The COVID-19 Numbers are just the tip of the Iceberg*

The data we simply can't trace

Bo Phi

06 February 2022

Abstract

During the recent pandemic, the datasets surrounding the COVID-19 virus have played an important role in navigating the future to move forward around the pandemic, however, the manner in which cases are reported could have effects on the credibility and integrity of the data. This statistical report uses data gathered by the city of Toronto to (1) examine the ways in which the cases have been recorded over time along with the metadata that is attached to each case, and (2), discuss the ways in which the data has an influence on peoples actions, and how those actions are further reflected within the data. With the number of cases having reached a record amount at the end of 2021, it is important to explore possible explanations for why that is the case. Additionally, with many of the direct sources of cases, and many cases also going unreported, it makes it incredibly difficult to estimate the true magnitude of the pandemic in Toronto.

1 Introduction

Since the beginning of 2020, the world has been struck by a global pandemic. A handful of politicians took the early signs of the pandemic very seriously and went into early preventative measures while many politicians did not. As a result, the virus began to spread like wildfire. Before very long, nearly all countries began the process of shutting down, closing nearly all parts of the economy to protect the world from this virus. In the United States, the President openly expressed control over the situation, assuring the cases were being reported but outbreaks were being contained. Soon after that, hospitals began to be flooded with COVID-19 patients and schools began to close as a result. After the development of the vaccine, people have been in a rush to get their immunization but a reasonable chunk of people still opt out of getting vaccines. There are a multitude of reasons for opting out of the vaccine but one of the most common reasons is the lack of belief in the vaccine. People who opt out of vaccines also have a lack of belief in the data and the governmental bodies that collect the data as well. Are their reasons justified?

Two years later in 2022, the same problems are still present. The pandemic has seen no real end and cases have spiked up to the tail end of 2021. Millions of people have died and there are still skeptical people on the validity of the data associated with COVID-19. Good data would serve as an important tool to inform the public and governing officials on choices they make but as it stands, the effectiveness of the current data doesn't seem to be convincing enough people to work towards ending the pandemic. With new variants emerging and the world itching for things to go back to normal, examining why the current model isn't working in an effort to suggest a better one seems appealing but is it even possible?

The importance of truthful pandemic statistics in informing the public cannot be understated as it seemingly determines the rate at which the world can move past COVID-19. For this report, I will be using an open-access dataset from the Toronto Public Health organization to examine the metadata that is associated with a report of a case. More importantly, I will use the data to discuss the ways in which the data is somewhat incomplete and incapable of telling the full story of the pandemic. The data will be processed and analyzed in R (R Core Team 2020) using the 'tidyverse' (Wickham et al. 2019) amd 'dplyr' (Wickham et al.

*Code and data are available at: https://github.com/bophiwork/paper-1

2021) packages. The packages 'knitr' (Xie 2021a), 'bookdown' (Xie 2020), and 'tinytex' (Xie 2021b) are also used to generate the final markdown report.

2 Data

This report uses data on COVID-19 cases in Toronto obtained by the Toronto Public Health organization (Health 2022). Toronto Public Health is responding to an ongoing COVID-19 outbreak, in the context of an evolving global pandemic. This data set contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since the first case was reported in January 2020. This includes cases that are sporadic (occurring in the community) and outbreak-associated. The data are extracted from the provincial Case & Contact Management System (CCM). The data is updated on a weekly basis so the numbers may differ from other reports using the same dataset during different times.

The dataset features a good amount of metadata attached with each case. The age of the person infected, the outbreak associated, the name of the neighborhood, the perceived source of infection, along with the reported dates of infection are just some of the data collected as part of this dataset. The main pieces of data I will be focusing on is the age of the infected person per case reported, the perceived source, and the estimation of when the disease was acquired. This is a sample table of what our cleaned up dataset will look like. It is important that this data is collected from every reported case in Toronto so the sample size is large inevitably due to the large population of the city leading to a large amount of reported cases.

```
## # A tibble: 6 x 4
##   Assigned_ID `Age Group`    `Source of Infection` `Episode Date`
##         <dbl> <chr>          <chr>                 <date>
## 1           1 50 to 59 Years Travel                2020-01-22
## 2           2 50 to 59 Years Travel                2020-01-21
## 3           3 20 to 29 Years Travel                2020-02-05
## 4           4 60 to 69 Years Travel                2020-02-16
## 5           5 60 to 69 Years Travel                2020-02-20
## 6           6 50 to 59 Years Travel                2020-02-24
```

The graph below shows the raw amount of COVID-19 reported cases in Toronto. As seen in the graph, the daily cases never top 2000 new cases in one day until recently at the end of 2021. It is hard to pinpoint determinants for this trend. We know that during Q1 of 2021, the various vaccines were done developing and released to the public which could explain the downtick in reported cases during the middle part of 2021. There also seems to be an uptick in reported cases between winter of 2020 into early 2021. Towards the tail end of 2021, the cases skyrocket reaching record numbers. (Figure 1).

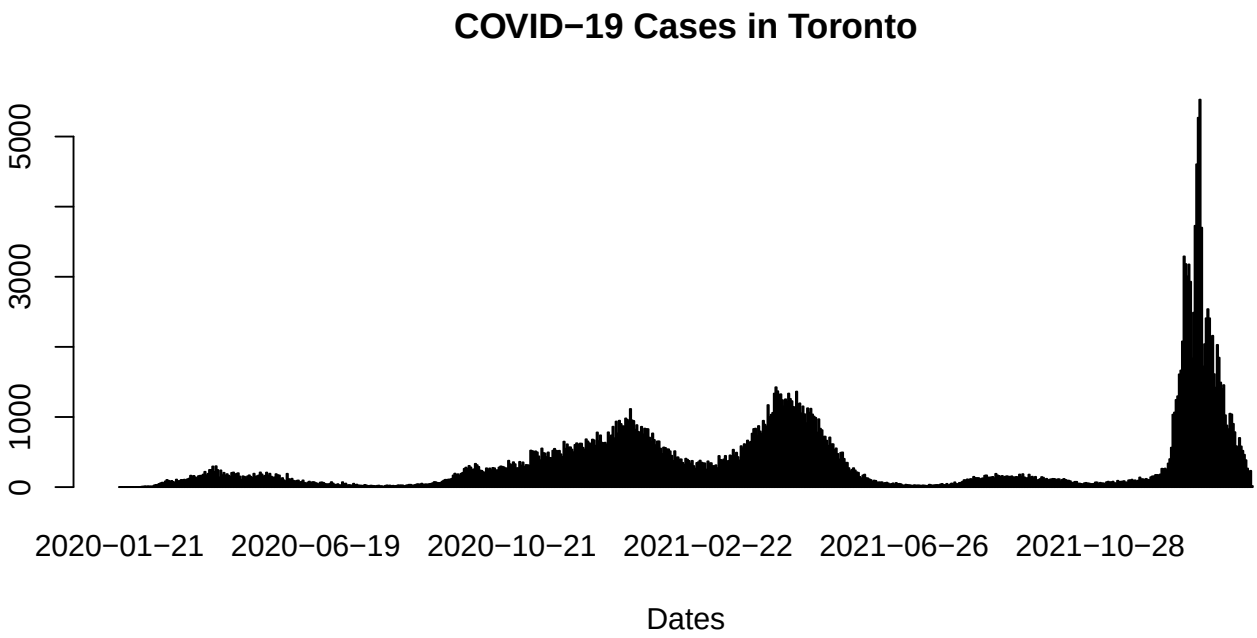

Figure 1: Episode Dates in Toronto

The plot below shows the various age groups people belonged to in all of the reported cases in Toronto. We find that people aged 40 and younger make up a large majority of all the total cases reported. The three age ranges within that 40 year range also have the 3 highest individual number of cases with people aged 20-29 testing positive the most. The data shows a dramatic dropoff in reported cases for people above the age of 70. It is important that this graph only features data from recorded and reported cases. The reality of cases that go unreported are not reflected in this plot. (Figure 2).

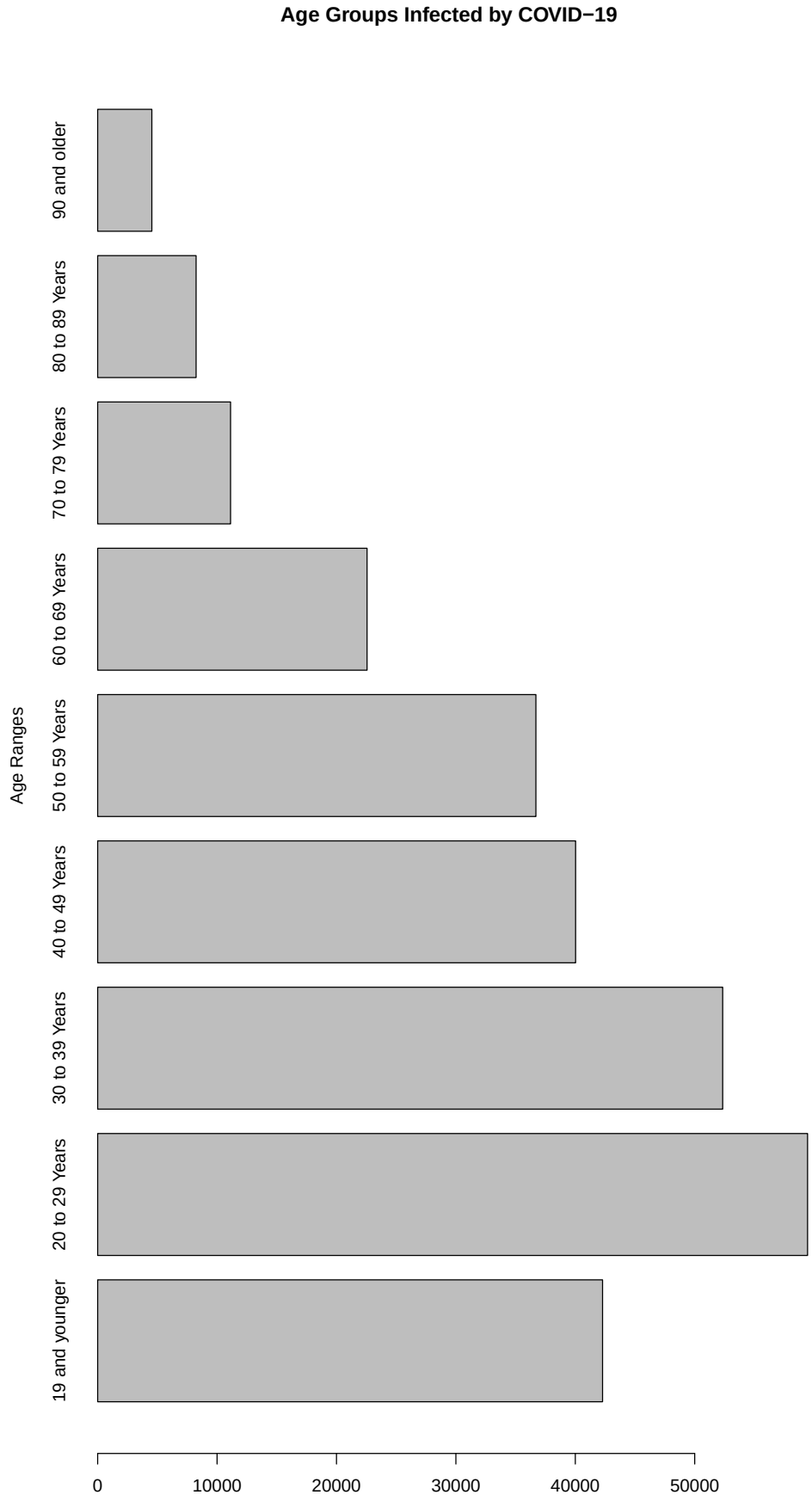

Figure 2: Age Ranges of Infected People

The plot below shows us the perceived sources of contraction for each person reporting a case. In the plot, we find that one bar sticks out much further than the others, and that outlier is "other." Other denotes that the reported case couldn't be attributed to a definitive source of contraction and was used when the person the case referred to had no knowledge of how they contracted the virus. The second most common source of contraction was found to be from the community. This was defined as cases who did not travel outside of Ontario, did not identify being in close contact with a COVID-19 case, and were not part of a known confirmed COVID-19 outbreak. Surprisingly, the lowest found sources of contraction were found to be through travel and through outbreak events. (Figure 3).

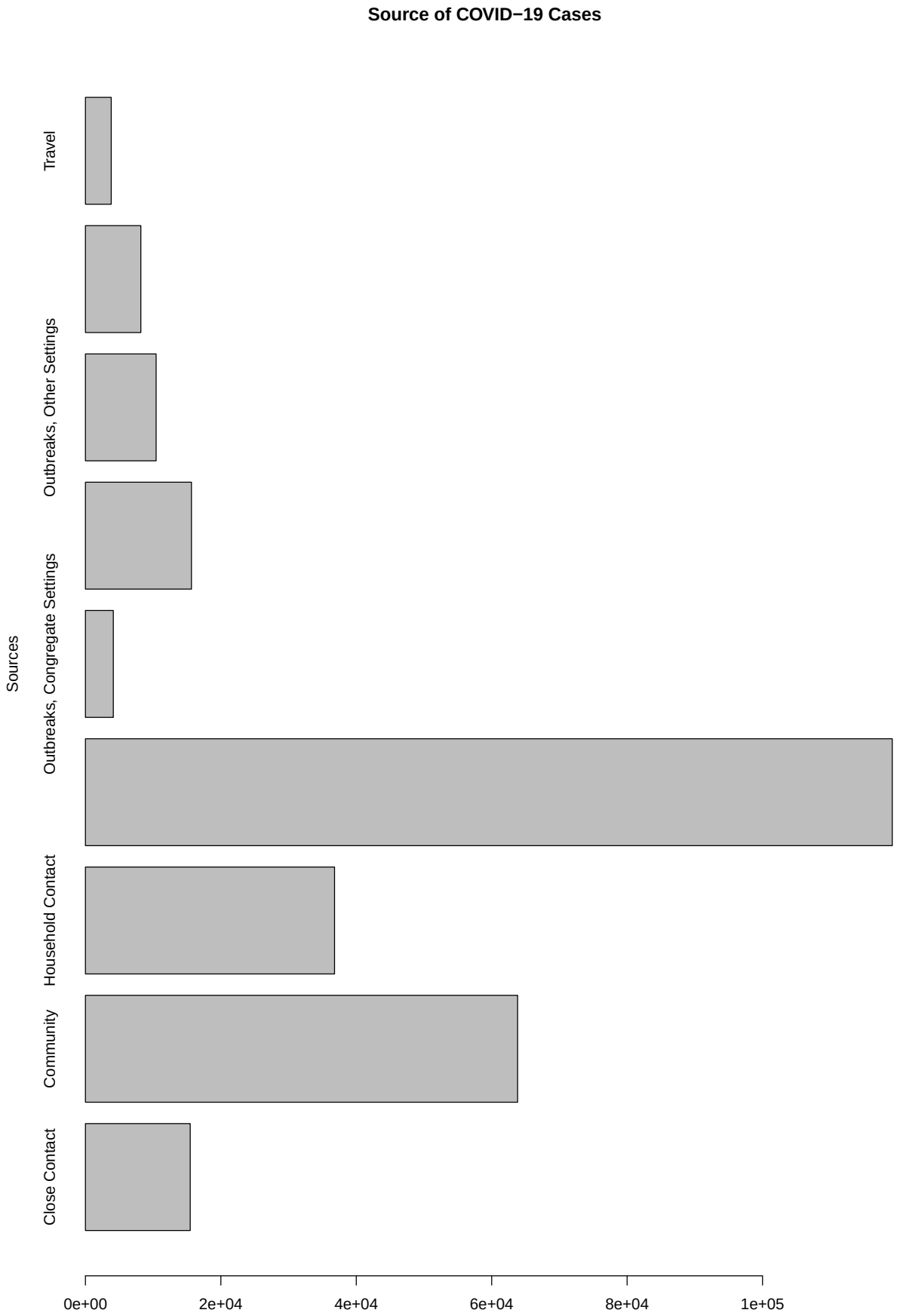

Figure 3: Source of COVID-19 Contractions

3 Results

The data plots allow us to pinpoint the number of cases during each stage of the pandemic, which age groups are primarily infected, and the alleged sources of contraction for each case. We find that new cases spiked at the tail end of 2021, the most common positive testing age group are found in people aged 40 and younger, and that most people have no information as to how they contracted the virus in the first place.

4 Discussion

The dataset leaves a lot to be wished for. Some of the missing points of the data could contribute to the skepticism towards the COVID-19 numbers in the first place. There is also the risk that the numbers could be inflated or misrepresent reality. I will be discussing the various forms of skepticism regarding the dataset in regards to uncollected data and the accuracy of the collected data.

4.1 Sources of Skepticism

Looking at the data and what we were able to produce with it, it seems largely incomplete and doesn't do a good job of telling the whole story of the pandemic. Let's first take a look at the number of reported cases. With travel restrictions loosening up since the mid-point of 2021, more people are traveling and required to test as a result. Many countries are enforcing rules that require travelers to have mandatory vaccines along with a negative test before they are able to board their flights. This increase in testing could have resulted in an increase in positive tests simply because there is more testing volume. In other words, the more testing that is conducted, the more likely it will be that positive tests be found. What this idea tells us more than anything, is that it is nearly impossible to get a true and accurate representation of how many people truly have COVID-19. The only way in which a case is recorded is if a positive test is found, so if people aren't getting tested, a large number of positive cases could be missed. What the data also misses is a distinction between total cases and total new cases. The dataset creates a unique ID for each report of a positive test case but doesn't account for the case where the same person can test positive twice during time periods and end up being recorded twice in the dataset. Depending on which one of these factors is more prevalent, the COVID-19 numbers could be largely seen as inflated or underrealized. I would argue this is the source of much of the skepticism surrounding the pandemic. Because the data isn't largely telling the whole story, people could argue that since the data isn't entirely the truth, in turn, the data is largely false. While there is not a feasible to achieve statistically perfect data because of the recording limitations, I would argue that statistical perfection exists in nearly every model, especially when done to a sample size. The sample size is meant to model the trends in a broader population so while we cannot definitively say that the numbers are completely true, we can use the data collected to develop inferences and inform us of the broader population.

4.2 Weaknesses and next steps

The biggest weaknesses in the dataset are what isn't included in the dataset. A distinction between new cases and total cases isn't represented, and there is no way of getting the full scope of the total cases in the city of Toronto. As a result, as it stands, the dataset is fairly weak. Depending on how you look at it, the results of the data can seem largely inflated or underrealized which poorly represents the true state of the pandemic. Moving forwards, it would be incredibly insightful to have more metadata exist alongside the existing dataset. Being able to distinguish unique cases versus total cases would be helpful. Not being limited to the city of Toronto would also be helpful. Gaining more information on the demographic of people getting tested and the test-sites people were going to. Having information on the reasons people were getting tested would provide insight on the motivations behind getting tested and could help us understand some of the reasons why people aren't getting tested. The next steps of this analysis would be scaling it to a broader scale. Comparing the dataset with nationwide statistics and potentially global statistics would help us understand if the trends observed in Toronto are mimicked in datasets globally beyond the scope of Toronto. Having information on the average income of different areas can give us insight as to whether there is a class element to being more or less likely to test positive. It would be interesting to see if certain areas of Toronto, or the world, are more susceptible to contracting COVID-19 and finding the underlying reasons behind that. In the

case of Toronto being a densely populated city, it would make sense that the cases would be high but within Toronto, there are likely neighborhoods and areas that are linked to higher rates of contraction which could be the result of a housing or education issue. The data could be looked at through so many different lenses that comparing it with different data sets can lead to different conclusions.

References

Health, Toronto Public. 2022. *COVID-19 CASES IN TORONTO*. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://github.com/rstudio/bookdown.

———. 2021a. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2021b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. https://github.com/yihui/tinytex.