Hospitalization Prediction Modeling for COVID

Alexander Ruse, Luke Wilsen, and Ben Gerber

Summary

Hospitals in the U.S. were overrun during the COVID pandemic due to limited resources and increased burden of illness. It is important for public health and healthcare organizations to prioritize care (using hospital beds for severe cases). An accurate prediction model may help identify those at high risk and need for hospitalization. This may improve resource efficiency.

The *primary outcome* of interest is hospitalization (binary).

The *research question* is: can we predict the risk of hospitalization for COVID based on demographic and clinical factors, including:

- Age group
- Sex
- Race
- Ethnicity
- Exposure
- Symptom Status
- Underlying Conditions

The *Bayesian models* we plan to produce include:

- Bayesian Logistic Regression
- Bayesian Hierarchical Model (By State)

Given the significant amount of missing data in reports, we will explore various methods including multiple imputation (e.g., mice package) and/or `brm_multiple` function for multiple imputed data sets (see vignette).

Example resources:

Bayesian Hierarchical Spatial Model to Correct for Misreporting in Count Data: Application to State-Level COVID-19 Data in the United States (Chen, Song, and Stamey 2022)

Bayesian Hierarchical models incorporating study-level covariates for multivariate meta-analysis of diagnostic tests without a gold standard with application to COVID-19 (Wang et al. 2023)

A Bayesian hierarchical model for estimating the statistical parameters in a three-parameter log-normal distribution for monthly average streamflows (Li, Zhou, and Yeh 2020)

Data

We will use data from the CDC available here. This includes case surveillance (public use) with for all COVID-19 cases shared with the CDC, including demographics, geography (e.g., state), exposure history, disease severity indicators, and outcomes (e.g., hospitalization). It also includes underlying medical conditions. Reporting of data was discontinued July 1, 2024.

```r
library(RSocrata)
library(tidyverse)

# CDC dataset with geographic data (this one includes state/county and person-level)
df <- read.socrata(
  url = "https://data.cdc.gov/resource/n8mc-b4w4.json?$limit=1000"
)

# Data structure
glimpse(df)
```

```
Rows: 1,000
Columns: 19
$ case_month               <chr> "2021-10", "2022-02", "2020-09", "2021-10", "~
$ res_state                <chr> "NC", "GA", "MO", "MI", "WI", "IN", "AL", "GA~
$ state_fips_code          <chr> "37", "13", "29", "26", "55", "18", "01", "13~
$ res_county               <chr> "DAVIE", "BULLOCH", "POLK", "SANILAC", "OZAUK~
$ county_fips_code         <chr> "37059", "13031", "29167", "26151", "55089", ~
$ age_group                <chr> "0 - 17 years", "18 to 49 years", "18 to 49 y~
$ sex                      <chr> "Female", "Female", "Female", "Female", "Fema~
$ race                     <chr> "NA", "Unknown", "NA", "NA", "NA", "NA", "Mis~
$ ethnicity                <chr> "NA", "Missing", "NA", "NA", "NA", "NA", "Mis~
$ case_positive_specimen   <chr> "0.0", NA, "0.0", NA, NA, "0.0", NA, NA, NA, ~
$ process                  <chr> "Missing", "Missing", "Missing", "Missing", "~
$ exposure_yn              <chr> "Missing", "Missing", "Missing", "Missing", "~
$ current_status           <chr> "Laboratory-confirmed case", "Laboratory-conf~
$ symptom_status           <chr> "Unknown", "Symptomatic", "Symptomatic", "Mis~
$ hosp_yn                   <chr> "Unknown", "Missing", "Unknown", "Missing", "~
$ icu_yn                    <chr> "Unknown", "Missing", "Missing", "Missing", "~
$ death_yn                  <chr> "No", "Missing", "Unknown", "Unknown", "Unkno~
$ case_onset_interval       <chr> NA, "0.0", "0.0", NA, NA, NA, NA, NA, "0.0", ~
$ underlying_conditions_yn <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

Action plan

We will consider a "Bayesian Workflow" (Gelman et al. 2020) with an iterative process, and will build a model, evaluate (diagnostics), and try to improve and expand the model in different ways.

1. We will clean, recode, and explore the data. This includes descriptive analyses (frequencies, summary statistics), histograms, scatterplots, and related covariates to outcomes.
2. We will explore missingness of data. This includes amount of missingness, and pattern of missingness (is it at random?).
3. We will establish an initial model (will start simple and expand). This will include those covariates deemed worthy based on missingness as well as relationship with outcome. A potential candidate model may include:

Let Y_i denote the binary outcome (hospitalization), where:

$$Y_i = \begin{cases} 1 & \text{if the individual is hospitalized,} \\ 0 & \text{otherwise.} \end{cases}$$

and:

$$Y_i | p_i \sim \text{Bern}(p_i)$$

Let $\mathbf{X}_i = (X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i})$ include all of the covariates for individual i, where:

- $X_{1,i}$: Underlying condition
- $X_{2,i}$: Age group
- $X_{3,i}$: Exposure
- $X_{4,i}$: Symptoms

Then the posterior distribution would resemble:

Our logistic Bayesian model would then be:

$$\text{logit}(p_i) = \log\left(\frac{P(Y_i = 1|\mathbf{X}_i)}{1 - P(Y_i = 1|\mathbf{X}_i)}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i}$$

We will use normal priors for all covariates, and then re-visit the model to experiment with other generic priors (i.e. a student t(3, 0 ,1) preferred by Aki Vehtari (https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations))

The prior distribution would assume normal for all covariates:

$$\beta_j \sim \mathcal{N}(0, \sigma^2) \quad \text{for} \quad j = 0, 1, 2, 3, 4$$

Note this initial model does not include state for hierarchical expression.

4. We will pursue prior predictive checks.
5. We will validate results and address any issues that arise. we will try to fit fast and fail fast. We will explore MCMC diagnostics. If there are computational issues, will consider subsetting data or using specific date range from CDC.
6. We will modify the model and explore additional situations. For example, we will consider the hierarchical model (state as group) once we have confidence based on other covariates and models.
7. We will validate the model using LOO internally. In addition, we will assess the model's performance on new (unseen) data, but using a different date range of data.
8. If time allows, we will consider interaction terms in the model and compare its performance. We will also investigate performance of prediction across different states, particularly those with less cases reported. Other investigations have been considered, depending on time, just for the additional experience.
9. Our group will work together in parallel with shared data and code in github. This will allow for independent trial/error exploration and comparison of approaches. We will check in with our results each week on Tuesdays and Thursdays. The final product will incorporate our consensus on best model and approach.

References

Chen, Jinjie, Joon Jin Song, and James D. Stamey. 2022. "A Bayesian Hierarchical Spatial Model to Correct for Misreporting in Count Data: Application to State-Level COVID-19 Data in the United States." *International Journal of Environmental Research and Public Health* 19 (6): 3327. https://doi.org/10.3390/ijerph19063327.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." https://doi.org/10.48550/ARXIV.2011.01808.

Li, Jinshu, Qing Zhou, and William W.-G. Yeh. 2020. "A Bayesian Hierarchical Model for Estimating the Statistical Parameters in a Three-Parameter Log-Normal Distribution for Monthly Average Streamflows." *Journal of Hydrology* 591 (December): 125265. https://doi.org/10.1016/j.jhydrol.2020.125265.

Wang, Zheng, Thomas A Murray, Mengli Xiao, Lifeng Lin, Demissie Alemayehu, and Haitao Chu. 2023. "Bayesian Hierarchical Models Incorporating Study-Level Covariates for Multivariate Meta-Analysis of Diagnostic Tests Without a Gold Standard with Application to COVID-19." *Statistics in Medicine* 42 (28): 5085–99. https://doi.org/10.1002/sim.9902.