people will

not spread the virus anymore. Future works may concern logistic growth instead. The model also ignores the chance that a non-negligible portion of asymptomatic individuals carries the virus to get COVID tests, and the limiting testing resource is only offered to the most symptomatic patients and front line workers, resulting in relatively large sampling variability. Future works may concern adding these considerations as variables. As the Monte Carlo (MC) simulation model purely relies on randomness, we can build a bayesian model later based on data provided by retrospective studies in the future to have a better understanding of the actual infected population behind COVID-19.

To reduce current sampling bias, one possible public policy is to introduce a free nationwide testing program. It could help a better understanding of the infection in a community, identify hidden patients as soon as possible to cut off the chain of transmission in the community, and eliminate the financial concern of citizens, and bias towards people who went to do a COVID. It is a viable opinion, but it will result in lower national savings and higher taxes in the future.

One suggestion for policy makers is to double the number of confirmed cases to get a vague sense of the actual infected population. As a result, powerful public policies made with consideration of actual infected populations can better control the spread of COVID before herd immunity(Please see Appendix 2 for herd immunity details).

Another opinion for policy makers to better control the spread of COVID is to increase the effectiveness of contact tracing. Contact tracing is the process of identifying, assessing, and managing people who have been exposed to a disease to prevent onward transmission[WHO, 2020]. Confirmed cases, asymptomatic infections, and clusters of epidemics are all targets of contact tracing. Due to COVID person-to-person transmission characteristics , contact tracing is a critical part of strategy to analyze and reduce the spread. If contact tracing is applied properly with corresponding effective regulations in an area, the spread of the virus would be stopped shortly. To conduct a more effective contact tracing to control the spread of COVID, government could expand the covering range in terms of scope, that is, trace close contacts and sub-close contacts of cases, and expand the scope of isolation and medical observation of relevant personnel; in terms of intensity, it is to expand the sampling size and monitor community activities, and increase COVID testing in high-risk areas and populations in Canada.

When a case is diagnosed, contact tracers have to contact the patient himself/herself, or the treating doctor, family members to inquiry, and access to information related to the patient's travelling history, clarify how the infection was transmitted, list close contacts, and confirm whether the virus has been transmitted to other people. However, current situations indicate convenient sampling occurs when contact tracers select patients to be traced. Contact tracers choose patients in their convenience for a limited level of effort, and only a small portion of confirmed cases are traced.

The expected contact tracing should involve the patient himself and all the people who have contacted the patient in the past 2 days. However, it's impossible to implement in reality for now since the majority of patients can't provide details of all the travel history. Safely assuming, Ontario contact tracers have reached their limit just to contact patients themselves, no need to say if the patients' close contacts will be traced. Interestingly, there're 6992 new reported cases in Canada on Dec 06, but only 2070 calls were made by Statistics Canada for contact tracing(See Appendix 3 for more details). The probability of contacting close contacts of confirmed cases is incredibly low at this time. What we expect for an effective contact tracing process is to contact all the patients' closed contacts. The current estimated contact tracing rate for close contacts would be around 15-30%. The tiny non-random sample is highly unlikely to be representative[Andrew Whitby, 2020].

References

Andrew Whitby (2020). Contact tracing can give a biased sample of COVID-19 cases. Retrieved from https://andrewwhitby.com/2020/11/24/contact-tracing-biased/

Affan S., Pratha S., Abhishek P., Chad R W., Yaning W., Zheng W., Burton H. Singer, Alison P. Galvani, Seyed M. Moghadas(2020). NCCID Special Post: Transmissibility of the Initial Cluster of COVID-19 Patients in Wuhan, China. Retrieved from: nccid.ca/publications/nccid-special-post-transmissibility-of-the-initial-cluster-of-covid-19-patients-in-wuhan-china/

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R. Bhattacharya, J. (2020). COVID-19 Antibody Seroprevalence in Santa Clara County, California. doi:10.1101/2020.04.14.20062463

Casella, G. & Berger, R. L.(2002) Statistical Inference 2nd edn. (Thompson Learning Inc., The Wadsworth Group, Stamford, 2002).

CDC(2020).Centers for Disease Control and Prevention Coronavirus Disease 2019 (COVID-19). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html#Asymptomatic

Daniel Andrés Díaz-Pachón & J Sunil Rao(2020). "A simple correction for covid-19 testing bias." ArXiv, arXiv:2007.07426v2.

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

European Centre for Disease Prevention and Control(2020).Retrieved from https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016, https://ggplot2.tidyverse.org

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr:A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

Joseph Workman (2020). The proportion of COVID-19 cases that are asymptomatic in South Korea: Comment on Nishiura et al.DOI:https://doi.org/10.1016/j.ijid.2020.05.037

Kevin Systrom, Thomas Vladeck, Mike Krieger(2020), Rt.live. Retrieved from: https://rt.live/us/WA

Lonnie Chrisman (2014). Latin Hypercube vs. Monte Carlo Sampling. Retrieved from https://analytica.com/latin-hypercube-vs-monte-carlo-sampling/

Martin Haugh(2017). MCMC and Bayesian Modeling. Retrieved from http://www.columbia.edu/~mh2078/MachineLearningORFE/MCMC_Bayes.pdf

M. D. Mckay, R. J. Beckman, W. J. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer CodeTechnometrics, Vol. 42, No. 1, Special 40th Anniversary Issue (Feb., 2000), pp. 55-61 Retrieved from : http://www.jstor.org/stable/1271432

Nishiura H, Kobayashi T, Suzuki A, Jung S-Mok, Hayashi K, Kinoshita R, Yang Y, Yuan B, Akhmetzhanov AR, Linton NM and Miyama T. 2020. "Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)." International Journal of Infectious Diseases.

Public Health Ontario (2020). "COVID-19: Epidemiologic Summaries from Public Health Ontario." Retrived from: covid-19.ontario.ca/covid-19-epidemiologic-summaries-public-health-ontario.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabail Chaudhry, George Dranitsaris, Talha Mubashir, Justyna Bartoszko, Sheila Riazi (2020). A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. doi:https://doi.org/10.1016/j.eclinm.2020.100464

Ruiz, D. A. P. (2017).Statistical Rethinking: A Bayesian Course with Examples in R and Stan Richard McElreath CRC Press, 2015, 469 pages, £67.99, hardcover ISBN: 978-1-482-25344-3 . International Statistical Review, 85(2), 379–380. https://doi.org/10.1111/insr.12229

Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., & Bhattacharya, J. (2020). Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. Jama, 323(23), 2425. doi:10.1001/jama.2020.8279

Statistics Canada (2020). Statistics Canada and contact tracing. Retrieved from https://www.statcan.gc.ca/eng/transparency-accountability/contact-tracing

Statistics Canada(2020). Canadian Perspectives Survey Series (CPSS). Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5311#a2

Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., & Bhattacharya, J. (2020). Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. Jama, 323(23), 2425. doi:10.1001/jama.2020.8279

Spychalski, P., Błażyńska-Spychalska, A., & Kobiela, J. (2020). Estimating case fatality rates of COVID-19. The Lancet Infectious Diseases, 20(7), 774-775. doi:10.1016/s1473-3099(20)30246-2

Vanessa Bates Ramirez.(2020) "What Is R0? Gauging Contagious Infections." Healthline, Healthline Media, 20 Apr. 2020, Retrieved from www.healthline.com/health/r-nought-reproduction-number.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, https://yihui.org/knitr/

Appendices

1. Github link which contains all the code, dataset, and report for the project: https://github.com/rubytianxiaoma/The-Usage-of-Biased-Smapling-Causing-Underestimation-of-Infected-Population-and-Weak-Policies

2. Karen Steward discussed her interesting perspective about herd immunity as a viable option to combat the Covid-19 pandemic in "Is a Herd Immunity Approach to the Coronavirus Outbreak a Viable Option?". URL as followed: https://www.technologynetworks.com/immunology/articles/is-a-herd-immunity-approach-to-the-coronavirus-outbreak-a-viable-option-332199 As stated in the article, herd immunity through natural infection is not the proposed solution as stated in the article. However, herd immunity through vaccination is the strongest way to prevent the spread and contain the covid-19 virus. Based on the article the basic reproductive ratio of the virus is about 2.5 and this equates to at least 60% of the population needing to be vaccinated for herd immunity to be effective. The paper also states herd immunity depending on mutation rates and the infectivity of the virus . If a virus has very fast rates of mutation, someone who had prior immunity to a virus does not necessarily have immunity to the newly mutated virus. An example of this is the flu shot in which you must get annually due to the high mutation rate of the influenza virus. Another factor affecting herd immunity is the infectivity of a virus. If the virus is highly infective, a higher percent of the population will need to become vaccinated to attain herd immunity. However, if the virus is not highly infective and does not transmit easily, then only a lower percentage of the population will need to get vaccinated in order to attain herd immunity.

3. Data used for estimate the contact tracing rate is adapted from the observational study conducted by Statistic Canada recording their calls completed for contact tracing for provinces Ontario, Alberta, Quebec, New Brunswick, Manitoba, Saskatchewan, British Columbia, Nova Scotia, Newfoundland and Labrador, and Nunavut. The daily capacity column is the number of calls which Statistics Canada is actually able to conduct for the province should the need arise.(Statistics Canada, 2020).