Lessons from COVID-19

Non-random Missing Data and Its Consequences

Mitzi Morris

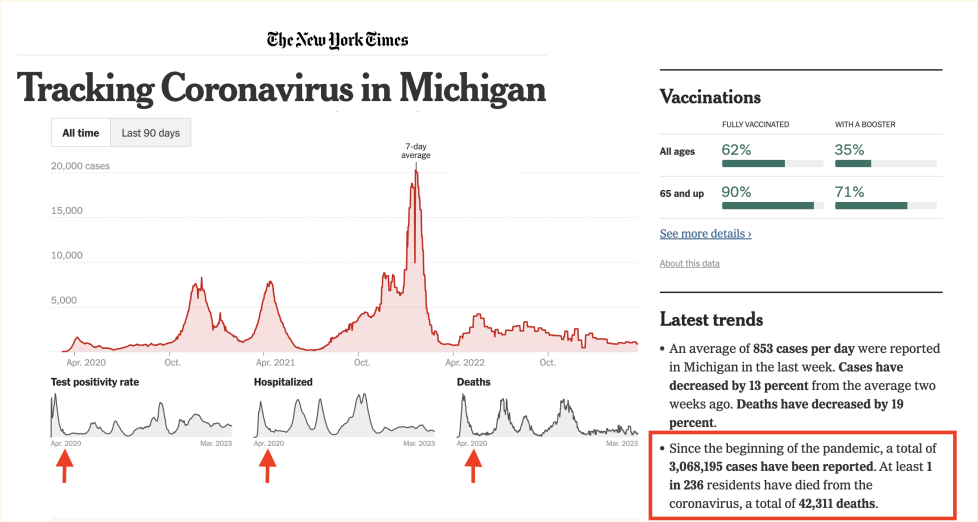
2023-08-08

Stan Development Team

Talk Outline

- Covid-19 in Michigan
- Missing data and ways to handle it
- Study: simulate data, compare model inferences
- Joint model of Covid-19 incidence and non-random missingness

Covid-19 pandemic in Michigan



- Trangucci et al data: Covid-19 positive PCR tests March 1 - June 31 2020

Michigan mortality rates March-October 2020

Research Paper

Racial disparities in COVID-19 mortality across Michigan, United States

Alyssa S. Parpia[a], Isabel Martinez[a], Abdulrahman M. El-Sayed[b,c], Chad R. Wells[a], Lindsey Myers[d], Jeffrey Duncan[d], Jim Collins[e], Meagan C. Fitzpatrick[a,f], Alison P. Galvani[a,*], Abhishek Pandey[a]

ARTICLE INFO

Article History:
Received 2 December 2020
Revised 3 February 2021
Accepted 3 February 2021
Available online 26 February 2021

Keywords:
Covid-19
Pandemics
Race factors
Racism
Emerging infectious diseases
United States
Michigan

ABSTRACT

Background: Black populations in the United States are being disproportionately affected by the COVID-19 pandemic, but the increased mortality burden after accounting for health and other demographic characteristics is not well understood. We examined characteristics of individuals who died from COVID-19 in Michigan by race stratified by their age, sex and comorbidity prevalence to illustrate and understand this disparity in mortality risk.

Methods: We evaluate COVID-19 mortality in Michigan by demographic and health characteristics, using individual-level linked death certificate and surveillance data collected by the Michigan Department of Health and Human Services from March 16 to October 26, 2020. We identified differences in demographics and comorbidity prevalence across race among individuals who died from COVID-19 and calculated mortality rates by age, sex, race, and number of comorbidities.

Findings: Among the 6,065 COVID-19 related deaths in Michigan, Black individuals are experiencing 3·6 times the mortality rate of White individuals ($p<0.001$), with a mortality rate for Black individuals under 65 years without comorbidities that is 12·6 times that of their White counterparts ($p<0.001$). After accounting for age, race, sex, and number of comorbidities, we find that Black individuals in all strata are at higher risk of COVID-19 mortality than their White counterparts.

EClinicalMedicine

- During the first 6 months of the pandemic
  - Mortality rate for Black individuals was 3.6 times that of White individuals.
  - Rate for Black individuals under 65 without comorbidities was 12.6 times greater.

Covid-19 Imperative: What's the Infection Rate?

Need to estimate trends in the general population and vulnerable subpopulations.

- Available data is PCR test data, may have missing demographics
  - Always present: age, sex, address
  - Sometimes missing: race/ethnicity
- Disadvantaged subpopulations may get fewer tests.
- Economic incentives to avoid testing.
- ***Consequence: unreliable estimates of Covid-19 prevalence***
  - Difficult to make informed policy decisions.
  - Erosion of confidence in public health officials and health care system.

Missing Data

Wikipedia article "Missing Data"

- ***Missing Completely at Random (MCAR)*** - *the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random.*
- ***Missing at random (MAR)*** - *missingness is not random, but may be possible to infer.* ("multiple imputation")
- ***Not Missing at Random (NMAR)*** - "nonignorable nonresponse" - *the value of the variable that's missing is related to the reason it's missing.*

The more data is NMAR, the more biased are the model estimates.

Background - Survey Data

- A *stratum* is any subset of the data based on some categorical feature(s).
- For a diverse population, stratification provides better inferences about subgroups.
- Michigan study data has the following features:
  - age (9 levels), sex assigned at birth (2 levels), geographic location (13 levels), race/ethnicity (4 levels)
- Trangucci et al start from stratification by age x sex: 18 strata
- The most fine-grained level of stratification is the cartesian product (cross-product) of all categorial variables.
  - age x sex x location - 234 strata
  - age x sex x location x race - 936 strata

Trangucci et al - Simulation Study Data

Compare modeling choices for missing data using simulated data.

- Simulated datasets are constructed Not Missing at Random (NMAR)
  - Missing race/ethnicity information varies solely by race/ethnicity, i.e., cannot be imputed from other variables

- Dataset construction
  - Same age, sex, race/ethnicity percentages as Wayne County, MI (Detroit)
  - Cases with race/ethnicity known: 90%, 80%, 60%, 20%
  - Race/ethnicity missingness varies by race, baseline White, ratio known Black/White $\frac{.75}{.9}$, Other/White $\frac{.6}{.9}$, (Latino/White 1 : 1)
  - Rate of Covid-19 varies by age and sex, but doesn't vary by race/ethnicity.

2010 Census Data for Wayne County, MI

TABLE 1

Population summary in Wayne County, Michigan as of the 2010 Decennial Census

Race/Ethnicity	Total Pop.	Mean Age×Sex×Race/Eth. ×PUMA Pop.	Std. dev. PUMA Pop.	100× Ratio to White
Black	732801	3132	3152	81
Hispanic/Latino	95260	407	757	11
Other	90343	386	397	10
White	902180	3855	3225	100

Wayne County includes the city of Detroit and close-in suburbs.

Simulated Datasets

TABLE 2

The table summarises the simulation study by missingness scenario by race/ethnicity. 200 datasets were simulated in each scenario. The column "Mean Obs." gives the average proportion of cases observed with race/ethnicity data across 200 simulated datasets. Similarly, "Mean True Inc." is the mean true incidence by group, and "Mean Obs. Inc." is the mean observed incidence by group.

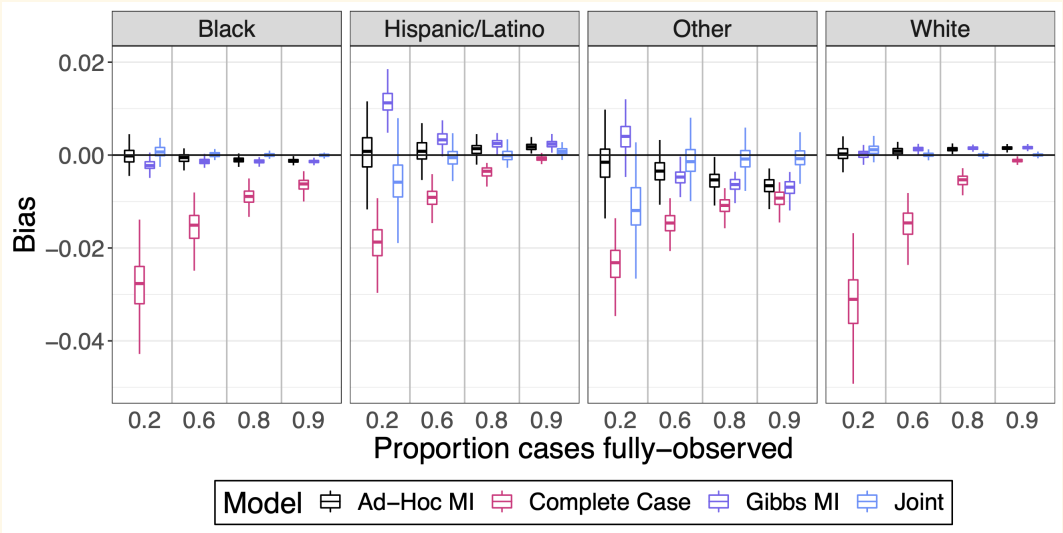
Proportion cases w/ race/ethnicity	Race/Ethnicity	Mean Obs.	Std. dev.	Mean True Inc.	Std. dev.	Mean Obs. Inc.	Std. dev.
90%	Black	80.7%	(2.4%)	3.4%	(0.9%)	2.8%	(0.8%)
	Hispanic/Latino	96.7%	(0.7%)	2.4%	(0.7%)	2.3%	(0.7%)
	Other	63.9%	(3.0%)	2.6%	(0.6%)	1.7%	(0.4%)
	White	97.1%	(0.5%)	4.4%	(1.8%)	4.3%	(1.7%)
80%	Black	72.7%	(3.0%)	3.4%	(1.0%)	2.5%	(0.8%)
	Hispanic/Latino	85.0%	(2.4%)	2.4%	(0.6%)	2.1%	(0.5%)
	Other	57.4%	(3.2%)	2.6%	(0.6%)	1.5%	(0.3%)
	White	86.5%	(2.1%)	4.2%	(1.2%)	3.7%	(1.0%)
60%	Black	53.7%	(4.6%)	3.5%	(1.1%)	1.9%	(0.7%)
	Hispanic/Latino	60.3%	(4.3%)	2.4%	(0.6%)	1.5%	(0.4%)
	Other	42.3%	(3.7%)	2.6%	(0.5%)	1.1%	(0.3%)
	White	64.4%	(4.4%)	4.3%	(1.3%)	2.8%	(1.0%)
20%	Black	17.2%	(3.5%)	3.4%	(0.8%)	0.6%	(0.2%)
	Hispanic/Latino	18.4%	(3.2%)	2.4%	(0.7%)	0.4%	(0.2%)
	Other	12.9%	(2.3%)	2.7%	(0.5%)	0.4%	(0.1%)
	White	21.7%	(4.5%)	4.4%	(1.5%)	1.0%	(0.6%)

Trangucci et al - Simulation Study Models

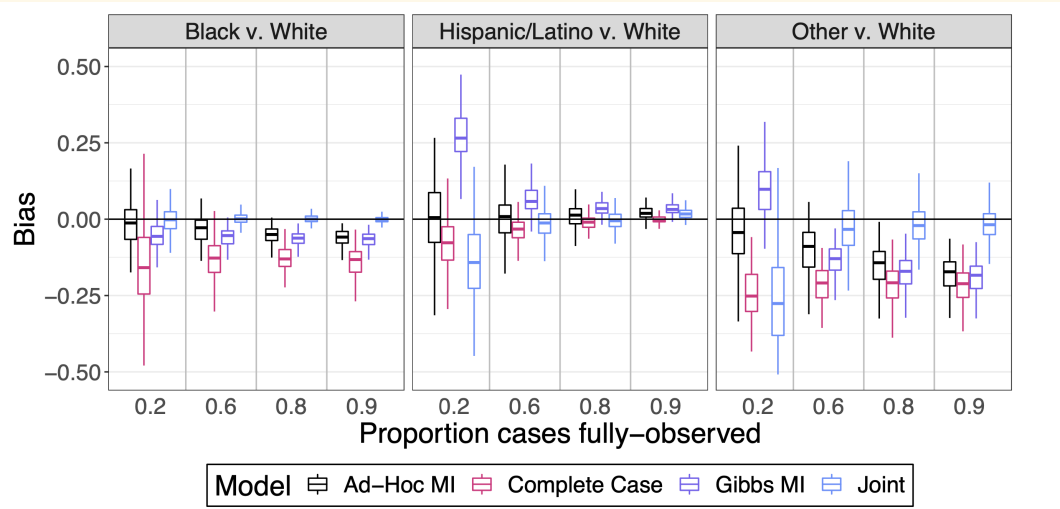
Compare 3 modeling approaches

- Complete case model
  - discard observations with missing datas
  - fit data using a Poisson GLM
- Missing at Random
  - first impute race/ethnicity from other variables,
    use 2 different imputation techniques: "Ad-Hoc MI" and "Gibbs MI"
  - fit imputed dataset using above Poisson GLM
- Not Missing at Random
  - tabulate data: complete case counts, missing race/ethnicity case counts
  - fit tabulated data using a joint hierarchical model

Trangucci et al - Fig. 1 Model Bias for Disease Prevalence



Trangucci et al - Fig. 2 Model Bias for Relative Risk



Modeling Disease Prevalence

When y is the number of cases and E is the size of the general population, we model the per-capita disease rate as:

$$y/E \sim Poisson(\lambda)$$

- Parameter λ is positive; it is the average expected number of events in an interval.
- The conjugate prior for λ is a Gamma distribution (shape parameter α, rate parameter β).

The Poisson GLM adds a set of linear predictors to the model via the exponential link function

$$\lambda = exp(X\,\beta)$$

With a little algebra, this becomes:

$$y \sim log\_poisson(log\_E + X\,\beta)$$

Modeling Missingness

- Code missingness as a binary indicator: 1 if present, and 0 if missing.
- Use the Bernoulli distribution for a single binary trail and the Binomial distribution for a series of trials.
  - The Bernoulli has a single parameter p, probability of trial outcome 1.
  - The Binomial has additional parameter, N, the number of trials

$$\text{complete-cases} \sim \text{Binomial}(N, p)$$

$$\text{incomplete-cases} \sim \text{Binomial}(N, (1 - p))$$

- Parameter p is a the probability of success, range 0-1.
- Conjugate prior for p is the Beta distribution with parameters α and β which represent prior (or pseudo) observations.
  Beta(1, 1) is the equivalent of a uniform prior.

Modeling Missingness

Tabulate by total positive tests per stratum, resulting in two sets of counts:

- Matrix X of counts over I strata by J race/ethnicity, where race/ethnicity is known.
- Vector M of counts over I stata where race/ethnicity is missing.

Compute likelihoods separately

$$X_{ij} \mid p_{ij}, \lambda_{ij}, E_{ij} \sim \text{Poisson}(p_{ij}\lambda_{ij}E_{ij})$$

$$M_i \mid (p_{i1}, \lambda_{i1}), \ldots, (p_{iJ}, \lambda_{iJ}), E_{ij} \sim \text{Poisson}\left(\sum_j (1 - p_{ij})\, \lambda_{ij}\, E_{ij}\right)$$

- For X, weight Poisson variate μ_{ij}, the per-capita disease-rate by p_{ij}
- For M, marginalize over $1 - p_{ij}$
  - Stan does not support sampling discrete parameters. See: *Stan User's Guide*

Trangucci et al - Missingness independent of strata

Section 2.1 presents the simplest possible model

- Missingness at the individual level depends only on race/ethnicity: $p_{ij} = p_j, \forall i$
- Infection probability is constant across strata: $\mu_{ij} = \lambda_j E_{ij}$

This simplifies the model; letting $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$ and $\boldsymbol{p} = (p_1, \ldots, p_J)$

$$X_{ij} \mid p_j, \lambda_j, E_{ij} \sim \text{Poisson}(p_j \lambda_j E_{ij})$$

$$M_i \mid \boldsymbol{p}, \boldsymbol{\lambda}, E_{ij} \sim \text{Poisson}\left(\sum_j (1 - p_j)\, \lambda_j\, E_{ij}\right)$$

Trangucci et al - Priors for Model 2.1

- Put informative priors on all variables.
- Choice of hyper-priors given as data.

$$p_j \overset{\text{iid}}{\sim} \text{Beta}(\alpha_j, \beta_j)$$
$$\lambda_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_j + \beta_j, r_j)$$

Trangucci et al - Missingness varies by race, age-sex strata

Model 2.2

$$X_{ij} \mid \lambda_j, z_i, \beta, p_{ij}, E_{ij} \sim \text{Poisson}(\exp(z_i^T \beta) p_{ij} \lambda_j E_{ij})$$

$$M_i \mid \lambda, z_i, \beta, p_i, E_i \sim \text{Poisson}\left(\exp(z_i^T \beta) \sum_j (1 - p_{ij}) \lambda_j E_{ij}\right)$$

$$p_{ij} = \left(1 + \exp\left(-(z_i^T \gamma + \eta_j)\right)\right)^{-1}$$

- Assume no interaction between race and age-sex strata for predicting incidence (λ) and missingness (p)
- Several conditions on amount data - Theorem 2.2
- More priors, hyperpriors

Trangucci et al - Theorem 2.2

THEOREM 2.2. *Let the model be defined as in (10) and let* $\mathbf{E}$ *be the* I *by* J *matrix where the* i*-th row is* $\mathbf{e}_i = (E_{i1}, E_{i2}, \ldots, E_{iJ})^T$*, and let* $\mathbf{Z}$ *be the* I *by* K *matrix where the* i*-th row is* $\mathbf{z}_i$*. If all of the following conditions hold:*

(S.a) $\mathbf{E}$ *is rank* J
(S.b) $\mathbf{Z}$ *is rank* K
(S.c) $I \geq J + K$
(S.d) $p_j \in (0, 1)$ *for all* j
(S.e) $\lambda_j \in (0, \infty)$ *for all* j
(S.f) $e^{\mathbf{z}_i \boldsymbol{\beta}} \sum_j^J E_{ij} \in (0, \infty)$ *for all* i
(S.g) $\operatorname{rank}\left(\left[\operatorname{diag}(\mathbf{E}_{[:,1]})\mathbf{Z} \ldots \operatorname{diag}(\mathbf{E}_{[:,J]})\mathbf{Z}\ \mathbf{E}_{[:,1]}\ \mathbf{E}_{[:,2]} \ldots\ \mathbf{E}_{[:,J]}\right]\right) > J + K$

the model is locally identifiable.

- "full rank" means that no rows are linear combinations of any other rows.
- E are per-strata, per-race population counts
- Z are per-strata, regression coefficients

Trangucci et al - Priors for Model 2.2

Given section 2.1.3, it is important to use prior information for minority groups when possible. To that end, the following priors can be employed:

$$\lambda_j \sim \text{LogNormal}(\mu_{\lambda_j}, s_\lambda^2) \quad \forall j \in [1, \ldots, J],$$

$$\eta_j \sim \text{Normal}(\mu_{\eta_j}, s_\eta^2) \quad \forall j \in [1, \ldots, J],$$

$$\boldsymbol{\beta} \sim \text{MultiNormal}(\boldsymbol{\mu}_\beta, \Sigma_\beta)$$

$$\boldsymbol{\gamma} \sim \text{MultiNormal}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma)$$

where $\mu_{\lambda_j}, \mu_{\eta_j}, s_\lambda, s_\eta, \boldsymbol{\mu}_\beta, \Sigma_\beta, \boldsymbol{\mu}_\gamma$ and Σ_γ are known hyperparameters.

2.1.3. Bayesian inference and prior sensitivity.

The two group setting in which one group's population is small compared to the other group's population motivates the careful choice of priors when doing Bayesian inference. . . . in this setting the posterior mean for the rate of disease in the minority group is sensitive to priors.

Trangucci et al - Model 2.3 -Adding Geo-location

Missingness depends on race, age-sex, geo-location

Model specification has more variables, more indices, more priors - *see Section 2.3!*

Conclusion

- Model the true data generating process
  - Socio-economic disparities in the system are reflected in the data quality
  - Missingness is ***not*** at random
- Establish model limits (proofs are less daunting than you think)
- Use simulated data to test your assumptions
- Use Stan to run simulations, fit models.
  - Models *are* complex and may take time to fit.
- Let's do better next time!

Thanks

Thanks to Rob Trangucci and Jon Zelner for many discussions.

Thank you Data Umbrella and R Ladies for hosting me.

Thank to CZI for support.

Thank you for being here!

References

- Trangucci, Rob, Yang Chen, and Jon Zelner. "Modeling rates of disease with missing categorical data." *arXiv preprint arXiv:2206.08161 (2022).*
- GitHub repository for example code from arxiv paper
- Stan User's Guide
- These slides

Questions?