CoVid-19 Outbreak in the subpopulation of long term care homes in Ontario

Fang Wang * Ruyi Pan * Hongyi Ding † Yufeng Li †
Patrick Brown ‡

January 15, 2021

Abstract

In this study, we used negative binomial generalized addictive model (GAM) to fit the daily death data due to Covid in Ontario from April to December,2020. We showed there is an outbreak in the subpopulation of Long Term Care (LTC) homes prior to the outbreak of the whole population.

1 Introduction

As of January 1 - 2021,7 pm EST, 652,473 people in Canada had tested positive for the coronavirus disease 2019 (CoViD-19).Among these people, 16,833 deaths had occurred. The majority of cases (67.1 percent) and deaths (80.6 precent) have been reported by Ontario and Quebec. Specifically, more than four thousands of deaths in Ontario are linked to the ongoing pandemic caused by(CoViD-19) [1], many of which are the residents of long term care centers (LTC). CoViD-19 is known to be more fatal to older adults with chronic diseases [2], which gives it a different epidemiological characteristics in the subpopulation of LTC residents, which is denoted by $\mathcal{P}_{\text{LTC}}$.

In our study, we modeled daily death count with negative binomial generalized addictive model (GAM);to account the difference in fatality of CoViD-19 across different populations, we allow distinct mortality rate time series across the $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}_{\text{LTC}}^{C}$, where the universal set is taken to be all susceptible people.

1.1 Long Term Care Homes

In Ontario, LTC homes are facilities that provide support to adults that can no longer live independently [3]; they have more than 14.5 millions of residents which is an important part

*MsC Student at Department of Statistics Science-University of Toronto
†Undergrad at Department of Mathematics and Statistics-McMaster University
‡Professor at Department of Statistics Science-University of Toronto

of social welfare program [4]. Many of them provide government funded medical services, but are less regulated than hospitals and are poorly coordinated [5]. Ontario administration started their supporting plan for LTC workers on mid April [6], which then effectively controlled the first wave of pandemic in LTC homes, but such action is too late comparing with other province [5]. More than half of residents are over 85 years old and 63% of them are sharing rooms [5], which makes them susceptible population in a high risk settings. Some researchers warned LTC homes are dangerous to the public heath system [7], which is consistent with our finding.

1.2 Outbreak in Long Term Care Homes

Some authors suggest the risk of LTC home proposed to the whole population [7], but very little was known whether a outbreak occurs in the subpopulation of LTC homes, due to a dearth of accurate data [4].

In our study, we wish to answer the following question: *Is there a outbreak in the subpopulation of LTC homes prior to the outbreak of the whole population ?* Thereafter, we will use H_0 to denote the hypothesis that such outbreak does not occur, latter we will show that H_0 does not hold.

Showing H_0 is not true can help researchers to answer other critical questions about CoViD-19. In [5], the author estimated the mortality rate of CoViD-19 using historical data and implicitly assumed H_0, which leads to potential inaccuracy as discussed in section 4; with our finding, a better approach may be adapted by identifying outbreak of the pandemic in each subpopulation.

2 Data

Due to the nature of our study, publicly available data can be directly used for our model is very rare, so a composed data set obtained from multiple sources undergoing extensive data cleaning procedures is used. The composed data set is a daily record of number of deaths due to CoViD-19 from April 5^{th} to December 5^{th} in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}^{C}_{\text{LTC}}$ respectively; the data set is summarized in Figure 1.

2.1 Data Sources

The data set is composed from three different sources, which are the daily cumulative death number in Ontario due to CoViD-19 given in [8]; the cumulative death number in LTC homes due to CoViD given in [9]; the daily death number obtained from daily CoViD report published by the Ontario government. Then, by taking the lag difference in the cumulative data sets we obtained the daily death number; finally, daily death number of non-LTC residents are obtained by taking the difference between the total daily death number and LTC residents death number due to CoViD-19.

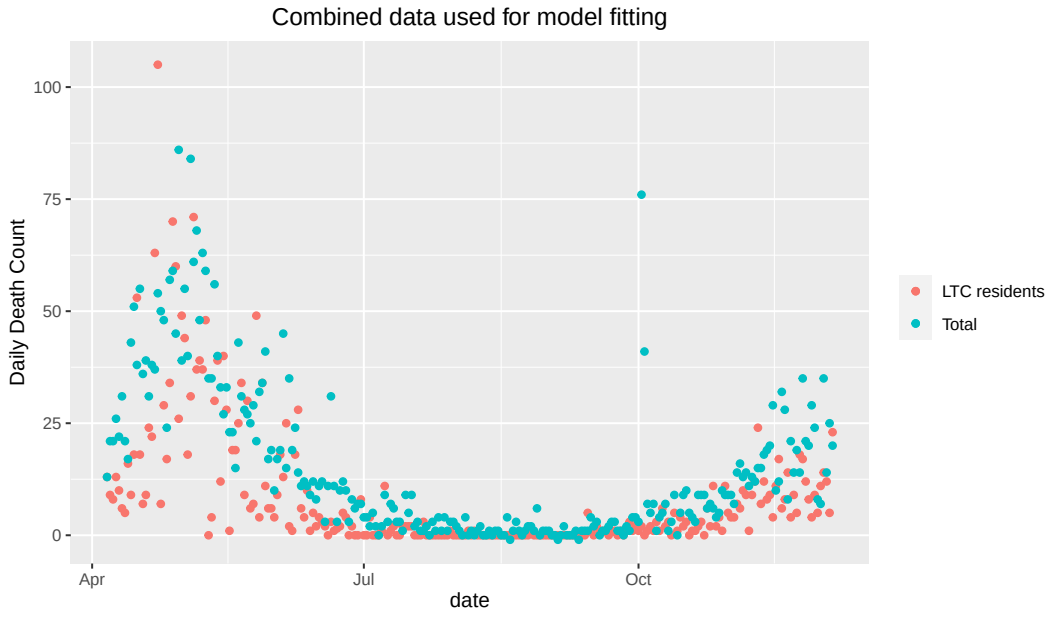

Figure 1: Combined data set obtained from multiple sources we used

2.2 Data Cleaning

The source data sets used have the problem of incompleteness and inconsistencies. Daily death count in LTC homes prior to April 25th is not given in [9], so such information is obtained from cite3. Notwithstanding possible inconsistencies introduced by composing two sources, daily death count in LTC homes in April is of critical importance for the understanding of first wave of pandemic outbreak in Canada; therefore, such data cleaning procedure is justified.

Data publishers often correct their result in the latest data count without changing the errors in the previous published data, so we have encountered negative daily death number when using lag difference on the cumulative data set ($n = 3$); such negative death counts are replaced by the mean of the death counts in two consecutive dates.

The daily death count of people not in LTC homes are obtained by taking difference between data from multiple sources, which also introduced negative count numbers ($n = 40$); we tried to treat them as 0 and missing, the difference in the resulting model is discussed in the section 4.

The daily death count of people not in LTC homes on October 2nd and 3rd are found to be 74 and 41 respectively, which are considered to be outliers and hence removed for model fitting.

3 Methods

In this study, we will use generalized addictive model (GAM) to model CoViD death data, which will be introduced in this section.

3.1 Negative Binomial GAM Model

Let $Y_{i,t}$ be the number of death in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}^C_{\text{LTC}}$ for $i = 0, 1$ respectively. It's a common to model the daily death count using Poisson model, but we do not assume that the variance and mean are the same in our death data, hence a negative binomial model is used.

We assume $Y_{i,t}$ has the following hierarchical structure:

$$\begin{aligned} Y_{i,t} &\sim \text{NB}(\lambda_{i,t}, \kappa) \\ \log(\lambda_{i,t}) &= \mu + I_{i=0}\beta_1 + U_i(t; \alpha_i), \end{aligned} \tag{1}$$

where U_i are the non-parametric smoothing functions that are estimated using `MGCV` package [10]; $\kappa, \mu, \lambda_{i,t}, \beta_1, \alpha_i$ are unknown parameters estimated using maximum likelihood method.

In the model (1), the parameter $\lambda_{1,t}$ and $\lambda_{0,t}$ described the daily death rate in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}^C_{\text{LTC}}$ at time t. We assume at any fixed time t_0, the death rate in two subpopulation is roughly differed by a scalar factor due to epidemiological characteristic difference and is captured by the parameter β_1 and the difference between U_1 and U_0.

4 Results and Discussions

4.1 Results

Let $D_1(t)$ and $D_2(t)$ be the time series of death count in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}^C_{\text{LTC}}$ respectively. To examine our hypothesis H_0, We fit the model (1) with `gam` function in `MGCV` [10] and summarized the fitted daily death time series D_1 and D_2 in Figure 2.

As can be seen from Figure2, both time series have similar pattern with respect to time: both D_1 and D_2 reached a peak at May, had a relative flat curve during August to October and started to increase again after September. However, deaths of the first outbreak between April and June most came from $\mathcal{P}_{\text{LTC}}$ but this is not true for the second outbreak starting from October.

To further examine H_0, we have plotted the normalized difference of fitted time series D_1 and D_2 after the log transformation and summarized in Figure 3. It is clear from Figure 3 that $\log(D_1) - \log(D_2)$ is not a constant time series.

We remark that D_1 and D_2 are similar in scale, which implies the mortality rate of $\mathcal{P}_{\text{LTC}}$ is much higher than that of $\mathcal{P}^C_{\text{LTC}}$ due to difference in the population size, and this is consistent with the results obtained from other studies [2].

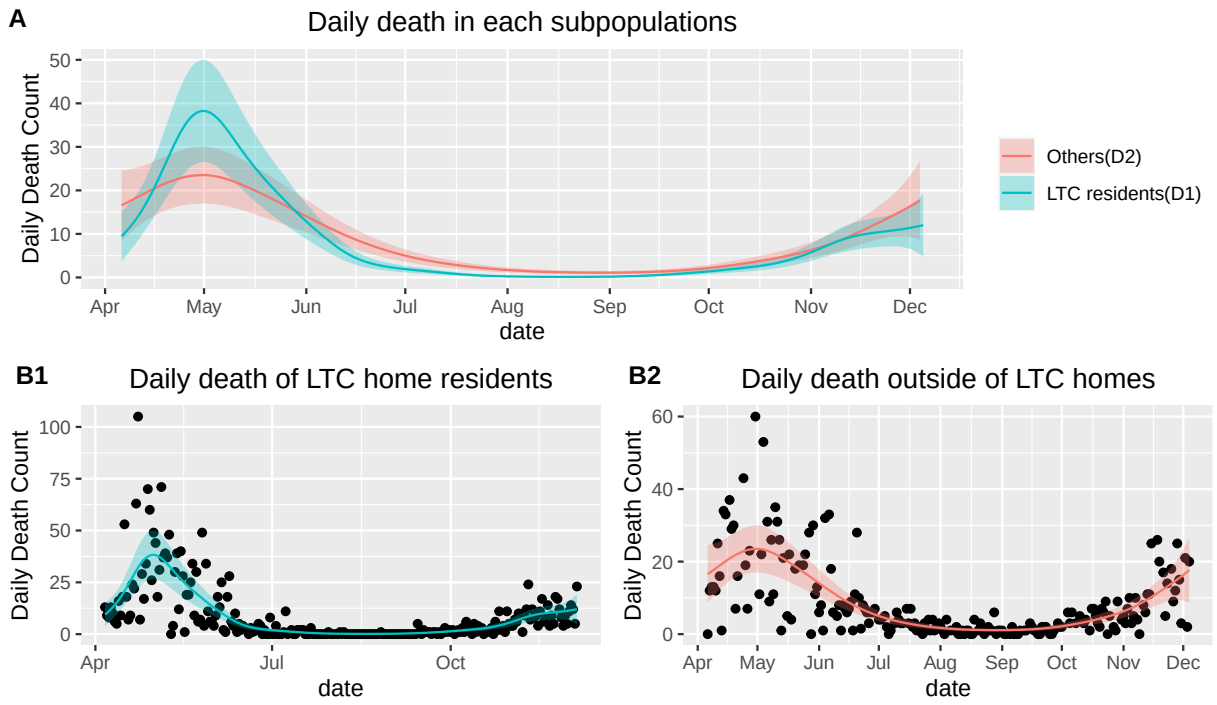

Figure 2: Fitted D_1 and D_2 using Gam model. Fitted time series is plotted using solid line with shaded confidence interval.

A: Fitted D_1 and D_2 with confidence interval

B1, **B2**: Fitted D_1 and D_2 with observed data in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}_{\text{LTC}}^{C}$ as a reference.

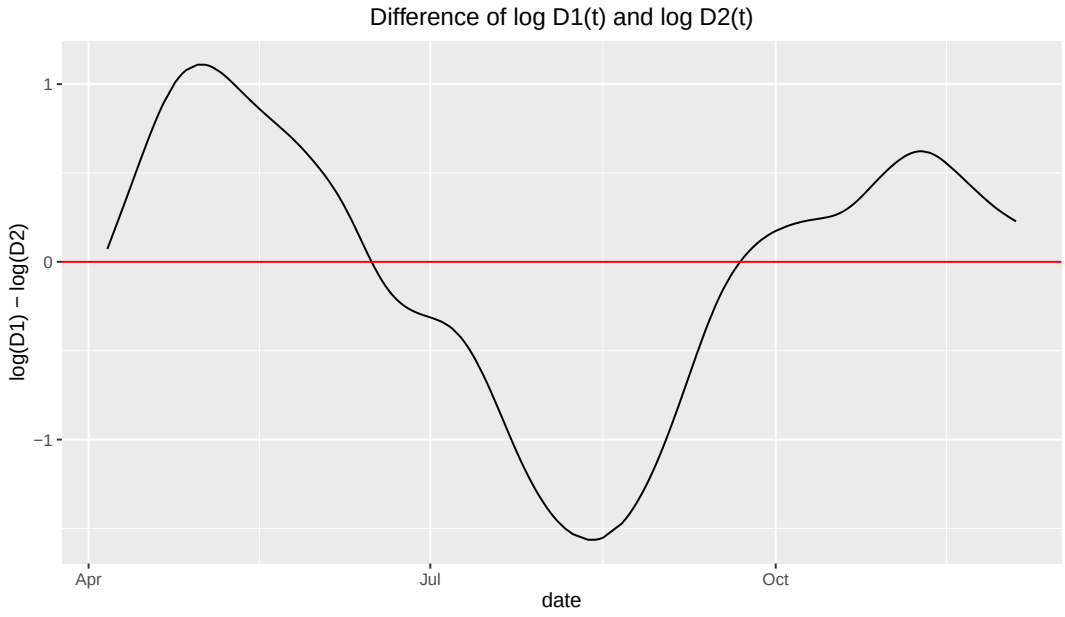

Figure 3: The normalized difference between $\log\left(\widehat{D}_1\right)$ and $\log\left(\widehat{D}_2\right)$, where $\widehat{D}_1$ and $\widehat{D}_2$ are fitted time series using model (1). Note that we have normalized the difference so the plotted time series is centered at 0

4.2 Discussions

4.2.1 Evidence Against H_0

Under H_0, the ratio time series $D_1(t)/D_2(t)$ time series would be a constant, if we reasonably assume the epidemiological characteristic of $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}_{\text{LTC}}^C$ does not change over time. It can be seen from Figure 2 **A** and Figure 3 that H_0 can not be true. In particular, if H_0 is true, we would have observed a horizontal line around $y = 0$ in Figure 3. So, we conclude that, there are two outbreaks occurred at different times, in the different subpopulations $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}_{\text{LTC}}^C$.

4.2.2 Suggestions for Policy Makers

By noting outbreaks occur in different subpopulations, some important implications and suggestions can be made for policy makers. One should be careful when estimating the mortality rate of CoViD-19, since outbreaks in different time may occur in the different subpopulations, then the estimated mortality rate from the historical data may not be a good predictor for future outbreaks.

For the reopening process in the future, restrictions on facilities with high risk population like schools and nursing home may be extended to reduce the chance and costs of outbreaks in such population. Finally, by closely monitoring and vaccinating the critical subpopulations like children in the day care and residents in the nursing home, we may effectively control the scale of the outbreak in the whole population.

We further remark that, multiple government agencies are publishing similar but incomplete and often conflicting data about CoViD-19, which jeopardized practical researches on the pandemic, as has pointed out in [5] and [4]. Hence, we urge the government to ensure departmental and interdepartmental communication so that a more consistent and comprehensive data source will be available for future researches.

4.2.3 Limitations

Despite its significance, our study have some major limitations. The final data set used is composed from multiple sources, which introduced a lot of inconsistencies besides the errors presented in each source. In particular, the daily death counts of $\mathcal{P}_{\text{LTC}}$ in April are obtained from daily reports [9] published and are more error-prone than other sources for two reasons: the short time between obtaining and publishing data for the data publisher makes it difficult to check and correct mistakes; data publisher would not correct published daily report, since mistakes can be corrected in the future report.

The generalized addictive model is used to model the time series D_1 and D_2, which is a trade off between explainability and flexibility. Gam models are semi-parametric, so it's more susceptible to overfit the observed data than the parametric model. Overfitting might be eased with Jackknife method; we did not do so since forecasting is not of the central purpose of our model.

Our model assumed the death rate in $\mathcal{P}_{\text{LTC}}$ and $\mathcal{P}\text{LTC}^C$ on average is different by a factor of $\exp\{\beta_1\}$, which is a fixed parameter to be estimated. However, $\mathcal{P}\text{LTC}$ and $\mathcal{P}\text{LTC}^C$ are vastly different in population size and so β_1 are difficult to interpret.

4.2.4 Future Work

Our method can be generalized in several ways. We have restricted ourselves to the pandemic occurs in Ontario, but similar method can be applied to the data for other regions and a comparison study may be conducted. In our model, the time series D_1 and D_2 does not explicitly depend on each other, which might not be realistic, so for future work a model incorporate the knowledge of corresponding subpopulations can be used. Since residents in LTC homes are often of high risk of CoViD-19, so future study can incorporate information such as age and social economic status in the model to reduce the possible confounding effect.

References

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[2] H. D'Adamo, T. Yoshikawa, and J. G. Ouslander, "Coronavirus disease 2019 in geriatrics and long-term care: the abcds of covid-19," *Journal of the American Geriatrics Society*, vol. 68, no. 5, pp. 912–917, 2020.

[3] "A guide to the long-term care homes act, 2007 and regulation 79/10." [Online]. Available: http://www.health.gov.on.ca/en/public/programs/ltc/docs/ltcha_guide_phase1.pdf

[4] A. T. Hsu, N. Lane, S. K. Sinha, J. Dunning, M. Dhuper, Z. Kahiel, and H. Sveistrup, "Understanding the impact of covid-19 on residents of canada's long-term care homes—ongoing challenges and policy responses," *Londres (Royaume-Uni): International Long-Term Care Policy Network*, vol. 24, 2020.

[5] M. Liu, C. J. Maxwell, P. Armstrong, M. Schwandt, A. Moser, M. J. McGregor, S. E. Bronskill, and I. A. Dhalla, "Covid-19 in long-term care homes in ontario and british columbia," *CMAJ*, vol. 192, no. 47, pp. E1540–E1546, 2020.

[6] (2020, Apr) Covid-19 action plan: long-term care homes. version 1. [Online]. Available: https://www.ontario.ca/page/covid-19-action-plan-long-term-care-homes

[7] W. Gardner, D. States, and N. Bagley, "The coronavirus and the risks to the elderly in long-term care," *Journal of Aging & Social Policy*, pp. 1–6, 2020.

[8] (2020, Dec). [Online]. Available: https://covid-19.ontario.ca/data#testing-data-reporting-sources

[9] (2020, Mar) Daily epidemiologic summary. [Online]. Available: https://covid-19.ontario.ca/covid-19-epidemiologic-summaries-public-health-ontario

[10] S. Wood and M. S. Wood, "Package 'mgcv'," *R package version*, vol. 1, p. 29, 2015.