

UCLA

Research Reports

Title

Pseudo-Likelihood Based Logistic Regression for Estimating COVID-19 Infection and Case Fatality Rates by Gender, Race, and Age in California

Permalink

<https://escholarship.org/uc/item/9hw6s7ks>

Authors

Xiong, Di
Zhang, Lu
Watson, Gregory L
et al.

Publication Date

2021-02-05

Pseudo-Likelihood Based Logistic Regression for Estimating COVID-19 Infection and Case Fatality Rates by Gender, Race, and Age in California

Di Xiong^{a,1}, Lu Zhang^{a,1}, Gregory L. Watson^a, Phillip Sundin^a, Teresa Bufford^a, Joseph A. Zoller^a, John Shamsboian^a, Marc A. Suchard^{a,b,c}, Christina M. Ramirez^{a,*}

^a*Department of Biostatistics, Jonathan and Karen Fielding School of Public Health, University of California, Los Angeles, California, United States of America*

^b*Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America*

^c*Department of Computational Medicine, University of California, Los Angeles, California, United States of America*

Abstract

In emerging epidemics, early estimates of key epidemiological characteristics of the disease are critical for guiding public policy. In particular, identifying high risk population subgroups aids policymakers and health officials in combatting the epidemic. This has been challenging during the coronavirus disease 2019 (COVID-19) pandemic, because governmental agencies typically release aggregate COVID-19 data as marginal summary statistics of patient demographics. These data may identify disparities in COVID-19 outcomes between broad population subgroups, but do not provide comparisons between more granular population subgroups defined by combinations of multiple demographics.

We introduce a method that overcomes the limitations of aggregated summary statistics and yields estimates of COVID-19 infection and case fatality rates — key quantities for guiding public policy related to the control and prevention of COVID-19 — for population subgroups across combinations of

*Corresponding author at: Department of Biostatistics, Jonathan and Karen Fielding School of Public Health, University of California, Los Angeles, 650 Charles E. Young Drive South, CHS 51-254, Los Angeles, CA 90095

Email address: cr@ucla.edu (Christina M. Ramirez)

¹These authors contributed equally to this work.

demographic characteristics. Our approach uses pseudo-likelihood based logistic regression to combine aggregate COVID-19 case and fatality data with population-level demographic survey data to estimate infection and case fatality rates for population subgroups across combinations of demographic characteristics.

We illustrate our method on California COVID-19 data to estimate test-based infection and case fatality rates for population subgroups defined by gender, age, and race and ethnicity. Our analysis indicates that in California, males have higher test-based infection rates and test-based case fatality rates across age and race/ethnicity groups, with the gender gap widening with increasing age. Although elderly infected with COVID-19 are at an elevated risk of mortality, the test-based infection rates do not increase monotonically with age. LatinX and African Americans have higher test-based infection rates than other race/ethnicity groups. The subgroups with the highest 5 test-based case fatality rates are African American male, Multi-race male, Asian male, African American female, and American Indian or Alaska Native male, indicating that African Americans are an especially vulnerable California subpopulation.

Keywords: COVID-19, Infection Rate, Case Fatality Rate, California Health Interview Survey, Logistic Regression,

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread from its zoonotic origins in Hubei Province, China, causing a global pandemic of coronavirus disease 2019 (COVID-19) [1, 2]. As of June 12, 2020, COVID-19
5 has infected over 7 million people across 188 countries and regions [3]. In the early stages of an emerging epidemic such as COVID-19, estimating the infection rate (IR) and case fatality rate (CFR) of the infectious disease is of utmost importance to health officials, policy makers, and the population at large. Accurate population and subgroup estimates of CFRs provide an evidence-based
10 rationale for policies designed to mitigate the spread of the infectious disease,

help identify disparities in disease vulnerability, and inform resource allocation to communities in greatest need.

Official COVID-19 data released by governmental health agencies and other public sources are prohibited by U.S. law from containing personally identifiable information. Consequently, these data are generally summarized in an aggregate format that comprise only marginal or limited bivariate summary statistics of patient demographics, providing valuable but limited information on the heterogeneity of patient attributes. Indeed, in New York City, the epicenter of the COVID-19 outbreak in the U.S., the reported infection rates and case fatality rates for African Americans were disproportionately higher than other races, according to data released by the New York City Department of Health and Mental Hygiene [4]. Data from several other U.S. states, including New Jersey[5], California[6], and Illinois[7], exhibited similar trends. Gender and age-disaggregated national case data from a vast array of countries across the globe reveal that males and older individuals generally have substantially higher case fatality rates. Furthermore, evidence from numerous clinical studies of COVID-19 risk factors have established that gender and age are risk factors for COVID-19 infection mortality [8, 9, 10, 11]. However, by aggregating, data from governmental health agencies or other public sources do not provide granular information on the combined effect of the risk factors under consideration. In particular, how IRs and CFRs vary across population subgroups characterized by gender, age, and race jointly has not yet received substantial attention. Understanding the gender-age-race dynamics of COVID-19 infection and mortality would provide deeper insights into the disparities that exist in the effects of COVID-19 on the population.

Various methods for using information contained in aggregate data have been proposed in a wide array of applications [see, e.g. 12, 13, 14], and there is growing interest in leveraging marginal summary statistics in publicly released COVID-19 datasets to quantify the impact of various risk factors on COVID-19 mortality [15]. In this paper, we propose a method that helps overcome the limitations of having only aggregate summary statistics on COVID-19 cases

and fatalities to obtain early estimates of COVID-19 IRs and CFRs for population subgroups defined by combinations of risk factors. A major difference between the prevalent approaches to analyzing marginalized data and our proposed method is that we incorporate multivariate population demographic data, which provides estimates of the joint probability distribution of risk factors for the disease. Specifically, we propose a pseudo-likelihood based multivariable logistic regression approach that combines publicly released aggregate COVID-19 case and fatality data with multivariate population-level demographic survey data.

The proposed method is composed of two main steps. First, we model COVID-19 IRs using a multivariable logistic regression model, estimating its parameter values from publicly available COVID-19 case data. Second, we estimate COVID-19 CFRs based on the recovered IRs and publicly available COVID-19 fatality data. This paper uses California as an example case study, but the approach is easily generalized to other states. We carry out the analysis using the most recent COVID-19 case and fatality data from the California Department of Public Health (CDPH) [6, 16, 17] and population-level demographic data from the California Health Interview Survey (CHIS) [18] to obtain estimates of IRs and CFRs for subgroups of the California state population characterized by the joint distribution of gender, age, and race. Not every person who may be infected with COVID-19 is tested, and in some locations only people who are symptomatic are tested. This introduces sampling bias into the COVID-19 data that prevents straightforward estimation of the true IRs and CFRs. To circumvent this issue, we estimate test-based IR (T-IRs) and test-based CFRs (T-CFRs) that depend on the availability and use of testing and may differ from the true IRs and CFRs. In particular, we expect true IRs to be greater than T-IRs and true CFRs to be less than T-CFRs due to the presence of asymptomatic and undiagnosed infections. While the test-based rates do not estimate the overall population rates, they capture the vast majority of severe or fatal COVID-19 infections, because these individuals are very likely to be tested. Consequently, the T-IRs and T-CFRs estimated by our method provide

valuable insights into the disparities in COVID-19 outcomes that exist across gender, age, and race/ethnicity groups and furnish guidance for public policy
75 related to the control and prevention of COVID-19.

2. Data

Our method for estimating COVID-19 T-IRs and T-CFRs relies on two data sources: daily COVID-19 data for California from the California Department of Public Health (CDPH), and the 2017–2018 wave of the California Health
80 Interview Survey (CHIS).

CDPH data are publicly available and provide up-to-date information on the number of COVID-19 cases and fatalities in California by gender [16], age [17], and race/ethnicity [6], separately. The case and fatality data as of June 12, 2020 are presented in Table 1. CDPH divides the population into ten age
85 groups: less than 5, 5–17, 18–34, 35–49, 50–59, 60–64, 65–69, 70–74, 75–79, and 80 and above. Age group is missing from less than 0.1% of the confirmed cases and deaths reported by CDPH. The eight race and ethnicity groups in the publicly released dataset are LatinX/ Hispanic (LatinX), White/ Caucasian (White), Asian, African American/Black (AA), Multi-Race, American Indian or
90 Alaska Native (AIAN), Native Hawaiian and other Pacific Islander, and others. We combined the last two race and ethnicity groups due to their small size in the California population. Race and ethnicity are missing in almost 29% of confirmed cases and 1% of the deaths. The CDPH data also provide the number of COVID-19 cases and fatalities by race and age jointly [6], the only state in
95 the U.S. doing so at the time of writing.

To supplement the CDPH COVID-19 data, we used demographic data on the California population collected by the California Health Interview Survey (CHIS). CHIS is the largest state health survey in the U.S., conducted by the UCLA Center for Health Policy Research in collaboration with the California
100 Departments of Public Health and Health Care Services. CHIS interviews over 20,000 Californians each year, collecting information on a wide range of demo-

Demographic	Group	Confirmed Cases	Deaths	Death Rate (%)
Gender	Male	73,480	2,727	3.71%
	Female	71,501	2,146	3.00%
Age Group	0–17	9,819	0	0.00%
	18–34	41,766	52	0.12%
	35–49	36,675	244	0.67%
	50–59	23,486	439	1.87%
	60–64	9,297	339	3.65%
	65–69	6,735	450	6.68%
	70–74	4,980	506	10.16%
	75–79	3,841	601	15.65%
	80+	8,847	2,247	25.40%
Race & Ethnicity	LatinX	58,415	1,941	3.32%
	White	18,783	1,594	8.49%
	Asian	8,271	709	8.57%
	AA	4,906	458	9.34%
	Multi-race	775	33	4.26%
	AIAN	201	13	6.47%
	Other	12,389	58	0.47%

Table 1: Confirmed COVID-19 cases and fatalities by gender, age and race/ethnicity in California as of June 13, 2020 [16, 17, 6]

graphic and health variables. CHIS oversamples certain population subgroups to achieve more reliable and precise estimates for these subgroups, and estimates a sampling weight for each respondent to represent the reciprocal of the probability of selection. We use the 2017–2018 wave of CHIS in our analysis, which consists of 45,369 subjects interviewed, focusing on the following three demographic variables recorded: gender, age, and race and ethnicity groups.

3. Methods

3.1. Infection Rate Estimation Procedure

We propose estimating COVID-19 T-IRs given gender, age and race using a multivariable logistic regression model. The variables we use in our analysis are listed in Table 2. Letting age 0–17, female, and LatinX be the reference categories, we let $\mathbf{z} \in \{0, 1\}^p \in \mathbb{R}^p$ denote the gender-age-race covariate setting of the covariates $\mathbf{Z}$ in Table 2, where p is 15. The postulated IR model follows

$$\log \left[\frac{\mathbb{P}(\mathcal{I} = 1 \mid \mathbf{z})}{1 - \mathbb{P}(\mathcal{I} = 1 \mid \mathbf{z})} \right] = \gamma_0 + \mathbf{z}^\top \boldsymbol{\gamma}, \quad (1)$$

where $\mathcal{I} \in \{0, 1\}$ represents infection status, γ_0 is the log odds of infection for the female age 0–17 Latinx group, and $\boldsymbol{\gamma} \in \mathbb{R}^p$ are the log odds ratios of infection associated with the other demographic categories.

The CDPH data provide the gender, age, and race distributions of COVID-19 infections separately [16]. To estimate the T-IRs given gender, age, and race jointly, we employ a pseudo-likelihood approach that maximizes a likelihood function constructed from univariate logistic regression models obtained by marginalizing over the covariates. The proposed method begins by first expressing Equation (1) in terms of the probability of infection conditional on the covariates,

$$\mathbb{P}(\mathcal{I} = 1 \mid \mathbf{z}) = \frac{\exp(\gamma_0 + \mathbf{z}^\top \boldsymbol{\gamma})}{1 + \exp(\gamma_0 + \mathbf{z}^\top \boldsymbol{\gamma})}. \quad (2)$$

We introduce $\mathbb{P}_{\mathbf{X}}(\mathbf{x})$ as the probability mass function of a p^* -dimensional discrete random variable $\mathbf{X}$ with support $\mathcal{X} \subset \{0, 1\}^{p^*}$, $p^* > p$, that represents the proportion of the California population with gender-age-race attributes $\mathbf{x}$,

Variable	Definition
Infection Status	Dichotomous outcome indicating COVID-19 infection (0 = No, 1 = Yes)
Gender	Dichotomous covariate indicating male or female (0 = Female, 1 = Male)
Age	Categorical covariate with the following age groups: 0–17 (reference level), 18–34, 35–49, 50–59, 60–64, 65–69, 70–74, 75–79, 80+
Race/Ethnicity	Categorical covariate with the following race categories: LatinX (reference level), White, Asian, African American/Black, Multi-Race, American Indian or Alaska Native, Other

Table 2: Variables used in the infection rate estimation procedure

which is simply an augmentation of the covariate setting $\mathbf{z}$ in (1) to include the reference levels listed in Table 2. In other words, there exists a bijection $\mathbf{z} = \mathbf{z}(\mathbf{x})$ from the $\mathbf{X}$ -space to the $\mathbf{Z}$ -space. We then define the conditional probability mass function of $\mathbf{X}_{-i}$ given $\mathbf{X}_i = 1$ to be $\mathbb{P}_{\mathbf{X}_{-i}|\mathbf{X}_i}(\mathbf{x}_{-i}|x_i = 1)$, where X_i is the i^{th} element of $\mathbf{X}$, and $\mathbf{X}_{-i}$ is the subset of $\mathbf{X}$ that omits X_i . Defining $\mathcal{X}_{(X_i=1)}$ to be the subset of $\mathcal{X}$ with the constraint that $X_i = 1$ and taking the expectation of both sides of Equation (2) conditional on $X_i = 1$, by the Law of Iterated Expectations we have

$$\begin{aligned}
\mathbb{P}(\mathcal{I} = 1 \mid X_i = 1) &= \mathbb{E}_{\mathbf{X}_{-i}|\mathbf{X}_i=1} [\mathbb{P}(\mathcal{I} = 1|\mathbf{z})] \\
&= \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \mathbb{P}(\mathcal{I} = 1|\mathbf{z}(\tilde{\mathbf{x}})) \mathbb{P}_{\mathbf{X}_{-i}|\mathbf{X}_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1) \\
&= \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \frac{\exp(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{z}(\tilde{\mathbf{x}}))}{1 + \exp(\gamma_0 + \boldsymbol{\gamma}^\top \mathbf{z}(\tilde{\mathbf{x}}))} \mathbb{P}_{\mathbf{X}_{-i}|\mathbf{X}_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1) .
\end{aligned} \tag{3}$$

Next, we construct the individual pseudo-log-likelihoods corresponding to each univariate logistic regression of $\mathcal{I}$ on $X_i = 1$ for each $X_i \in \mathbf{X}$. Let N

denote the total population size, N_{i1} denote the number of individuals in the population with $X_i = 1$, and $N_{i1}^{(\mathcal{I})}$ denote the total number of individuals with $X_i = 1$ who have been or will be infected with COVID-19. Therefore, $N_{i1}^{(\mathcal{I})}$ follows a binomial distribution,

$$\text{Binomial} \left(N_{i1}, \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \frac{\exp(\gamma_0 + \gamma^\top \mathbf{z}(\tilde{\mathbf{x}}))}{1 + \exp(\gamma_0 + \gamma^\top \mathbf{z}(\tilde{\mathbf{x}}))} \mathbb{P}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1) \right) \quad (4)$$

for $i = 1, \dots, p^*$. We define the individual pseudo-log-likelihood of (γ_0, γ) for X_i corresponding to the binomial distribution (4) as $\mathcal{L}_{X_i}^{(\mathcal{I})}(\gamma_0, \gamma | N_{i1}^{(\mathcal{I})}, N_{i1})$, and we define the full pseudo-log-likelihood of (γ_0, γ) as the sum of the individual pseudo-log-likelihoods

$$\mathcal{L}(\gamma_0, \gamma) = \sum_{i=1}^{p^*} \mathcal{L}_{X_i}^{(\mathcal{I})}(\gamma_0, \gamma | N_{i1}^{(\mathcal{I})}, N_{i1}). \quad (5)$$

We use the CHIS data to approximate $\mathbb{P}_{\mathbf{X}}(\mathbf{x})$, which we denote $\hat{\mathbb{P}}_{\mathbf{X}}(\mathbf{x})$. Let $N^{(\mathcal{I})}$ denote the total number of individuals in the population who have
115 been or will be infected with COVID-19, and let $\pi_{\mathcal{I}} = \mathbb{P}(\mathcal{I} = 1)$ denote the overall infection rate in the population. Thus, the total population size is $N = N^{(\mathcal{I})}/\pi_{\mathcal{I}}$. From the CDPH data presented in Table 1, we have the cumulative number of reported COVID-19 infections as of June 12, 2020, which we denote $\hat{N}^{(\mathcal{I})}$. Because $\hat{N}^{(\mathcal{I})}$ measures the cumulative number of COVID-19 infections
120 up to June 12, 2020, and increases daily, $\hat{N}^{(\mathcal{I})}$ is smaller than $N^{(\mathcal{I})}$, perhaps substantially. Furthermore, $\pi_{\mathcal{I}}$ is unknown, and for a given estimate $\hat{\pi}_{\mathcal{I}}$ of $\pi_{\mathcal{I}}$, we define $\hat{N}$ to be $\hat{N} = \hat{N}^{(\mathcal{I})}/\hat{\pi}_{\mathcal{I}}$. Therefore, even for accurate estimates of $\pi_{\mathcal{I}}$, $\hat{N}$ will be smaller, perhaps substantially, than the total number of individuals in the population. However, we assume here that the relative size of $\hat{N}$ to N
125 is approximately equal to the relative size of $\hat{N}^{(\mathcal{I})}$ to $N^{(\mathcal{I})}$. Hence, $\hat{N}$ may be interpreted as an appropriately scaled version of N with respect to $\hat{N}^{(\mathcal{I})}$ and $\hat{\pi}_{\mathcal{I}}$ as of June 12, 2020. Likewise, we define $\hat{N}_{i1} = \hat{N} \times \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \hat{\mathbb{P}}_{\mathbf{X}}(\tilde{\mathbf{x}})$, with $\hat{N}_{i1}$ having the same interpretation as $\hat{N}$ but for the subset of the population with $X_i = 1$. We denote $\hat{N}_{i1}^{(\mathcal{I})}$ to be the cumulative number of infected individuals

130 with $X_i = 1$ as of June 12, 2020, and present $\{\hat{N}_{i1}^{(T)} : i = 1, \dots, p^*\}$ in Table 1.

Substituting $\hat{\mathbb{P}}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1)$ for $\mathbb{P}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1)$, $\{\hat{N}_{i1}^{(T)} : i = 1, \dots, p^*\}$ for $\{N_{i1}^{(T)} : i = 1, \dots, p^*\}$, and $\{\hat{N}_{i1} : i = 1, \dots, p^*\}$ for $\{N_{i1} : i = 1, \dots, p^*\}$ in the pseudo-log-likelihood (5), we obtain an approximate pseudo-likelihood

$$\tilde{\mathcal{L}}(\gamma_0, \gamma) = \sum_{i=1}^{p^*} \tilde{\mathcal{L}}_{X_i}^{(T)}(\gamma_0, \gamma | \hat{N}_{i1}^{(T)}, \hat{N}_{i1}) . \quad (6)$$

We maximize the approximate pseudo-likelihood (6) with respect to (γ_0, γ) to obtain our estimates

$$(\hat{\gamma}_0, \hat{\gamma}) = \arg \max_{\gamma_0, \gamma} \{\tilde{\mathcal{L}}(\gamma_0, \gamma)\} . \quad (7)$$

Lastly, by plugging $(\hat{\gamma}_0, \hat{\gamma})$ into Equation 2, we obtain the predicted test-based infection probabilities for individuals with gender-age-race covariate setting $\mathbf{z}$

$$\hat{\mathbb{P}}(\mathcal{I} = 1 | \mathbf{z}) = \frac{\exp(\hat{\gamma}_0 + \mathbf{z}^\top \hat{\gamma})}{1 + \exp(\hat{\gamma}_0 + \mathbf{z}^\top \hat{\gamma})} . \quad (8)$$

3.2. Case Fatality Rate Estimation Procedure

Similar to the T-IR estimation method, we model the T-CFRs given gender, age, and race using a multivariable logistic regression model. The gender-age-race covariate we use for CFR estimation (see Table 3) is the same as the covariate we use for IR estimation, except that we combined the 0–17 and 18–34 age groups due to low numbers of fatalities among the 0–17 age group. With a slight abuse of notation, we denote $\mathbf{z} \in \{0, 1\}^q \in \mathbb{R}^q$ to be the covariate setting of the vector of non-reference group covariates $\mathbf{Z}$, where $q = 14$. The corresponding random variable $\mathbf{X}$ and its covariate setting $\mathbf{x}$ are as defined in the preceding subsection and have dimension q^* , where $q^* = 17$. We give the T-CFR model as

$$\log \left[\frac{\mathbb{P}(\mathcal{M} = 1 | \mathbf{z}, \mathcal{I} = 1)}{1 - \mathbb{P}(\mathcal{M} = 1 | \mathbf{z}, \mathcal{I} = 1)} \right] = \delta_0 + \mathbf{z}^\top \boldsymbol{\delta}, \quad (9)$$

where $\mathcal{M} \in \{0, 1\}$ represents mortality status, δ_0 is the log odds of mortality for the LatinX female age 0–34 group, and $\boldsymbol{\delta} \in \mathbb{R}^q$ are the log odds ratios of mortality for other covariate settings.

Variable	Definition
Mortality Status	Dichotomous outcome indicating fatality status (0 = No, 1 = Yes)
Gender	Dichotomous covariate indicating gender (0 = Female, 1 = Male)
Age	Categorical covariate with the following age groups: 0–34 (reference level), 35–49, 50–59, 60–64, 65–69, 70–74, 75–79, 80+
Race/Ethnicity	Categorical covariate with the following race categories: LatinX (reference level), White, Asian, African American/Black, Multi-Race, American Indian or Alaska Native, Other

Table 3: Variables used in the case fatality rate estimation procedure

We again employ a pseudo-likelihood approach to estimate $(\delta_0, \boldsymbol{\delta})$ that maximizes a likelihood function constructed from univariate logistic regression models. Following similar steps as shown in the preceding subsection, we have

$$\begin{aligned} & \mathbb{P}(\mathcal{M} = 1 \mid X_i = 1, \mathcal{I} = 1) \\ &= \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \frac{\exp(\delta_0 + \boldsymbol{\delta}^\top \mathbf{z}(\tilde{\mathbf{x}}))}{1 + \exp(\delta_0 + \boldsymbol{\delta}^\top \mathbf{z}(\tilde{\mathbf{x}}))} \mathbb{P}_{\mathbf{X}_{-i} | X_i}(\tilde{\mathbf{x}}_{-i} | x_i = 1, \mathcal{I} = 1). \end{aligned} \quad (10)$$

We use the CHIS data and the IR model (1) with coefficient estimates (7) to estimate $\mathbb{P}_{\mathbf{X}_{-i} | X_i}(\tilde{\mathbf{x}}_{-i} | x_i = 1, \mathcal{I} = 1)$. First, we estimate $\mathbb{P}(\mathbf{z} | \mathcal{I} = 1)$ using Bayes' Rule

$$\hat{\mathbb{P}}(\mathbf{z} | \mathcal{I} = 1) = \frac{\hat{\mathbb{P}}(\mathcal{I} = 1 | \mathbf{z}) \hat{\mathbb{P}}_{\mathbf{X}}(\mathbf{x}(\mathbf{z}))}{\hat{\pi}_{\mathcal{I}}}, \quad (11)$$

where $\hat{\mathbb{P}}(\mathcal{I} = 1 | \mathbf{z})$ comes from Equation (8), and $\hat{\mathbb{P}}_{\mathbf{X}}(\mathbf{x}(\mathbf{z}))$ is obtained from the CHIS dataset. Then, by the definition of conditional probability, we estimate $\mathbb{P}_{\mathbf{X}_{-i} | X_i}(\tilde{\mathbf{x}}_{-i} | x_i = 1, \mathcal{I} = 1)$ by

$$\hat{\mathbb{P}}_{\mathbf{X}_{-i} | X_i}(\tilde{\mathbf{x}}_{-i} | x_i = 1, \mathcal{I} = 1) = \frac{\hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}}) | \mathcal{I} = 1)}{\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}}) | \mathcal{I} = 1)}, \quad (12)$$

135 where $\hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}})|\mathcal{I} = 1)$ comes from Equation (11).

Analogous to the IR model, we denote $N_{i1}^{(\mathcal{M})}$ to be the number of individuals with $X_i = 1$ who have died or will die from COVID-19. Therefore, $\{N_{i1}^{(\mathcal{M})} : i = 1, \dots, q^*\}$ each follows a binomial distribution

$$N_{i1}^{(\mathcal{M})} \sim \text{Binomial}(N_{i1}^{(\mathcal{I})}, \sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(X_i=1)}} \frac{\exp(\delta_0 + \boldsymbol{\delta}^\top \mathbf{z}(\tilde{\mathbf{x}}))}{1 + \exp(\delta_0 + \boldsymbol{\delta}^\top \mathbf{z}(\tilde{\mathbf{x}}))} \mathbb{P}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1, \mathcal{I} = 1)). \quad (13)$$

We then construct the full pseudo-log-likelihood of $(\delta_0, \boldsymbol{\delta})$ as the sum of the individual pseudo-log-likelihoods of $(\delta_0, \boldsymbol{\delta})$ for X_i corresponding to binomial distribution (13)

$$\mathcal{L}(\delta_0, \boldsymbol{\delta}) = \sum_{i=1}^{q^*} \mathcal{L}_{X_i}^{(\mathcal{M})}(\delta_0, \boldsymbol{\delta} | N_i^{(\mathcal{M})}, N_i^{(\mathcal{I})}). \quad (14)$$

From the CDPH data presented in Table 1, we have the cumulative number of COVID-19 deaths by gender, age, and race. We denote $\hat{N}_{i1}^{(\mathcal{M})}$ to be the cumulative number of reported deaths of infected individuals with $X_i = 1$ as of June 12, 2020. Analogous to the infection risk model, we assume that the relative size of $\hat{N}_i^{(\mathcal{I})}$ to $N_i^{(\mathcal{I})}$ is approximately equal to the relative size of $\hat{N}_i^{(\mathcal{M})}$ to $N_i^{(\mathcal{M})}$. Substituting $\hat{N}_i^{(\mathcal{I})}$ for $N_i^{(\mathcal{I})}$, $\hat{N}_i^{(\mathcal{M})}$ for $N_i^{(\mathcal{M})}$, and $\hat{\mathbb{P}}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1)$ for $\mathbb{P}_{\mathbf{X}_{-i}|X_i}(\tilde{\mathbf{x}}_{-i}|x_i = 1)$, $i = 1, \dots, q^*$, in the pseudo-likelihood (14), we obtain an approximate pseudo-likelihood

$$\tilde{\mathcal{L}}(\delta_0, \boldsymbol{\delta}) = \sum_{i=1}^{q^*} \tilde{\mathcal{L}}_{X_i}^{(\mathcal{M})}(\delta_0, \boldsymbol{\delta} | \hat{N}_i^{(\mathcal{I})}, \hat{N}_i^{(\mathcal{M})}) \quad (15)$$

and maximize it with respect to $(\delta_0, \boldsymbol{\delta})$ to obtain our estimates

$$(\hat{\delta}_0, \hat{\boldsymbol{\delta}}) = \arg \max_{\delta_0, \boldsymbol{\delta}} \{\tilde{\mathcal{L}}(\delta_0, \boldsymbol{\delta})\}. \quad (16)$$

Lastly, from $(\hat{\delta}_0, \hat{\boldsymbol{\delta}})$, we can obtain the predicted COVID-19 test-based case fatality rates for individuals with gender-age-race covariate setting $\mathbf{z}$,

$$\hat{\mathbb{P}}(\mathcal{M} = 1 | \mathbf{z}, \mathcal{I} = 1) = \frac{\exp(\hat{\delta}_0 + \mathbf{z}^\top \hat{\boldsymbol{\delta}})}{1 + \exp(\hat{\delta}_0 + \mathbf{z}^\top \hat{\boldsymbol{\delta}})}. \quad (17)$$

3.3. Monte Carlo Simulation Procedure

To quantify the uncertainty of the T-IR and T-CFR estimates in (8) and (17), respectively, we carry out a Monte Carlo procedure that repeatedly performs the T-IR and T-CFR estimation procedures described in Sections 3.1 and 3.2 sequentially, introducing sampling variation in the data in three stages. The first stage bootstraps the CHIS data with selection probabilities proportional to the sampling weights. The second stage introduces variation in $\{\hat{N}_{i1}^{(T)}\}$ immediately prior to maximizing the approximate pseudo-log-likelihood (9), by simulating values of $\hat{N}_{i1}^{(T)}$ for each i independently from a binomial distribution with success probability equal to $\hat{N}_{i1}^{(T)}/\hat{N}_{i1}$, i.e.,

$$\tilde{N}_{i1}^{(T)} \stackrel{\text{ind}}{\sim} \text{Binomial} \left(\hat{N}_{i1}, \frac{\hat{N}_{i1}^{(T)}}{\hat{N}_{i1}} \right). \quad (18)$$

Similarly, the third stage introduces variation in the $\{\hat{N}_{i1}^{(\mathcal{M})}\}$ prior to maximizing the approximate pseudo-log-likelihood (15) by simulating values of $\hat{N}_{i1}^{(\mathcal{M})}$ for each i independently from a binomial distribution with success probability equal to $\hat{N}_{i1}^{(\mathcal{M})}/\hat{N}_{i1}^{(T)}$,

$$\tilde{N}_{i1}^{(\mathcal{M})} \stackrel{\text{ind}}{\sim} \text{Binomial} \left(\hat{N}_{i1}^{(T)}, \frac{\hat{N}_{i1}^{(\mathcal{M})}}{\hat{N}_{i1}^{(T)}} \right). \quad (19)$$

The entire Monte Carlo simulation procedure can be summarized in 5 steps:

Monte Carlo Simulation Procedure

140

Step 1: Bootstrap the CHIS dataset with selection probabilities proportional to the sampling weights.

145

Step 2: Perform the T-IR estimation procedure, simulating a value of $\hat{N}_{i1}^{(T)}$ from (18) for each X_i , subsequently obtaining estimates $(\hat{\gamma}_0, \hat{\gamma})$.

Step 3: Using $(\hat{\gamma}_0, \hat{\gamma})$ obtained in Step 2, perform the T-CFR estimation procedure, simulating a value of $\hat{N}_{i1}^{(\mathcal{M})}$ from (19) for all X_i , subsequently obtaining estimates $(\hat{\delta}_0, \hat{\delta})$.

Step 4: Repeat Steps 1–3 to obtain a total of $\mathcal{B}$ estimates of (γ_0, γ) and (δ_0, δ) , which we denote $\{(\hat{\gamma}_0^{(b)}, \hat{\gamma}^{(b)}) : b = 1, \dots, \mathcal{B}\}$ and $\{(\hat{\delta}_0^{(b)}, \hat{\delta}^{(b)}) : b = 1, \dots, \mathcal{B}\}$, respectively.

150

Step 5: For each set of bootstrap coefficient estimates $(\gamma_0^{(b)}, \gamma^{(b)}, \delta_0^{(b)}, \delta^{(b)})$, $b = 1, \dots, \mathcal{B}$, estimate the T-IR and T-CFR for covariate setting $\mathbf{z}$ using Equations (8) and (17), respectively.

In Figure 1, We illustrate the Monte Carlo simulation procedure in a flow chart.

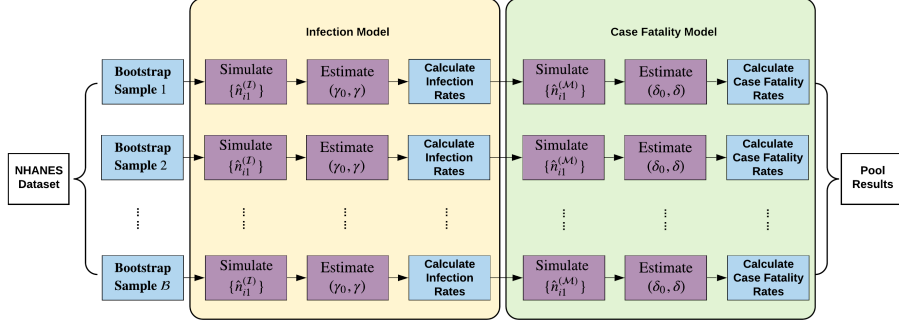


Figure 1: Flow chart depicting the Monte Carlo simulation procedure

155

3.4. Summary Statistics for Infection and Case Fatality Rate Estimates

In addition to estimating the T-IR and T-CFR for specific covariate settings $\mathbf{z}$ through Equations (8) and (17), respectively, we can provide collapsed estimates of T-IRs and T-CFRs for specific values of any subset $\{X_{j_1}, \dots, X_{j_r}\}$ of $\mathbf{X}$. Let $J_r = \{j_1, \dots, j_r\}$, where $J_r \subset \{1, \dots, p^*\}$; $\mathbf{c}_r = (c_{j_1}, \dots, c_{j_r})$, where $c_{j_1}, \dots, c_{j_r} \in \{0, 1\}$; and $\mathcal{X}_{(J_r)}^{(\mathbf{c}_r)}$ denotes the subset of $\mathcal{X}$ with the constraint that $X_{j_1} = c_{j_1}, \dots, X_{j_r} = c_{j_r}$. Estimates of collapsed T-IRs given $X_{j_1} = c_{j_1}, \dots, X_{j_r} = c_{j_r}$ can be obtained using the marginalization formula

$$\hat{\mathbb{P}}(\mathcal{I} = 1 | X_{j_1} = c_{j_1}, \dots, X_{j_r} = c_{j_r}) = \frac{\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(J_r)}^{(\mathbf{c}_r)}} \hat{\mathbb{P}}(\mathcal{I} = 1 | \mathbf{z}(\tilde{\mathbf{x}})) \hat{\mathbb{P}}_{\mathbf{X}}(\mathbf{z}(\tilde{\mathbf{x}}))}{\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(J_r)}^{(\mathbf{c}_r)}} \hat{\mathbb{P}}_{\mathbf{X}}(\mathbf{z}(\tilde{\mathbf{x}}))}. \quad (20)$$

Likewise, collapsed estimates of T-CFRs given $X_{j_1} = c_{j_1}, \dots, X_{j_r} = c_{j_r}$ can be obtained using the marginalization formula

$$\begin{aligned} & \hat{\mathbb{P}}(\mathcal{M} = 1 | X_{j_1} = c_{j_1}, \dots, X_{j_r} = c_{j_r}, \mathcal{I} = 1) \\ &= \frac{\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(J_r)}^{(\mathbf{c}_r)}} \hat{\mathbb{P}}(\mathcal{M} = 1 | \mathbf{z}(\tilde{\mathbf{x}}), \mathcal{I} = 1) \hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}}) | \mathcal{I} = 1)}{\sum_{\tilde{\mathbf{x}} \in \mathcal{X}_{(J_r)}^{(\mathbf{c}_r)}} \hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}}) | \mathcal{I} = 1)}, \end{aligned} \quad (21)$$

where $\hat{\mathbb{P}}(\mathbf{z}(\tilde{\mathbf{x}})|\mathcal{I} = 1)$ comes from Equation (11).

4. Results

We present select estimates of T-IRs and T-CFRs obtained from our IR (1) and CFR (8) models fit to the California data described in Section 2 and summarized according to Section 3.4. All standard errors were computed using a bootstrap size of $\mathcal{B} = 100$. As a baseline estimate, we assume an overall California COVID-19 infection rate $\pi_{\mathcal{I}}$ of 2%, a rough estimate of the infection rate of the 1918 Influenza Pandemic [19] in the U.S., that is estimated to have had a basic reproductive number ($\mathcal{R}_0$) comparable to that of COVID-19 [20, 21]. However, there is still substantial uncertainty surrounding the true COVID-19 infection rate primarily due to the lack of testing and the large prevalence of asymptomatic cases. Recent studies suggest that the true overall infection rate in the U.S. is much higher than what was initially hypothesized [22, 23].

Figure 2 depicts T-IR estimates and error bars indicating two bootstrap standard errors (SEs) for different combinations of gender and age group under the assumption of an overall California infection rate of 2%. The T-IR estimates range from 0.1% to 8.3% for female and 0.1 % to 8.9% for male. The figures present 6 different race/ethnicity groups: LatinX/ Hispanic (LatinX), White/ Caucasian (White), Asian, African American / Black (AA), Multi-Race, and American Indian or Alaska Native (AIAN). LatinX has the highest T-IRs, followed by African American. Both females and males age 80 and older have extremely high T-IRs compared with other age groups across race/ethnicity groups. T-IRs were non-monotonic at younger ages, with age groups 60–64 and 70–74 having slightly lower T-IRs than the preceding age groups, 50–59 and 65–69 respectively. Males had higher T-IRs than females across all age groups with the gender gap slightly increasing with age.

We also considered alternate values for the overall California IR. Table 4 presents the point estimates and associated two SE intervals of the marginal T-IRs for gender and age group obtained from marginalization formula (20)

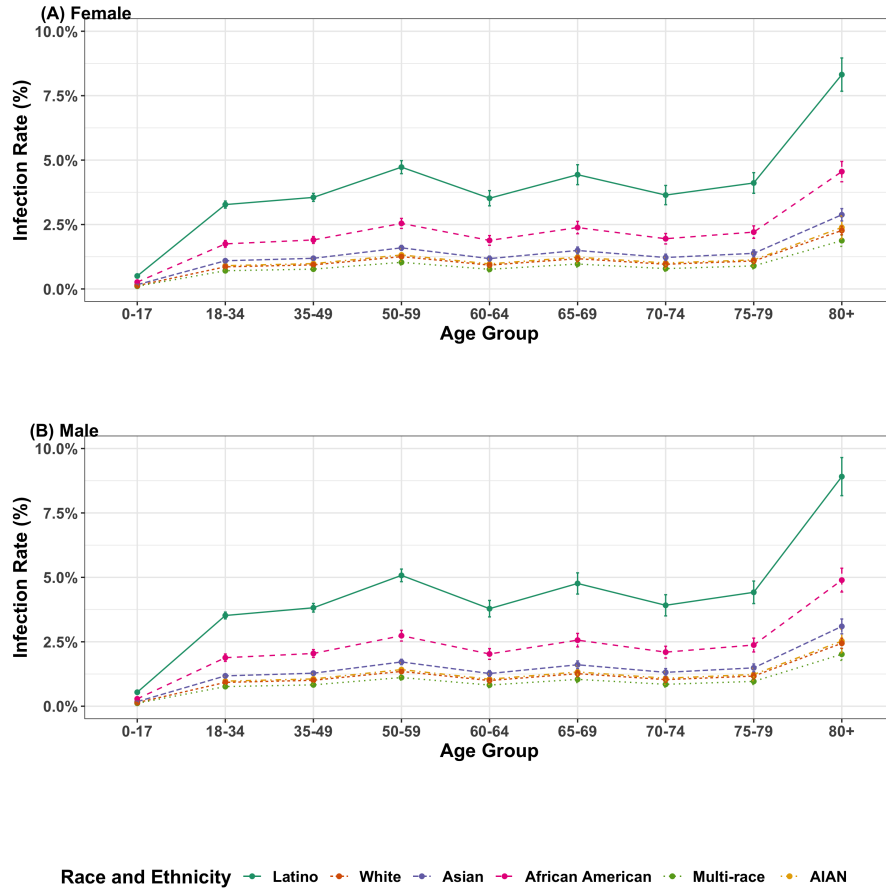


Figure 2: Estimated test-based infection rates given age and race/ethnicity, stratified by gender. (A) and (B) present the bootstrapped mean infection rates for female and male respectively. Only 6 racial and ethnicity groups are considered in the figures, including LatinX/Hispanic (LatinX), White/ Caucasian (White), Asian, African American / Black (AA), Multi-Race, and American Indian or Alaska Native (AIAN). The overall infection rate was assumed to be 2%, and the error bars denote two bootstrap standard errors.

assuming overall IRs of 1%, 2%, and 5%. The estimated marginal T-IRs for gender, age groups and race and ethnicity groups are consistent with the results presented in Figure 2, including males and older individuals having higher estimated T-IRs.

190 Table 5 presents the point estimates and associated two SE intervals of

		Default Infection Rate (Mean $\pm$ 2 SE)		
Variable	Category	1%	2%	5%
	Overall	1.00% (0.00%)	2.00% (0.01%)	5.00% (0.02%)
Gender	Male	1.03% (0.01%)	2.05% (0.02%)	5.13% (0.06%)
	Female	0.97% (0.01%)	1.95% (0.02%)	4.87% (0.05%)
Age Group	0–17	0.29% (0.01%)	0.57% (0.01%)	1.43% (0.04%)
	18–34	1.21% (0.02%)	2.43% (0.04%)	6.07% (0.10%)
	35–49	1.29% (0.03%)	2.58% (0.06%)	6.46% (0.13%)
	50–59	1.34% (0.04%)	2.69% (0.07%)	6.73% (0.19%)
	60–64	0.98% (0.04%)	1.96% (0.08%)	4.90% (0.20%)
	65–69	1.06% (0.05%)	2.13% (0.11%)	5.31% (0.27%)
	70–74	0.89% (0.05%)	1.78% (0.10%)	4.46% (0.25%)
	75–79	0.90% (0.06%)	1.79% (0.10%)	4.47% (0.29%)
	80+	1.73% (0.10%)	3.46% (0.20%)	8.66% (0.52%)
Race & Ethnicity	LatinX	1.41% (0.02%)	2.82% (0.03%)	7.06% (0.08%)
	White	0.48% (0.01%)	0.97% (0.02%)	2.42% (0.04%)
	Asian	0.59% (0.02%)	1.17% (0.04%)	2.92% (0.10%)
	AA	0.92% (0.05%)	1.84% (0.10%)	4.62% (0.21%)
	Multi-race	0.27% (0.02%)	0.55% (0.04%)	1.36% (0.11%)
	AIAN	0.57% (0.11%)	1.14% (0.20%)	2.87% (0.57%)
	Other	13.92% (1.61%)	27.85% (3.20%)	69.68% (8.01%)

Table 4: Estimated marginal test-based infection rates for different overall infection rates

the bootstrap estimated marginal T-CFRs obtained from marginalization formula (21), assuming an overall infection rate of $\pi_{\mathcal{I}} = 2\%$; T-CFR estimates do not vary in expectation for different values of $\pi_{\mathcal{I}}$. Males have a mean T-CFR 0.78% higher than females, and T-CFRs increase with age, ranging from less than 0.11% for the 0-34 age group to over 27.45% for the 80+ age group. Among 6 race and ethnicity groups, African American, Asian, and White are high-risk groups with mean T-CFRs as 7.42%, 6.80%, and 6.78% respectively.

Other, LatinX, and Multi-race subgroups have T-CFRs below the overall 3.68% T-CFR for California.

Variable	Category	Case Fatality Rate (MEAN $\pm$ 2 SE)
Overall	Overall	3.68% (0.06%)
Gender	Male	4.06% (0.13%)
	Female	3.28% (0.12%)
Age Group	0–34	0.11% (0.03%)
	35–49	0.74% (0.10%)
	50–59	2.07% (0.19%)
	60–64	4.08% (0.46%)
	65–69	7.41% (0.70%)
	70–74	11.25% (0.91%)
	75–79	17.19% (1.21%)
	80+	27.45% (0.87%)
Race & Ethnicity	LatinX	2.62% (0.10%)
	White	6.78% (0.28%)
	Asian	6.80% (0.50%)
	AA	7.42% (0.71%)
	Multi-race	3.31% (1.05%)
	AIAN	5.19% (2.54%)
	Other	0.37% (0.10%)

Table 5: Estimated marginal test-based case fatality rates

Figure 3 presents the estimated T-CFRs, obtained from formula (21), with error bars displaying two SEs of uncertainty for different combinations of gender and age groups, stratified by 6 race and ethnicity groups as shown in Figure 2. Males have higher estimated T-CFRs than females across all age-race levels, and the gender gap increases with age. African American and then Multi-race have higher estimated T-CFRs than other race groups in general across different

age groups based on the stratified results. African American female even has a higher of T-CFRs than AIAN, LatinX and White male for each age group correspondingly.

Although Multi-race has higher T-CFRs at each age groups than White, the overall marginal T-CFRs for Multi-race is lower than White as shown in Table 5. The reversal of the inequality between the size ratios is an example of Simpson's paradox. The Multi-race population is younger than White population in California. For example, we can compare the number of adults (18 and above) and the number of children and adolescents (0 to 17 years old) within each race and ethnicity group. An overall adult-child ratio is defined as the total number of adults over that of children and adolescents ignoring race and ethnicity groups. Multi-race population in California has 0.43 times overall adult-child ratio (1.7% of adults and 4% of children and adolescents), while it is 1.3 times overall adult-child ratio for White population (38.8% of adults and 29.2% of children and adolescents)[6]. Since the T-CFRs increase with the age, a higher adult-child ratio in the age structure for Multi-race leads to a higher marginal T-CFR. The case still holds when we have multiple age groups. A similar paradox happens for LatinX and Asian. Meanwhile, the small proportions of Multi-race and AIAN in the general population also result in large error bars.

5. Discussion

In this paper, we combined aggregate COVID-19 case and fatality data with population demographic data in a pseudo-likelihood based multivariable logistic regression approach for obtaining early estimates of COVID-19 T-IRs and T-CFRs for subgroups of the California population. Overall, our results revealed that males, the elderly, and LatinX are marginally at relatively higher risk of COVID-19 infection, and that males, the elderly, Africa Americans, Asians, and Whites are marginally at elevated risk of mortality after COVID-19 infection. However, due to the imbalance in the age distribution of different races in California, the subgroups with the top 5 T-CFRs are Africa American male,

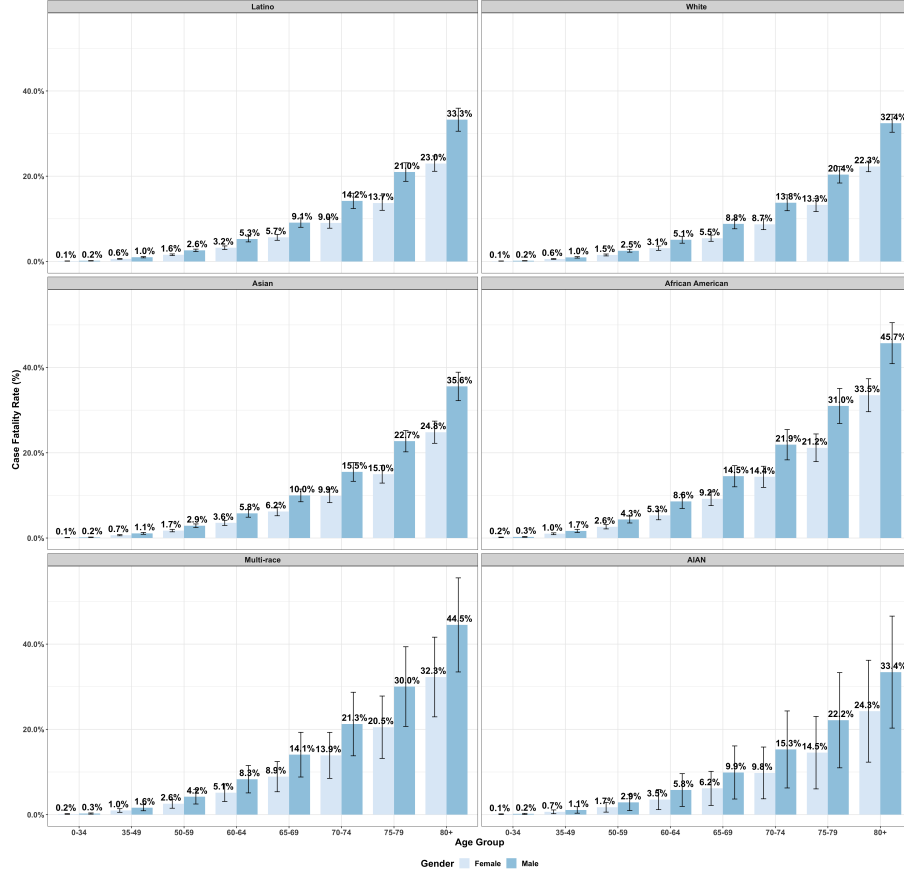


Figure 3: Estimated test-based case fatality rates by age and gender, stratified by race/ethnicity. The bootstrap mean case fatality rates are presented separately for LatinX, White, Asian, African American, Multi-Race, and AIAN groups. The overall infection rate was assumed to be 2%, and the error bars denote two bootstrap standard errors.

235 Multi-race male, Asian male, African American female, and AIAN male for each age group. Overall, therefore, African Americans are the race/ethnicity group most vulnerable to COVID-19 in California. We also found that the elevated infection and mortality risk for males and the greater mortality risk for all races increase with age.

240 The proposed methods are subject to three general limitations. First, the analysis is based on publicly available test-based infection rates and case fatality

rates. It has been well documented that the lack of testing for COVID-19 in the U.S. has hindered efforts to estimate the true COVID-19 infection rate. Further compounding this issue is the high prevalence of asymptomatic COVID-19 cases. These two issues may lead to substantial underestimates of the infection rates and/or substantial overestimates of the case fatality rates from our analyses. Second, race/ethnicity is missing in 29% of the reported cases from CDPH, which may bias our estimates. Even though CDPH releases summary statistics for age-race covariates and our model can fit the finer data, we fit the marginal statistics for each risk factor in the analysis to minimize the impact of the potential sampling bias. Moreover, the case and fatality data released by CDPH provide marginal summary statistics for a subset of risk factors, and we do not have direct information on the joint distribution of all risk factors. Although the central goal of our proposed methods is to circumvent this limitation, the absence of direct multivariate information on the risk factors of COVID-19 infection and mortality as well as the sampling bias should be taken into account when interpreting the results of our models. Third, in this paper, we do not consider regularity conditions ensuring concavity associated with the pseudo-log-likelihood functions constructed in Equations (6) and (15), nor do we examine the asymptotic properties of the parameter estimates in Equations (7) and (16). Future research investigating the mathematical theory of the proposed methods is warranted.

Another promising avenue for future work is combining this method with a COVID-19 prediction model [24] to provide detailed demographic projections of COVID-19 cases and mortalities. This would be a substantial improvement over most COVID-19 prediction models, as they tend to be quite limited in their ability to forecast the demographic characteristics of the infected.

In summary, this paper provides a pragmatic tool for producing early estimates of COVID-19 T-IRs and T-CFRs for the California population, which offer valuable information to guide health policies concerning the control and prevention of COVID-19. In addition, our methods can be generalized into a general framework for early estimation of subpopulation IRs and CFRs from

aggregate case and fatality data in other locations and for future epidemics.

Acknowledgements

275 We thank Dr. Sudipto Banerjee and Jay J. Xu (University of California, Los Angeles) for their many helpful comments and assistance.

Author Contributions

DX, MAS, and CMR conceptualized the study design. DX, LZ, GLW, and CMR drafted the original manuscript. DX, PS, TB, and CMR performed the data collection and literature search. DX and LZ developed the methodology, 280 conducted the statistical analysis, and interpreted the results. DX, LZ, GLW, PS, TB, JZ, JS, and MAS performed the computer programming. All authors revised the manuscript and approved the final version.

Funding

285 The authors received no specific funding for this work.

Declaration of Competing Interest

The authors have declared that no competing interests exist.

References

- 290 [1] A. L. Phelan, R. Katz, L. O. Gostin, The novel coronavirus originating in wuhan, china: challenges for global health governance, *Jama* 323 (8) (2020) 709–710.
- [2] D. Cucinotta, M. Vanelli, Who declares covid-19 a pandemic., *Acta bio-medica: Atenei Parmensis* 91 (1) (2020) 157–160.

- 295 [3] C. COVID, global cases by the center for systems science and engineering
(csse) at johns hopkins university (jhu), ArcGIS. Johns Hopkins CSSE.
Retrieved April 8 (19) 2020.
- [4] S. Garg, Hospitalization Rates and Characteristics of Patients Hospitalized
with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14
States, March 1–30, 2020, MMWR. Morbidity and Mortality Weekly Re-
300 port 69.
- [5] [dataset]New Jersey Department of Health, COVID-19 Information Hub,
<https://covid19.nj.gov/> (2020 (accessed June 13, 2020)).
- [6] [dataset]California Department of Public Health, COVID-19 Race and
Ethnicity Data, [https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/](https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Race-Ethnicity.aspx)
305 [COVID-19/Race-Ethnicity.aspx](https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Race-Ethnicity.aspx) (2020 (accessed June 13, 2020)).
- [7] [dataset]Illinois Department of Public Health, Coronavirus Dis-
ease 2019 (COVID-19), [https://www.dph.illinois.gov/covid19/](https://www.dph.illinois.gov/covid19/covid19-statistics)
[covid19-statistics](https://www.dph.illinois.gov/covid19/covid19-statistics) (2020 (accessed June 13, 2020)).
- [8] Z. Zheng, F. Peng, B. Xu, J. Zhao, H. Liu, J. Peng, Q. Li, C. Jiang,
310 Y. Zhou, S. Liu, et al., Risk factors of critical & mortal COVID-19 cases:
A systematic literature review and meta-analysis, *Journal of Infection*.
- [9] J.-M. Jin, P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, S. Liu, J.-K. Yang,
Gender Differences in Patients With COVID-19: Focus on Severity and
Mortality, *Frontiers in Public Health* 8 (2020) 152.
- 315 [10] A. B. Docherty, E. M. Harrison, C. A. Green, H. E. Hardwick, R. Pius,
L. Norman, K. A. Holden, J. M. Read, F. Dondelinger, G. Carson, L. Mer-
son, J. Lee, D. Plotkin, L. Sigfrid, S. Halpin, C. Jackson, C. Gamble, P. W.
Horby, J. S. Nguyen-Van-Tam, A. Ho, C. D. Russell, J. Dunning, P. J.
Openshaw, J. K. Baillie, M. G. Semple, Features of 20 133 uk patients in
320 hospital with covid-19 using the isaric who clinical characterisation proto-
col: prospective observational cohort study, *BMJ* 369 (m1985).

- [11] R.-H. Du, L.-R. Liang, C.-Q. Yang, W. Wang, T.-Z. Cao, M. Li, G.-Y. Guo, J. Du, C.-L. Zheng, Q. Zhu, et al., Predictors of mortality for patients with covid-19 pneumonia caused by sars-cov-2: a prospective cohort study, *European Respiratory Journal* 55 (5).
325
- [12] D. Mavridis, G. Salanti, A practical introduction to multivariate meta-analysis, *Statistical methods in medical research* 22 (2) (2013) 133–158.
- [13] M. C. Simmonds, J. P. Higgins, A general framework for the use of logistic regression models in meta-analysis, *Statistical methods in medical research* 25 (6) (2016) 2858–2877.
330
- [14] B.-H. Chang, S. Lipsitz, C. Waternaux, Logistic regression in meta-analysis using aggregate data, *Journal of Applied Statistics* 27 (4) (2000) 411–424.
- [15] F. Caramelo, N. Ferreira, B. Oliveiros, Estimation of risk factors for covid-19 mortality-preliminary results, medRxiv.
- [16] [dataset]California Department of Public Health, California Coronavirus (COVID-19) Response, <https://update.covid19.ca.gov/#top> (2020 (accessed June 13, 2020)).
335
- [17] [dataset]California Department of Public Health, Cases and Deaths Associated with COVID-19 by Age Group in California, <https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID-19-Cases-by-Age-Group.aspx> (2020 (accessed June 13, 2020)).
340
- [18] [dataset]California Health Interview Survey, CHIS 2017-2018 Public Use File. Los Angeles, CA: UCLA Center for Health Policy Research, <https://healthpolicy.ucla.edu/chis/data/Pages/GetCHISData.aspx> (2020 (accessed June 13, 2020)).
345
- [19] E. Vynnycky, A. Trindall, P. Mangtani, Estimates of the reproduction numbers of spanish influenza using morbidity data, *International Journal of Epidemiology* 36 (4) (2007) 881–889.

- [20] Q. Liu, Z. Liu, J. Zhu, Y. Zhu, D. Li, Z. Gao, L. Zhou, J. Yang, Q. Wang,
350 Assessing the global tendency of covid-19 outbreak, medRxiv.
- [21] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao,
L. Yang, D. He, et al., Preliminary estimation of the basic reproduction
number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A
data-driven analysis in the early phase of the outbreak, International jour-
355 nal of infectious diseases 92 (2020) 214–217.
- [22] E. Bendavid, B. Mulaney, N. Sood, S. Shah, E. Ling, R. Bromley-Dulfano,
C. Lai, Z. Weissberg, R. Saavedra, J. Tedrow, et al., Covid-19 antibody
seroprevalence in santa clara county, california, medRxiv.
- [23] D. Sutton, K. Fuchs, M. D’alton, D. Goffman, Universal screening for sars-
360 cov-2 in women admitted for delivery, New England Journal of Medicine.
- [24] G. L. Watson, D. Xiong, L. Zhang, J. A. Zoller, J. Shamsioian, P. Sundin,
T. Bufford, A. W. Rimoin, M. A. Suchard, C. M. Ramirez, Fusing a
bayesian case velocity model with random forest for predicting covid-19
in the us, Available at SSRN 3594606.