Sampling Bias in COVID-19 Contact Tracing Leads to Overestimation of Case Numbers from Institutional and Congregate Settings

Annie Collins

December 22, 2020

Abstract

"In this paper, I examine the potential impact of sampling bias in contact tracing on data regarding sources of infection for COVID-19 in Toronto, Ontario, Canada. I ran several simulations modeling true versus observed proportions of cases resulting from institutional and congregate settings and explored these results in comparison with Toronto Public Health data from the duration of the pandemic. Although institutional outbreaks have consistently accounted for a large proportion of all COVID-19 cases in Toronto, the simulations show that these numbers are likely inflated due to the ease at which individuals operating in these settings can be contact traced in comparison to the general population. This has great implications for any policy that is based on these metrics on an institutional, municipal, and provincial level."

Keywords: COVID-19; Contact tracing; Sampling bias; Simulation

Code and data for this analysis can be found at https://github.com/anniecollins/sta304-ps5.

Introduction

Contact tracing is an integral component of public health response to the COVID-19 pandemic. It is the process through which individuals who have had close contact with someone infected with the virus are identified and monitored to assess other potential cases and limit further infections (Public Health Ontario, 2020). This is much easier to do in some contexts than in others. Individuals operating in institutional or congregate settings (schools, universities, long-term medical care facilities, retirement homes, prisons, etc.) often have some or all of their location or personal contact history well documented due to the administration of these institutions. Other sources of transmission, such as community spread, casual events, travel, and outpatient healthcare, are significantly harder to trace due to a lack of documentation as well as personal recall bias.

For these reasons, it is possible that individuals in congregate settings are more frequently and accurately traced, affiliated with their transmission source, tested, and thus identified as new cases than those in other settings, resulting in data that reports highly inflated proportions of cases transmitted in institutional contexts. This is referred to as sampling bias, wherein the sample of the population that is tested for COVID-19 consists disproportionately highly of individuals affiliated with institutional or congregate settings. An additional troubling factor in this context is Toronto Public Health's temporary suspension of contact tracing initiatives outside of "high-risk" (potential outbreak) settings in October (Rocca, 2020).

In this paper, I aim to explore the impact this sampling bias may have on the data reported by Toronto Public Health (TPH) through the use of simulations. These simulations develop on work by Andrew Whitby (2020) who considered this bias in a more theoretical framework. I first explore TPH COVID-19 case data as it has been reported throughout the pandemic (January 23 through December 6, 2020) and some of its relevant statistics. I then simulate a two-step contact tracing procedure at varying tracing accuracies and examine the proportion of cases that appear to have resulted from institutional settings in comparison with those from casual contact or community spread contexts. This is followed by a discussion on its implications for TPH data as well as the policy or attitudes it may influence.

Data

To support these simulations and research, I will primarily be examining TPH data reported throughout the duration of the pandemic (cases recorded January 23 through December 6, 2020). TPH has recorded a total of 45,971 COVID-19 cases during this time period, along with demographic, location, source of infection, and medical outcome information for each case. I will also be discussing data reported through the Ontario Laboratories Information System (OLIS) for metrics related to COVID-19 testing. This data includes slightly different dates, time lags, and reporting metrics than the data retrieved from TPH directly, however it still provides useful insight within the context of the other data.

"Outbreak associated" cases are reported as those relating to Toronto healthcare institutions (i.e. hospitals, long-term care facilities, retirement homes) and other congregate settings (i.e. homeless shelters, prisons). These cases are distinct from those labeled "Institutional setting" (where a case occurred in an institutional setting but was not linked to other cases in the context of an outbreak) or "Healthcare setting" (where the virus was transmitted through outpatient medical care, such as a dentist or family doctor).

Table 1

Source of Infection	Total Cases	Percent of Total
Unknown/Missing	13793	29.9
Close contact	12564	27.3
Outbreak associated	10053	21.8
Community	6165	13.4
Healthcare	1709	3.7
Travel	1100	2.4
Institutional	670	1.5
Pending	49	0.1

Figure 1

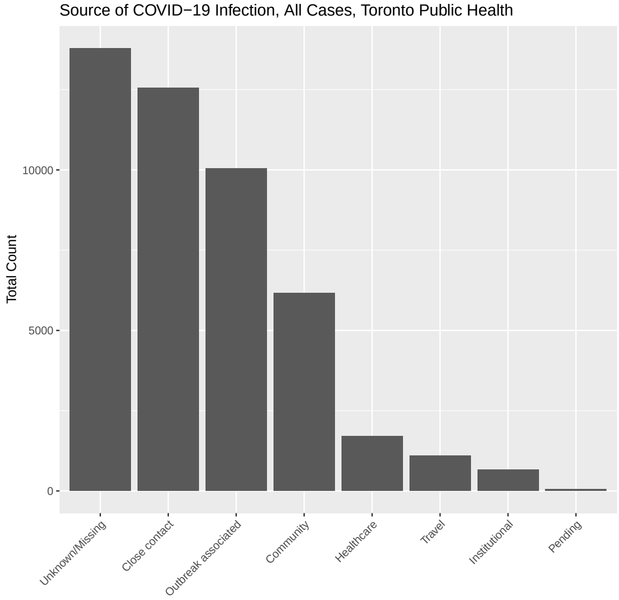
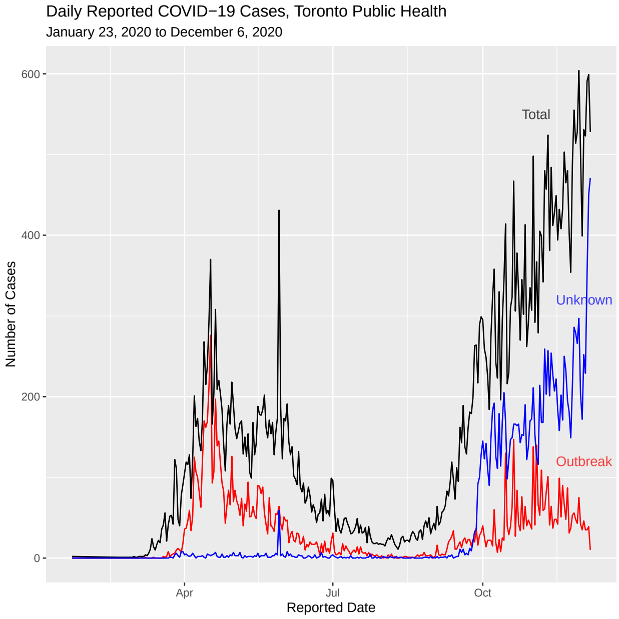
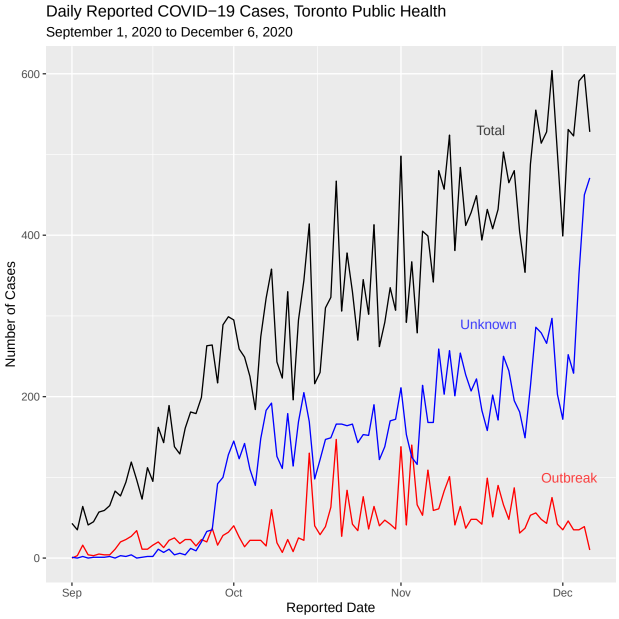


Figure 1 displays the total number of cases resulting from each source over the course of the pandemic in Toronto. Although the largest number of cases have not been traced to a specific source, a significant proportion (21.8 percent) of overall cases are "outbreak associated", in fact more than the lowest five sources combined (10,053 compared with 9,693). Close contact and community transmission, on the otherhand, are responsible for approximately 40 percent of total cases (18,729).

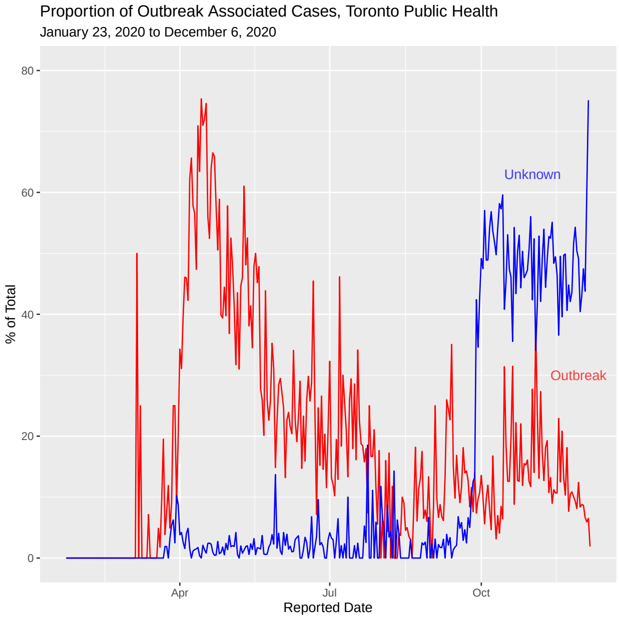
Figure 2 & 3

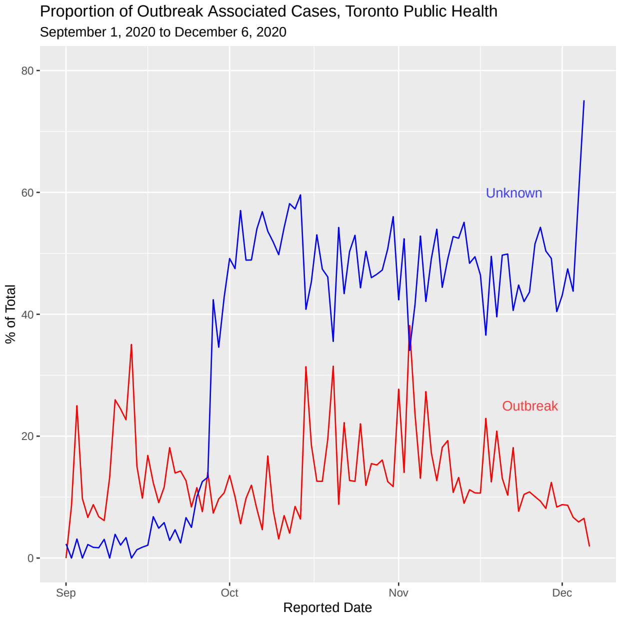




Examining daily case numbers throughout the pandemic (Figure 2), we see that the number of outbreak associated cases generally follows the same trends as the total cases over any given time period, with the exception of recent months. Since September, the total number of cases has drastically and consistently increased, while those resulting from outbreak settings appear to have increased in variance while maintaining relatively low overall case numbers. In contrast, cases with unknown or missing sources of infection increased drastically at the beginning of October hand have maintained high daily totals since. These trends are further highlighted in Figure 3, where TPH data has been restricted to cases reported on or after September 1. It is clear that the trend in outbreak-associated cases displays relatively little increase in comparison to all cases reported and those with missing or unknown data, meaning the proportion of cases traced to outbreak settings has generally decreased since September in comparison to earlier in the pandemic (the first wave). It also appears as though fewer cases are being traced (or traced effectively) with the increase in overall cases.

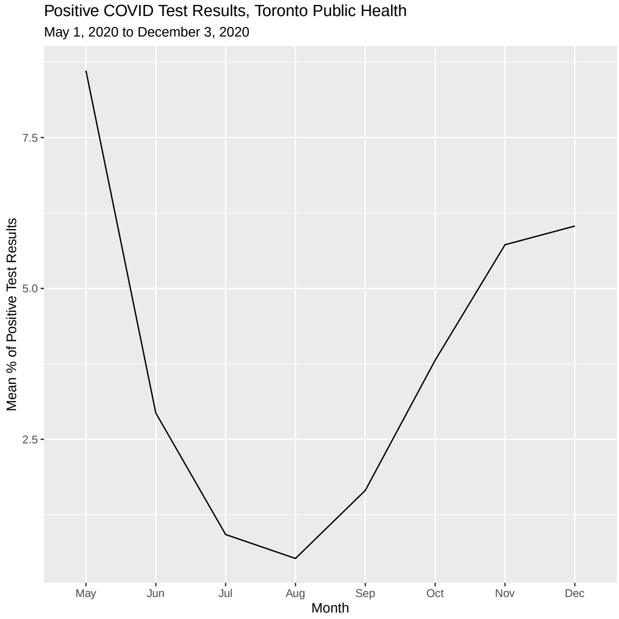
Figure 4 & 5





Figures 4 and 5 display trends in the proportion of total cases resulting from outbreak or unknown settings since January and since September. The average proportion of outbreak associated cases since September is 13.3 percent (compared to 20.4 percent overall). There has also been noticeable overall decrease in the proportion of cases resulting from outbreak settings, with peak values occurring in April and May despite peak total case numbers occurring in November and December. In contrast, the proportion of unknown sources of infection has increased dramatically, with a mean of 36.7 percent since October compared to 13.7 percent overall.

Figure 6



Finally, observing the changes in average proportion of positive test results over time, we see that it loosely follows the same trends as the total number of cases in Figure 2. We also observe that the mean value for the positive testing rate is approximately 3.5 percent across all available testing data (tests conducted since May 1), with the highest mean positive testing rates in May, November, and December. These proportions are highly susceptible to sampling bias, where individuals who are symptomatic for COVID-19 are more likely to pursue testing and thus increase the proportion of positive test results, likely above the rate of the population as a whole (Díaz-Pachón & Rao, 2020).

Simulation

The following simulations are intended to model the process of contact tracing and the resulting statistics for three different scenarios. Across the simulations, the contact tracing procedure remains the same but with varying efficacy. Our sample population consists of 2000 individuals: two institutional settings of 100 people each (200 individuals total), and 180 small gathering or community scenarios of 10 people each (1800 people total). I consider cases resulting from institutional and congregate settings in contrast with those traced to community spread and close contact scenarios as these are currently the top three sources of infection according to TPH data.

Each individual has a 5 percent chance of being infected independently of others in the sample, and these individuals have all been identified through public health testing procedures. 5 percent was chosen as the infection rate as it is similar to but slightly lower than the mean percent of positive tests observed by TPH in both November and December. As mentioned previously, the percent of positive tests is unlikely to represent the true rate of infection population-wide which is why a slightly lower infection rate was chosen for these simulations. The sampling, infection, and tracing process is simulated 50,000 times per scenario to gain insight into the resulting distributions of measurements, and all simulations were modeled using R software.

Table 2

Source of Infection	Distinct Settings	People per Setting	Total People	Approximate Number of Cases	Approximate Percentage of Total Cases	Simulated Tracing Accuracy
Institution	2	100	200	10	10	100%
Casual or Community	180	10	1800	90	90	20%, 50%, 80%
Total	182		2000	100		

Once individuals affiliated with an institution have tested positive, their COVID-19 positive status is linked directly back to their institutional setting in all simulated scenarios (due to a presumably well-defined record of populations and locations affiliated with both the individual and the broader institution). This also matches TPH protocol as of October, wherein only individuals linked to "high risk" settings are currently contact traced in an official capacity. Although this is no longer the case, the decision to halt contact tracing was made in response to overwhelming daily case numbers which have continued to rise, and limited or ineffective contact tracing remains a possible future scenario for TPH. 100 percent accuracy is an oversimplification for the purposes of simulation which aims to emphasize the relative simplicity of tracing institutional outbreaks compared to those in the broader community.

These simulations vary in the efficacy of contact tracing for potential community or small gathering transmission. There are several reasons why this is the case. In situations where contact tracing is delegated to the individual who tests positive, there comes a high degree of uncertainty and lack of accountability for accurately tracing one's recent contacts. Even in an official public health capacity, it is generally more difficult to trace individuals unaffiliated with institutions due to a lack of documentation affiliating a positive case with potential points of contact. There is also a high level of uncertainty in the efficacy of contact tracing apps, which are used voluntarily in Ontario and cannot guarantee that all at-risk contacts will receive a notification that they may be infected with the virus.

The three simulations model scenarios in which contact tracing for "low risk" individuals is completed with 20, 50, and 80 percent accuracy (ie. the percent of positive cases that get traced back to their associated event). In each scenario it is assumed that the contact tracing is conducted by the individual themselves, however it is possible that these levels of precision are also reflective of professional contact tracing as well.

Once the infected individuals have proceeded past the first stage of tracing, a secondary tracing process occurs, during which potential casual or community transmission events to which two or more infections have been traced are identified. Once these events are identified, any other un-traced cases that resulted from the event will be identified and traced to the event as well. For example, if three COVID-19-positive individuals attended the same party but only two of them identified the party as a potential site of transmission, the third individual will then be traced to the same party even though they did not identify the event as a point of transmission themselves.

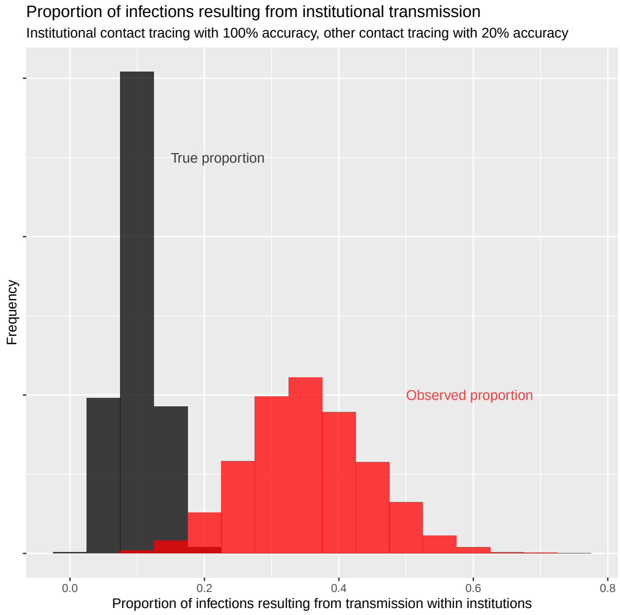
The two person per event threshold is derived from the Ontario COVID-19 case data glossary in which an outbreak is identified when two or more people who are not living together could reasonably have been infected in the same setting (Government of Ontario, 2020). This mainly applies to congregate and institutional contexts, however for the purposes of this simulation it will be used a threshold for all group settings.

This two-stage contact tracing process results in a number of COVID-19 cases that have all been traced to their respective institutions (one of two), and a number of COVID-19 cases that may or may not have been traced to a casual contact or community transmission event (one of 180), with those that have not been traced being excluded from the simulated "Source of Infection" data. The true proportion of cases from intitutions is measured as the number of positive cases traced to an institution divided by the total number of positive cases. The observed propotion of cases is the number of positive cases traced to an institution divided by the total number of positive cases that have been traced to any setting.

Results

Figure 7 displays the results of 20 percent efficacy in tracing casual contacts and community transmission. While the mean and modal true proportion of cases resulting from institutional settings is approximately 10 percent of all cases, the observed distribution has a much higher variance about a mean of approximately 35.5 percent. This means that if close contact and community-based infections are traced with an accuracy of 20 percent while institutional cases are traced with an accuracy of 100 percent, institutional cases could be measured to account for over triple their true proportion of total cases (and often even more).

Figure 7



If the accuracy of tracing is increased to 50 percent (Figure 8), the mean observed proportion of cases resulting from institutional settings decreases to 17.9 percent while the true proportion of approximately 10 percent is maintained. Finally, if the accuracy of tracing is increased to 80 percent (Figure 9), the mean proportion of infections traced to institutional settings is 12 percent, just slightly higher than the true proportion mean of 10 percent.

Due to the nature of the contact tracing process, it is much more difficult to connect individuals to small or casual events, and the resulting cases are hidden from statistics concerning sources of infection on a sample scale. This increases the observed proportion of cases from settings that are easy to trace, like those resulting from institutional settings. As the accuracy of contact tracing for casual and community settings increases between simulations, the proportion of cases traced to institutional settings approaches the true distribution. Although it is unlikely that a contact tracing accuracy of 100 percent for institutions is attainable, it is true that this subset of cases is more easily traced than those from other sources, and thus the observed number of COVID-19 cases resulting from institutional settings is likely to be inflated regardless of contact tracing methodology.

Figure 8

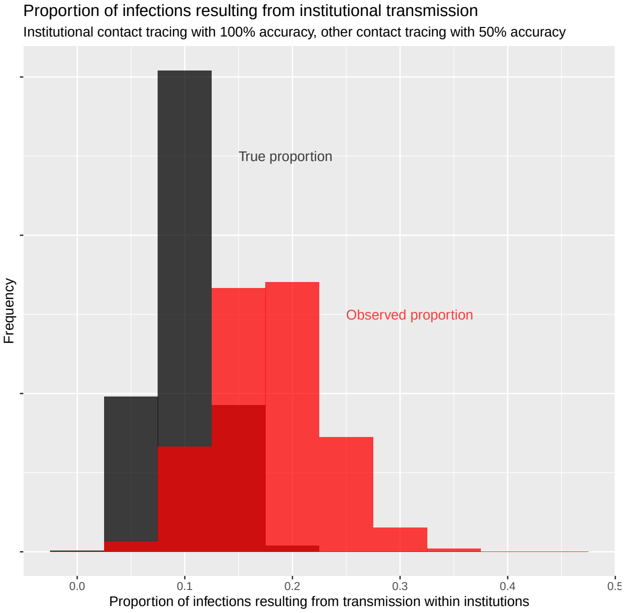
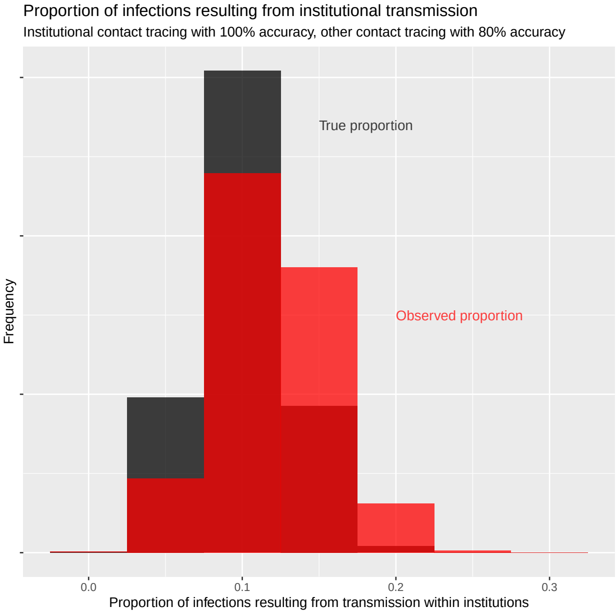


Figure 9



Discussion

As the results of simulations 1 and 2 (20 percent and 50 percent tracing accuracy) indicate, it is highly likely that the true proportion of cases resulting from institutional and congregate settings in Toronto is significantly lower than what is being measured. TPH data indicates that outbreak associated cases accounted for 21.8 percent of all cases as of December 6. Given a contact tracing methodology similar to the above simulations with an accuracy between 20 and 50 percent, it is very possible that the true statistic could fall at 10 percent or lower, reducing its significance by half and possibly more depending on accuracy and procedure. Considering the prevalence of unknown sources of infection in the data (29.9 percent overall, with significantly higher daily rates since October) and recent contact tracing priorities, it is likely that these cases could be largely attributed to non-outbreak sources and increase the proportion of casual contact and community transmission if they were properly traced.

It is also important to note that, even though contact tracing and thorough testing is conducted more easily and frequently in congregate settings, the proportion of these cases in recent months is relatively low compared to other stages of the pandemic and has stayed consistently low as daily cases have increased. This was true even during the month of October, during which TPH suspended contact tracing for cases outside of high-risk settings such as hospitals, long-term care facilities, and schools (Rocca, 2020). This seems to indicate that the spread of COVID-19 in institutional contexts may currently comprise a fairly small proportion of all transmission in Toronto. This is troublesome not only for members of the general populous, but also for those operating in long-term congregate living settings where high levels of community spread increase the likelihood for an outbreak (Crawley, 2020).

Sampling bias and its influence on contact tracing must be acknowledged if public health and government officials are to make well-informed decisions in response to the COVID-19 pandemic in Toronto. Limiting testing and tracing delays are key in preventing further transmissions of the virus, and ensuring thorough, well-funded, and well-staffed contact tracing efforts on a population scale is imperative to this goal (Reitjes, 2020; Kretzschmar et al., 2020). Beyond contact tracing, it should be recognized that community transmission and casual contact spread may be the source of the vast majority of COVID-19 cases in Toronto and stricter restrictions should be implemented to prevent these cases from increasing, such as closure of recreational venues, reducing legal capacities for indoor community settings, and enforcing stricter lockdown restrictions.

This analysis does not aim to minimize the importance of preventing the spread of COVID-19 in congregate or institutional settings. I merely suggest that, in order to mitigate the spread of COVID-19 throughout the population as a whole, we are in need of more robust contact tracing measures for *all* positive cases and stronger policy preventing transmission in casual and community settings.

Bibliography

City of Toronto Open Data. (2020). *COVID-19 Cases in Toronto* [CSV data file]. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Crawley, M. (2020, November 29). *How Ontario's second wave of COVID-19 is hitting long-term care.* CBC News. https://www.cbc.ca/news/canada/toronto/ontario-covid-19-long-term-care-1.5820944

Díaz-Pachón, D. A., & Rao, J. S. (2020). A simple correction for covid-19 testing bias. *ArXiv*, arXiv:2007.07426v2.

Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra

Government of Ontario. (2020). *COVID-19 case data: Glossary.* Government of Ontario. https://covid-19.ontario.ca/data/covid-19-case-data-glossary#outbreak

Kirill Müller (2020). here: A Simpler Way to Find Your Files. R package version 1.0.0. https://CRAN.R-project.org/package=here

Kretzschmar, M.E., Rozhnova, G., Bootsma, M., Van Boven, M., Van de Wijgert, J., & Bonten, M. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health, 5*(8), E452-E459. https://doi.org/10.1016/S2468-2667(20)30157-2

Ontario Laboratories Information System. (2020). *Positive Results, Toronto Public Health* [CSV data file]. https://covid-19.ontario.ca/data

Peter Solymos and Zygmunt Zawadzki (2020). pbapply: Adding Progress Bar to '*apply' Functions. R package version 1.4-3. https://CRAN.R-project.org/package=pbapply

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

Reintjes, R. (2020). Lessons in contact tracing from Germany. *The BMJ, 369*(m2522). https://doi.org/10.1136/bmj.m2522]

Rocca, R. (2020, October 3). *Toronto Public Health suspends some contact tracing due to 'high level' of coronavirus cases.* Global News. https://globalnews.ca/news/7376562/toronto-public-health-coronavirus-contact-tracing-suspension/

Whitby, A. (2020, November 25). Contact tracing can give a biased sample of COVID-19 cases [Blog post]. https://andrewwhitby.com/2020/11/24/contact-tracing-biased/

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

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