# Identifying Risk: Predictive Model of Suicide Among Toronto's Homeless*

Hailey Jang

2024-03-16

This study unveils a novel risk stratification model to assess suicide risks among Toronto's homeless population, using sociodemographic data within a generalized linear model. Our analysis, grounded in data from Toronto Public Health, reveals critical patterns and predictors of suicidal behaviour, providing pivotal insights for intervention strategies. The findings highlight the urgent need for targeted preventive measures, emphasizing the model's potential in shaping public health policies and social services, thereby contributing significantly to mitigating suicide risks in marginalized communities.

## Table of contents

**1 Introduction** **2**

**2 Data** **3**
2.1 Data Source . . . . . 3
2.2 Data Measurment . . . . . 3
2.3 Data Summary . . . . . 3

**3 Model** **4**
3.1 Model Set-up . . . . . 4
3.2 Model Equation . . . . . 4
3.3 Model Justification . . . . . 6

**4 Results** **6**
4.1 Model Coefficeints Interpretation . . . . . 6
4.2 Model Visualization . . . . . 7

*Code and data are available at: https://github.com/Hailey-Jang/Suicide-and-Homelessness.git

**5 Discussion** **8**
5.1 Overview of the Study . . . . . 8
5.2 Reflection on the Study . . . . . 9

**References** **9**

# 1 Introduction

Homelessness represents a significant public health crisis, with the intersection of inadequate housing and mental health issues exacerbating the risk of suicide among this population. Individuals experiencing homelessness face a 2- to 6-fold increased risk of suicide compared to the general population (Sinyor et al. (2017)). This alarming statistic not only highlights the vulnerability of homeless individuals but also underscores the urgent need for targeted interventions and preventive measures. This paper delves into the development of a risk stratification model designed to identify homeless individuals at an elevated risk of suicide. By integrating sociodemographic data within a generalized linear model (Bayesian logistic regression model) focusing on sex and age as primary factors, our research aims to provide a predictive tool for healthcare providers and social services. The motivation behind this study stems from the critical gap in existing research regarding effective, data-driven strategies for suicide prevention in the homeless community.

Our research methodology involved the analysis of comprehensive data collected by Toronto Public Health, starting from January 2017. This dataset not only provided a foundation for our model but also enriched our understanding of the prevalence and causes of suicide among the homeless. The outcomes of our study revealed significant patterns and predictors that can enhance the effectiveness of suicide prevention strategies, offering a new perspective on addressing this public health challenge. By identifying high-risk individuals, our model can inform targeted interventions, potentially saving lives and allocating resources more efficiently.

The structure of the paper is organized as follows: Section 2 (Data) delves into the broader context of the dataset, emphasizing the crucial aspects of measurement relevant to our study. Section 3 (Model) elucidates the setup and justification of our chosen logistic regression model, ensuring clarity and transparency of our methodology. Subsequently, Section 4 (Results) presents the coefficients of the regression model alongside detailed statistical analyses. Lastly, Section 5 (Discussion) articulates the study's contributions to our understanding of homelessness and suicide and outlines directions for future research, ensuring a thorough contemplation of the study's broader implications.

# 2 Data

## 2.1 Data Source

The study was performed using data from the City of Toronto's database portal,(Gelfand 2020), accessed through the 'opendatatoronto' package and processed using the statistical programming environment R (R Core Team 2023). The tidyverse (Wickham et al. 2019) package facilitated the data and the ggplot (Wickham 2016), knitr (Xie 2014), readr (Wickham, Hester, and Bryan 2022) and tibble (Müller and Wickham 2022) package was utilized for enhancing table presentations. Specific to this study, the kableExtra (Zhu 2021) package was implemented in the R markdown setting to ensure stable positioning of figures and tables.

## 2.2 Data Measurment

Initiated in January 2017, Toronto Public Health (TPH) embarked on a systematic record-keeping of homeless mortality to gain a clearer understanding of the prevalence and causative trends of these incidents. The dataset comprises variables such as the year of death, cause of death, age group, gender, and number of deaths. It classifies individuals into age categories spanning 20 years, starting from 20 years to 60+ years, with gender recorded as either Male or Female. The dataset enumerates the deaths annually from 2017 through 2023, categorizing the causes into distinct classifications like Accident, Drug Toxicity", Suicide, among others.

Focusing on suicide-related fatalities, the dataset was refined to exclude entries marked as Unknown or empty. This filtration led to the construction of two specialized datasets: one delineating the yearly suicide death toll segregated by gender and another by age group. This reorganization necessitated aggregating individual counts from each report into a consolidated figure for these subgroups, facilitating a focused examination of suicide trends across different demographics over the years.

## 2.3 Data Summary

Table Table 1 presents the refined data, specifically spotlighting instances of suicide by age groups, and depicting the numerical specifics of each case.

Table Table 2 presents the refined data, specifically spotlighting instances of suicide by gender, and depicting the temporal and numerical specifics of each case.

Table 2: Suicides Rates by Gender

| Year | sex | count |
|---|---|---|
| 2017 | Male | 69 |
| 2018 | Male | 65 |

| | | |
|---|---|---|
| 2019 | Male | 87 |
| 2020 | Male | 107 |
| 2021 | Male | 153 |
| 2022 | Male | 129 |
| 2023 | Male | 64 |
| 2017 | Female | 21 |
| 2018 | Female | 19 |
| 2019 | Female | 34 |
| 2020 | Female | 27 |
| 2021 | Female | 45 |
| 2022 | Female | 36 |
| 2023 | Female | 8 |

# 3 Model

## 3.1 Model Set-up

This section elucidates the development of a logistic regression model, tailored to predict the likelihood of suicide among homeless individuals using demographic factors. Before delving into the model, we visually explore the relationship between suicide rates and demographic variables—age and gender—using ggplot2.

We employ Figure 1 to create a comprehensive visual representation, showcasing the relationship between suicide rates and the key demographic variables, age and gender. This visualization aids in understanding the data distribution and any apparent trends that could influence the model.

## 3.2 Model Equation

The logistic regression model is designed to predict the probability of suicide cases among homeless individuals, utilizing sociodemographic factors such as age and gender. In a Bayesian logistic regression context, the model's setup can be represented with the following hierarchical structure:

Table 1: Suicides Rates by Age Group

[H]

| Age_group | Total Counts |
|---|---|
| 40-59 | 376 |
| 60+ | 255 |
| 20-39 | 225 |
| <20 | 8 |

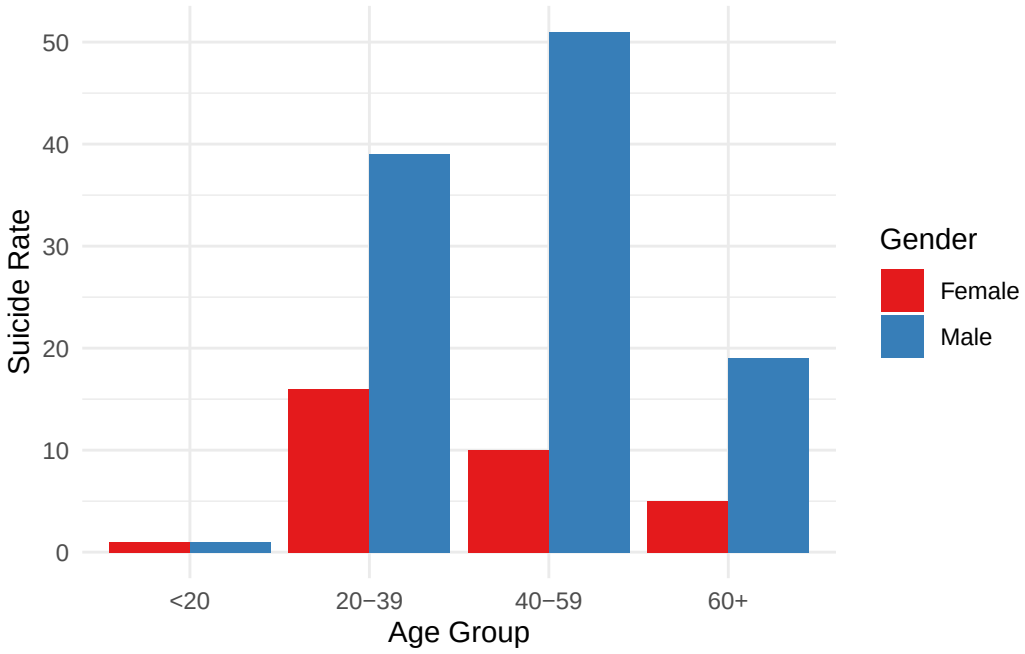


Figure 1: Summary of Suicide Rates by Age and Gender

$$y_i | p_i \sim \text{Bernoulli}(p_i) \tag{1}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 \times \text{AgeGroup}_{i1} + \beta_2 \times \text{AgeGroup}_{i2} + \beta_3 \times \text{Gender}_i \tag{2}$$

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_j \sim \text{Normal}(0, 2.5) \text{ for } j = 1, 2, 3 \tag{4}$$

$$\tag{5}$$

In this adapted model:

- $y_i$ represents the binary outcome for each individual (i.e., the presence or absence of a suicide case).
- $p_i$ is the probability of observing a suicide case for the individual, linked to the predictors through the logistic function.
- The coefficients $\alpha$,$\beta_1$, $\beta_2$, and $\beta_3$ are assigned Normal prior distributions, reflecting our prior beliefs about these parameters' distributions before observing the data. The Normal priors are centered at 0 with a standard deviation of 2.5, indicating moderate certainty in the prior information.
- The logit link function (log-odds) is the natural logarithm of the odds $p_i$ / $1 - p_i$ and linearly relates the predictors to the probability of the outcome.

### 3.3 Model Justification

The Bayesian logistic regression model is adeptly suited for the paper's objective of creating a risk stratification tool to identify homeless individuals at elevated suicide risk. It judiciously integrates sociodemographic factors like age and gender, utilizing a Bayesian framework to incorporate prior knowledge and present data, thus offering a robust predictive analysis. The model's hierarchical structure and the inclusion of Normal priors for coefficients ensure a nuanced analysis, balancing prior beliefs with empirical data, crucial for strategizing suicide prevention efforts among the homeless.

## 4 Results

### 4.1 Model Coefficeints Interpretation

Table 3 provides a detailed summary of the estimated effects of age groups and gender on the suicide risk among homeless individuals, as derived from the Bayesian regression analysis. Each row in the table represents a different predictor in the model, with the coefficients indicating the magnitude and direction of the association between that predictor and the observed counts of

suicide cases. For instance, if the coefficient for a certain age group is positive and statistically significant, it suggests that individuals in this age group have a higher risk of suicide compared to the baseline group, after controlling for other factors in the model. Conversely, a negative coefficient would suggest a lower risk. Similarly, the gender coefficients shed light on the differential risk of suicide between genders.

Table 3: Summary of Residuals for the Bayesian Model

| Observation | Residual |
|---|---|
| 1 | -3.542 |
| 2 | -1.788 |
| 3 | -1.178 |
| 4 | -0.522 |
| 5 | -3.542 |
| 6 | -0.788 |
| 7 | -0.178 |
| 8 | -4.650 |
| 9 | -1.542 |
| 10 | 0.212 |

Table 3 is crucial for justifying the thesis of the paper as it provides empirical evidence supporting the hypothesis that certain age groups or genders are at a higher risk of suicide among homeless individuals. By quantifying these associations, the model helps in identifying high-risk subgroups within the homeless population, thereby offering insights that can inform targeted interventions and preventive strategies. The Bayesian framework further enriches this analysis by incorporating prior knowledge and uncertainty into the estimation process, thereby enhancing the robustness and interpretability of the findings.

## 4.2 Model Visualization

Figure 2 effectively illustrates the relationship between demographic factors—specifically age groups and gender—and the probability of suicide among homeless individuals. This graphical representation is pivotal in understanding how these variables interplay to influence suicide risk.

Each point on the plot corresponds to a specific age group and is color-coded to represent gender, showcasing a nuanced view of how suicide counts vary across different demographic segments. The size of each point is proportional to the count of suicide cases, providing an immediate visual cue to the relative magnitude of suicide risk within each category. This size differentiation helps in identifying which combinations of age group and gender have higher or lower counts of suicide incidents, thereby indicating potential risk stratification within the homeless population.

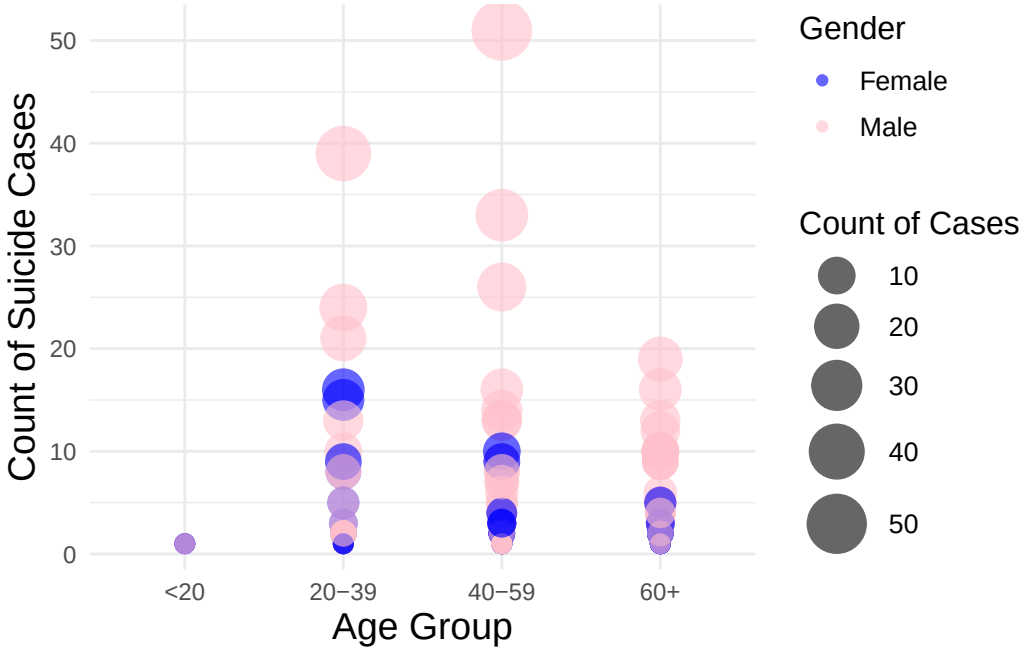


Figure 2: Suicide Cases by Age Group and Gender

# 5 Discussion

## 5.1 Overview of the Study

This research delved into the complex interplay between sociodemographic factors and the incidence of suicide within the homeless population of Toronto. Through a detailed regression analysis, we explored how variables such as age, gender, and time-related elements influence suicide rates, uncovering specific patterns and insights pivotal to understanding this critical public health concern. The precision of our model, grounded in solid statistical practices, offered an in-depth view of the demographic determinants of suicide risk among the homeless.

A key discovery of our analysis was the pronounced impact of gender and age on suicide risk, with the model indicating that males in the 40-59 age group are at a higher risk compared to other demographics. This group emerged as particularly susceptible, signaling an urgent need for targeted interventions and support mechanisms. Additionally, the model revealed that younger males, specifically those aged 20-39, also face a heightened risk of suicide attempts, pointing towards the necessity for early preventive measures tailored to this subgroup.

Contrastingly, the results showed that females generally have lower instances of suicide attempts, a finding that, while reassuring, also calls for a nuanced understanding of gender-specific vulnerabilities and protective factors in the context of homelessness.

Moreover, our analysis shed light on the changing trends in suicide rates over time, providing insights into the fluctuating dynamics of this issue. Such temporal insights are vital for adapting intervention strategies and policy formulations to the evolving landscape of societal, economic, and health-related influences affecting the homeless population.

These findings not only emphasize the importance of considering age and gender in suicide prevention strategies but also highlight the necessity for dynamic, evidence-based approaches to support Toronto's homeless community, ultimately aiming to reduce the prevalence of this tragic outcome.

### 5.2 Reflection on the Study

Our investigation yields valuable findings, yet it is important to acknowledge its limitations. The study's dependency on historical data, coupled with the intrinsic limitations associated with regression analysis, may have impacted the depth of our insights. The possibility of data reporting biases, along with the difficulty in fully encapsulating the complex realities of homelessness, might limit how broadly our conclusions can be applied. Moreover, the snapshot nature of our data constrains our capacity to establish causative links between the sociodemographic variables studied and the instances of suicide.

Looking ahead, addressing this intricate issue demands a comprehensive strategy. It's imperative for upcoming research to incorporate longitudinal designs, enabling a clearer understanding of the causative dynamics and enduring patterns connecting homelessness with suicide risks. This approach will enhance our grasp of the temporal aspects of these factors and their long-term consequences, paving the way for more effective interventions and policy-making.

## References


Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Müller, Kirill, and Hadley Wickham. 2022. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sinyor, Mark, Nicole Kozloff, Catherine Reis, and Ayal Schaffer. 2017. "An Observational Study of Suicide Death in Homeless and Precariously Housed People in Toronto." *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie* 62 (May): 706743717705354. https://doi.org/10.1177/0706743717705354.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.