TODO: CITE R: Unpacking why almost half of COVID-19 cases in Toronto have no information about their source*

Investigating possible data collection biases using patient demographics

Sidharth Gupta

06 February 2022

Abstract

Data about the sources of COVID-19 infection can be used to guide epidemic and lockdown policies as the pandemic evolves; however almost half of COVID-19 cases in Toronto have no information about their source. In this paper we analyze the patient demographics of these cases with no information, and when they occur. We identify both groups of patients and time-points in the pandemic where there is a large proportion of cases with no source information. Our results can be used to improve the current methods for collecting data about COVID-19 cases in Toronto and other regions so that more sources of infection can be identified.

1 Introduction

The COVID-19 pandemic has made a major global and societal impact. From its beginning to this time of writing, there have over 390 million positive recorded cases of COVID-19. The virus has a contagion that is exponential in nature, and spreads easily in most indoor and close-contact settings. To counteract the rapid and dangerous spread of COVID-19, lockdown and social isolation policies are regularly reviewed and enforced. These lockdowns are effective in controlling the spread of COVID-19, however they have shown to create many societal, economic, and global mental health consequences. For instance, (Le and Nguyen 2021) found that lockdowns cause people to experience adverse psychological events, such as anxiety, disinterest, depression, and an overall reduction in mental health. (Larue 2021) analyzed that lockdowns in Canada impact low-income workers and vulnerable communities significantly with employment instability, and (Kaur, Jagpal, and Paudyal, n.d.) discover that lockdowns reduce the ability to provide services for social causes such as homelessness. Many new methods have been introduced for improving lockdown design, such as modelling pandemic mobility with data (Dutta 2021), and developing public health messaging systems (Block et al. 2020). These methods stem from the prevalent topic in the literature that lockdown strategies should be smart – that is, they should minimize isolation consequences and maximize virus spread containment Olivier, Botha, and Craig (2020). Datasets that identify outstanding regions for COVID-19 spread are vital, because they can identify areas that most benefit from lockdown, and areas that least benefit. As of the time of this writing, the City of Toronto has been recording this kind of data from positive COVID-19 cases since the start of the pandemic. Each recorded COVID-19 case has a field that identifies the source of infection; however 42% of cases in this dataset have no information entered for this field. Such a large proportion of missing data about the source of infection significantly reduces this dataset's utility for aiding smart lockdown designs. Additionally, this large proportion of missing data suggests that there are flaws within the data collection process itself. Because this COVID-19 data is collected weekly and is used for high-stakes decision making, finding flaws in data collection and addressing them quickly is pivotal.

In this paper, we thoroughly describe how this dataset is collected, and propose three biases that can explain why data for the source of infection is missing. We affirm the plausibility of these proposed by visualizing the data in three perspectives. The first perspective visualizes how the proportion of cases with no source

*Code and data are available at: TODO.

Table 1: One record in the COVID-19 cases data

Outbreak Associated	Age Group	Neighbourhood Name	FSA	Source of Infection
Sporadic	30 to 39 Years	Brookhaven-Amesbury	M6M	Household Contact

Table 2: One record in the COVID-19 cases data, continued

Classification	Episode Date	Reported Date	Client Gender	Outcome
CONFIRMED	2021-04-07	2021-04-12	MALE	RESOLVED

information is related to the different months in the pandemic's lifetime. The second perspective illustrates the age groups of individuals who have a high proportion of cases with this field missing. Finally, we cross reference the neighbourhoods in this dataset with another dataset to find their income index values, and observe that neighbourhoods with a high proportion of missing fields in their data have a lower income index than neighbourhoods with a lower proportion. With data to support the plausiblity of these biases, we hope that our results can be used to further investigate the data collection process of COVID cases in Toronto and reduce the proportion of records with this field missing.

2 Data

The dataset of COVID-19 cases in Toronto is provided by the Open Data Toronto portal (Gelfand 2020). This dataset is curated by the City of Toronto (Toronto 2019) and it is refreshed with new records on a weekly bases. The raw data comes from The Public Health Case and Contact Management (CCM) group, who collect COVID-19 case and outbreak from all of Ontario's Public Health Units at 1:00 P.M Eastern time each day (Toronto, n.d.). The first row of the curated dataset are shown in Table 1 and Table 2; some fields are not shown in these tables due to their rare occurrence. The focus of our paper will be the header for "Source of Infection." As stated in the dataset's technical guide (Toronto 2022), this field is determined by a public health investigator's assessment. If this assessment is absent, then other data fields may have been used to estimate the source of infection, such as if there a recorded household positive case. A breakdown of responses for this field is shown in Table 3.

The described data collection process has many venues for biases to be introduced, and we will discuss three such biases. The first centers around traffic in different Public Health Units (PHUs). The recorded COVID-19 cases come from individual PHUs across the neighbourhoods of Toronto, and in an unbiased dataset, each COVID-19 case would contain equal information even if it was processed in a different PHU.

Table 3: Distribution of each response to the Source of Infection field.

Source of Infection	Proportion of response against all cases
No Information	0.43
Community	0.23
Household Contact	0.13
Outbreaks, Healthcare Institutions	0.06
Close Contact	0.06
Outbreaks, Other Settings	0.04
Pending	0.03
Outbreaks, Congregate Settings	0.01
Travel	0.01

However, different PHUs receive different amounts of traffic, especially during months when the pandemic shows a high surge of cases. Labour shortages in healthcare is a well-known issue in metropolitan cities that has been severely affected by COVID-19 (Mascha et al. 2020), and as such different PHUs may not have equal capacities to handle noise in data collection, or even collect fields like the source of infection at all. A second venue for bias comes with the process of identifying the sources of COVID-19 infection through human interactions. Symptoms for COVID-19 can occur several days after the virus is initially contracted, and so asking a patient where they think the source of infection depends entirely on their own memory and comfort levels. Several studies in cognitive psychology (Bard 2015) have shown that human memory is unreliable and irrational, and distrust can sway decision making (Platt, Jacobson, and Kardia 2018). The third bias digs deeper in the issues of trust; even if a patient is confident about the source of their infection, they may not feel comfortable sharing that information. Some lockdown policies have legal consequences if they are not followed, and so some patients may fear the potential consequences that come with sharing their information. We would expect to see higher levels of distrust in low-income neighbourhoods, because those are often populated by people who are most vulnerable to legal consequences.

Starting with the first bias of traffic in PHUs, we can explore how surges in cases influence the proportion of cases with no information about the source.

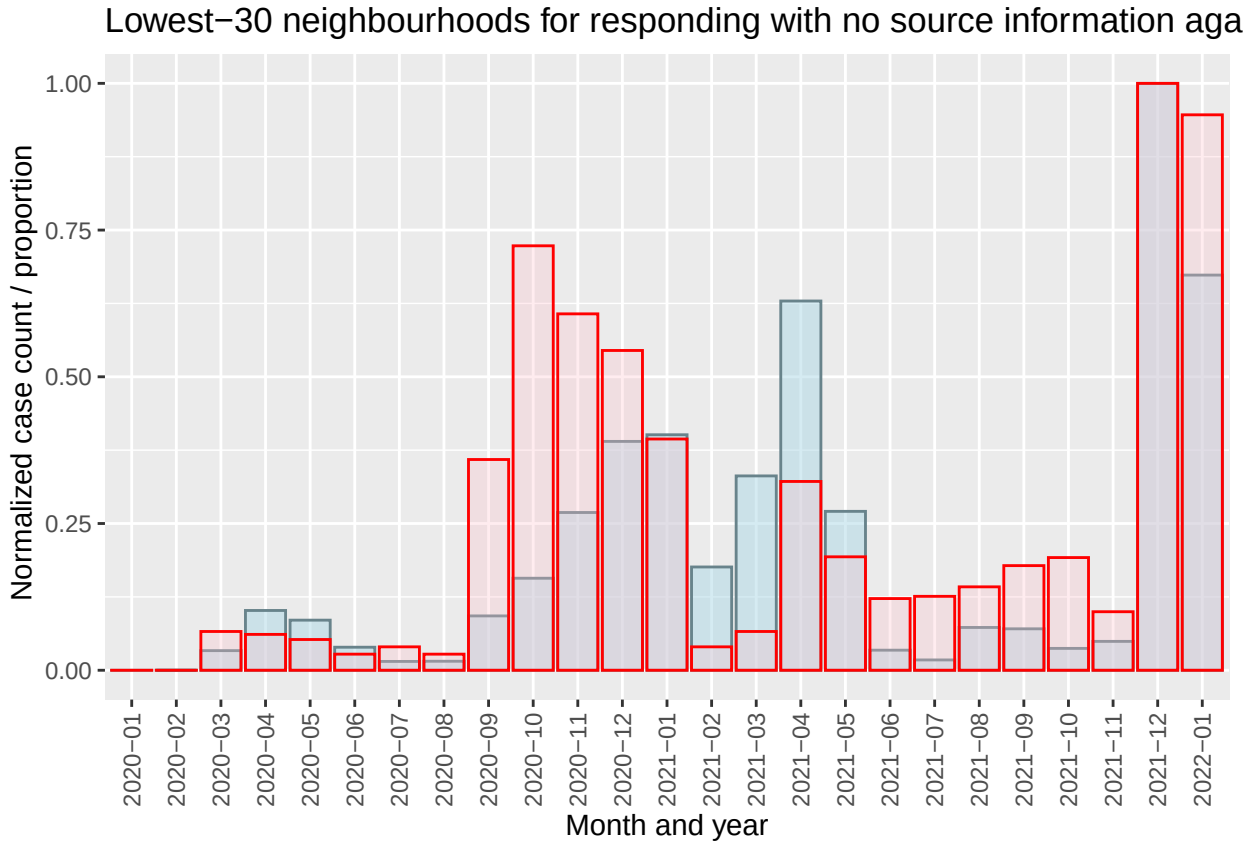
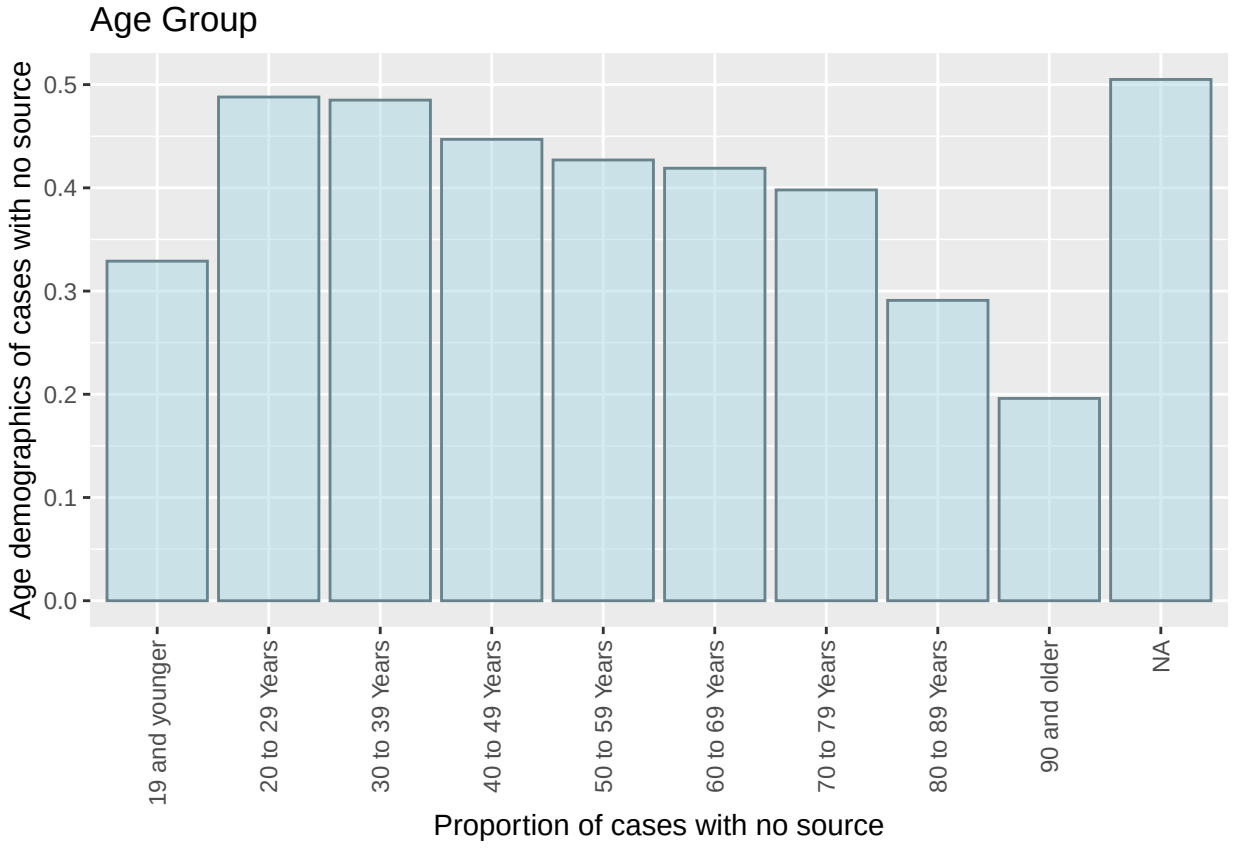


Figure 1: TODO



References

Bard, Jennifer S. 2015. "Ah Yes, i Remember It Well: Why the Inherent Unreliability of Human Memory Makes Brain Imaging Technology a Poor Measure of Truth-Telling in the Courtroom." *Oregon Law Review* 94 (2): 295–358. https://heinonline.org/HOL/PrintRequest?handle=hein.journals/orglr94&collection=journals&div=11&id=297&print=section&sction=11.

Block, Per, Marion Hoffman, Isabel J. Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C. Mills. 2020. "Social Network-Based Distancing Strategies to Flatten the COVID-19 Curve in a Post-Lockdown World." *Nature Human Behaviour* 4 (6): 588–96. https://doi.org/10.1038/s41562-020-0898-6.

Dutta, Susana N. AND Kalise, Ritabrata AND Gomes. 2021. "Using Mobility Data in the Design of Optimal Lockdown Strategies for the COVID-19 Pandemic." *PLOS Computational Biology* 17 (8): 1–25. https://doi.org/10.1371/journal.pcbi.1009236.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Kaur, Simran, Parbir Jagpal, and Vibhu Paudyal. n.d. "Provision of Services to Persons Experiencing Homelessness During the COVID-19 Pandemic: A Qualitative Study on the Perspectives of Homelessness Service Providers." *Health & Social Care in the Community* n/a (n/a). https://doi.org/https://doi.org/10.1111/hsc.13609.

Larue, Bruno. 2021. "COVID-19 and Labor Issues: An Assessment." *Canadian Journal of Agricultural Economics/Revue Canadienne d'agroeconomie* 69 (2): 269–79. https://doi.org/https://doi.org/10.1111/cjag.12288.

Le, Kien, and My Nguyen. 2021. "The Psychological Consequences of COVID-19 Lockdowns." *International Review of Applied Economics* 35 (2): 147–63. https://doi.org/10.1080/02692171.2020.1853077.

Mascha, Edward J., Patrick Schober, Joerg C. Schefold, Frank Stueber, and Markus M. Luedi. 2020. "Staffing with Disease-Based Epidemiologic Indices May Reduce Shortage of Intensive Care Unit Staff During the

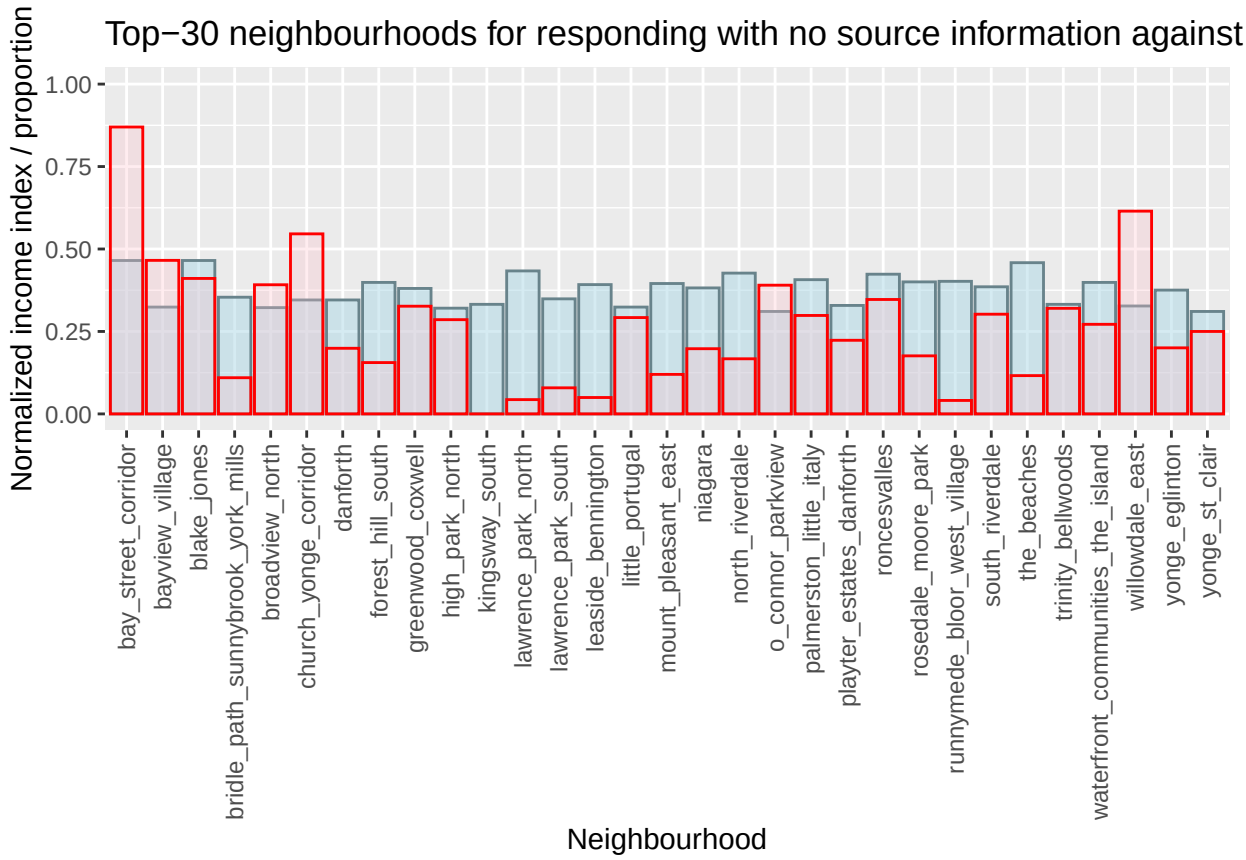


Figure 2: TODO

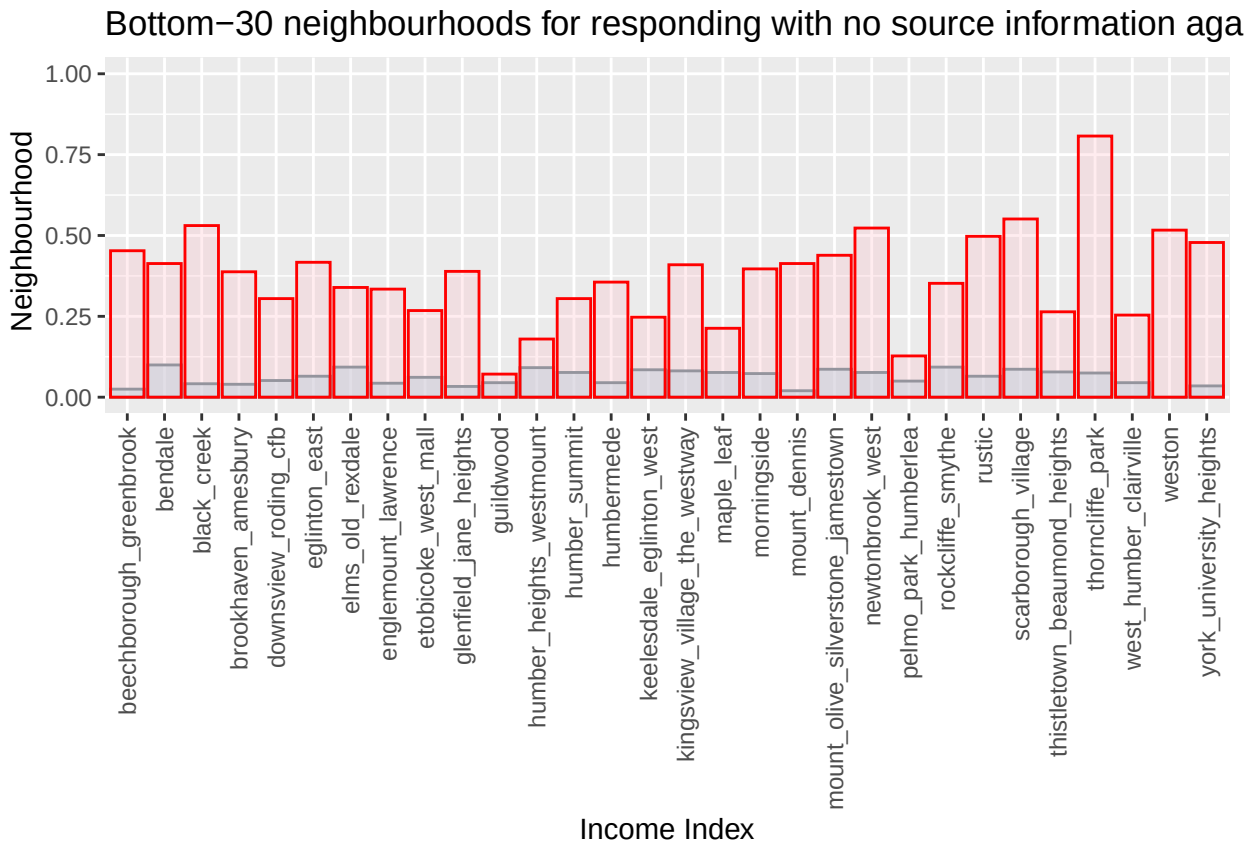


Figure 3: TODO

COVID-19 Pandemic." *Anesthesia and Analgesia* 131 (1): 24–30. https://doi.org/10.1213/ANE.0000000000004849.

Olivier, Laurentz E., Stefan Botha, and Ian K. Craig. 2020. "Optimized Lockdown Strategies for Curbing the Spread of COVID-19: A South African Case Study." *IEEE Access* 8: 205755–65. https://doi.org/10.1109/access.2020.3037415.

Platt, Jodyn E., Peter D. Jacobson, and Sharon L. R. Kardia. 2018. "Public Trust in Health Information Sharing: A Measure of System Trust." *Health Services Research* 53 (2): 824–45. https://doi.org/https://doi.org/10.1111/1475-6773.12654.

Toronto, City of. 2019. "Open Data." *City of Toronto*. https://www.toronto.ca/city-government/data-research-maps/open-data/.

———. 2022. "COVID-19: Epidemiological Summary of Cases." *City of Toronto*. https://www.toronto.ca/home/covid-19/covid-19-pandemic-data/covid-19-epidemiological-summary-of-cases-data/.

———. n.d. "All Ontario: Case Numbers and Spread." *COVID*. https://covid-19.ontario.ca/data/case-numbers-and-spread#where_numbers.

Wang, Chunlei, Dake Wang, Jaffar Abbas, Kaifeng Duan, and Riaqa Mubeen. 2021. "Global Financial Crisis, Smart Lockdown Strategies, and the COVID-19 Spillover Impacts: A Global Perspective Implications from Southeast Asia." *Frontiers in Psychiatry* 12 (September). https://doi.org/10.3389/fpsyt.2021.643783.