Bayesian Spatio-Temporal Modelling of Disease Incidence with Nonignorable Missingness

Miles Moran

Oregon State University

November 4, 2024

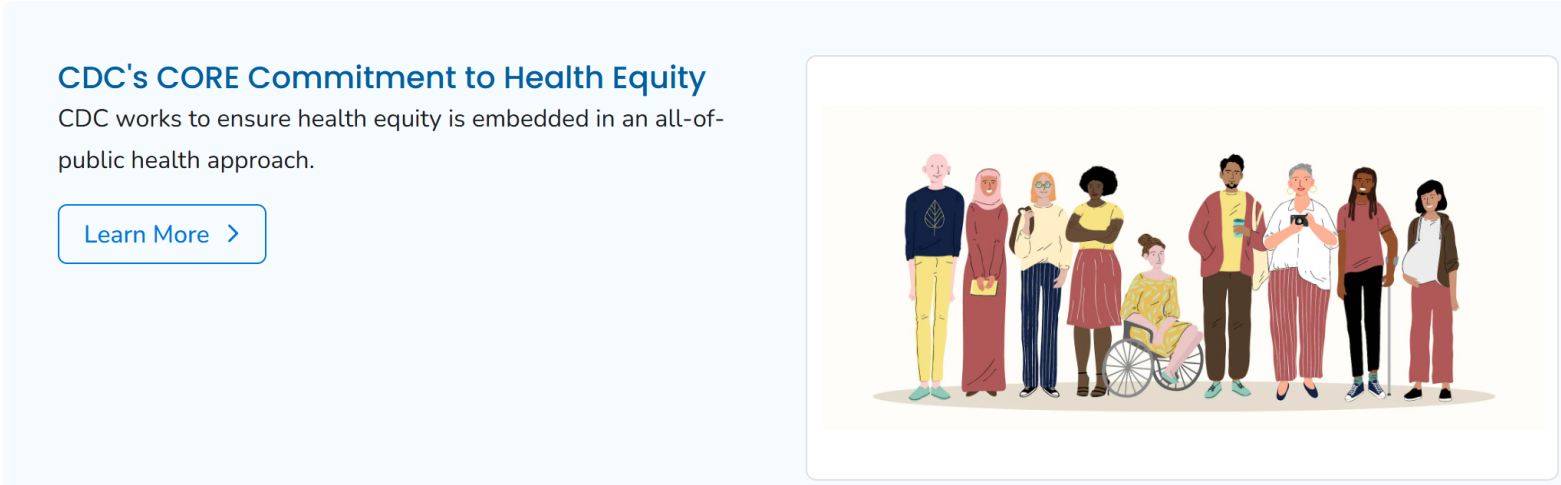
The Motivating Principle

CDC's CORE Commitment to Health Equity

CDC works to ensure health equity is embedded in an all-of-public health approach.

Learn More >

The Motivating Principle



CDC's CORE Commitment to Health Equity stands on four key pillars:

- **C**ultivating comprehensive health equity science
- **O**ptimizing interventions
- **R**einforcing and expanding robust partnerships
- **E**nhancing capacity and workplace diversity, inclusion, and engagement.

The Motivating Principle

CDC's CORE Commitment to Health Equity

CDC works to ensure health equity is embedded in an all-of-public health approach.

Learn More >

CDC's CORE Commitment to Health Equity stands on four key pillars:

- **C**ultivating comprehensive health equity science
- **O**ptimizing interventions
- **R**einforcing and expanding robust partnerships
- **E**nhancing capacity and workplace diversity, inclusion, and engagement.

What it Takes to Achieve Health Equity

- *More than ever before,*

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands
- *However,*

What it Takes to Achieve Health Equity

- *More than ever before*,
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands
- *However*,
  - The data are *incomplete* for some demographic covariates

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands
- *However,*
  - The data are *incomplete* for some demographic covariates
  - *If* these covariates are Missing Not-at-Random (MNAR), then models fit with imputed data (or complete-cases-only) will yield biased parameter estimates

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands
- *However,*
  - The data are *incomplete* for some demographic covariates
  - *If* these covariates are Missing Not-at-Random (MNAR), then models fit with imputed data (or complete-cases-only) will yield biased parameter estimates

Definition: MNAR Data

With $p = \Pr(\text{race is observed})$, we say that the race variable is Missing Not-at-Random (MNAR) if p depends on race itself or on other *un*observed variables

What it Takes to Achieve Health Equity

- *More than ever before,*
  - Disease surveillance data are more prevalent (and detailed!)
  - Models of disease spread (e.g. SIR-like) are well-studied
  - Modern computational capacity can meet our demands
- *However,*
  - The data are *incomplete* for some demographic covariates
  - *If* these covariates are Missing Not-at-Random (MNAR), then models fit with imputed data (or complete-cases-only) will yield biased parameter estimates
  - For many vulnerable or marginalized subpopulations, we can *expect* demographic membership to impact missingness ($\therefore$ MNAR)

Our Research Goals

- Develop a spatio-temporal model for the joint distribution of

$$X_{tgij} = \text{number of cases from } (t, g, i) \text{ observed with race } j$$
$$M_{tgi} = \text{number of cases from } (t, g, i) \text{ missing race}$$

for every time period t, geographic unit g, stratum i, and race j.

Our Research Goals

- Develop a spatio-temporal model for the joint distribution of

$$X_{tgij} = \text{number of cases from } (t, g, i) \text{ observed with race } j$$
$$M_{tgi} = \text{number of cases from } (t, g, i) \text{ missing race}$$

  for every time period t, geographic unit g, stratum i, and race j.
- Conduct a simulation study to demonstrate the model's advantage over SIR-like models that account for missingness in other ways (e.g. imputation or complete-case analysis)

Our Research Goals

- Develop a spatio-temporal model for the joint distribution of

$$X_{tgij} = \text{number of cases from } (t, g, i) \text{ observed with race } j$$
$$M_{tgi} = \text{number of cases from } (t, g, i) \text{ missing race}$$

  for every time period t, geographic unit g, stratum i, and race j.
- Conduct a simulation study to demonstrate the model's advantage over SIR-like models that account for missingness in other ways (e.g. imputation or complete-case analysis)
- Apply the model to real-world data (in our case, COVID-19 incidence in Michigan, disaggregated by race/ethnicity, age group, and sex)

Our Research Goals

- Develop a spatio-temporal model for the joint distribution of

$$X_{tgij} = \text{number of cases from } (t, g, i) \text{ observed with race } j$$
$$M_{tgi} = \text{number of cases from } (t, g, i) \text{ missing race}$$

  for every time period t, geographic unit g, stratum i, and race j.

- Conduct a simulation study to demonstrate the model's advantage over SIR-like models that account for missingness in other ways (e.g. imputation or complete-case analysis)
- Apply the model to real-world data (in our case, COVID-19 incidence in Michigan, disaggregated by race/ethnicity, age group, and sex)

Overview: Disease Process Models

- Compartment models where transition rates are based on assumptions of homogeneous mixing and the *law of mass action*

Overview: Disease Process Models

- Compartment models where transition rates are based on assumptions of homogeneous mixing and the *law of mass action*

- Deterministic Models: ODE-based, cts. in time. With infection rate β and recovery rate γ, under frequency-dependent transmission:

$$\begin{aligned}
\frac{dX(t)}{dt} &= -\frac{\beta X(t)Y(t)}{N} \\
\frac{dY(t)}{dt} &= \frac{\beta X(t)Y(t)}{N} - \gamma Y(t) \\
\frac{dZ(t)}{dt} &= \gamma Y(t)
\end{aligned}$$

Overview: Disease Process Models

- Compartment models where transition rates are based on assumptions of homogeneous mixing and the *law of mass action*

- Deterministic Models: ODE-based, cts. in time. With infection rate β and recovery rate γ, under frequency-dependent transmission:

$$\begin{aligned}\frac{dX(t)}{dt} &= -\frac{\beta X(t)Y(t)}{N} \\ \frac{dY(t)}{dt} &= \frac{\beta X(t)Y(t)}{N} - \gamma Y(t) \\ \frac{dZ(t)}{dt} &= \gamma Y(t)\end{aligned}$$

- Working with surveillance (i.e. population-level) data, We want a *stochastic*, *discrete-time* analog to this SIR model, e.g.

$$\begin{aligned}X_t &= X_{t-1} + B_{t-d} - Y_t \\ Y_t &\sim F(\cdots)\end{aligned}$$

Brief History of Relevant Stochastic Models

(1) Time-Series SIR ("TSIR") Models:

- Kendall (1949) *Stochastic Processes and Population Growth.*
- Bartlett (1956) *Deterministic and Stochastic Models for Recurrent Epidemics.*
- Bjørnstad, Finkenstädt, Grenfell (2002) *Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model*

Brief History of Relevant Stochastic Models

(1) Time-Series SIR ("TSIR") Models:
  - Kendall (1949) *Stochastic Processes and Population Growth.*
  - Bartlett (1956) *Deterministic and Stochastic Models for Recurrent Epidemics.*
  - Bjørnstad, Finkenstädt, Grenfell (2002) *Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model*

(2) Epidemic / Endemic (`hhh4/surveillance`) Models:
  - Held, Höhle, Hofmann (2005) *A Statistical Framework for the Analysis of Multivariate Infectious Disease Surveillance Counts*
  - Held & Paul (2012) *Modeling Seasonality in Space-Time Infectious Disease Surveillance Data*

Brief History of Relevant Stochastic Models

(1) Time-Series SIR ("TSIR") Models:
- Kendall (1949) *Stochastic Processes and Population Growth.*
- Bartlett (1956) *Deterministic and Stochastic Models for Recurrent Epidemics.*
- Bjørnstad, Finkenstädt, Grenfell (2002) *Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model*

(2) Epidemic / Endemic (`hhh4/surveillance`) Models:
- Held, Höhle, Hofmann (2005) *A Statistical Framework for the Analysis of Multivariate Infectious Disease Surveillance Counts*
- Held & Paul (2012) *Modeling Seasonality in Space-Time Infectious Disease Surveillance Data*

(3) Zelner Contact-Heterogeneity Models:
- Lloyd-Smith, et al. (2005) *Superspreading and the Effect of Individual Variation on Disease Emergence*
- Zelner et. al (2020) *Understanding the Importance of Contact Heterogeneity and Variable Infectiousness in the Dynamics of a Large Norovirus Outbreak*

Derivation #1: TSIR Models (1/2)

(1) Let Y_t denote # infected at time t. Denote $Y_0 = y_0$.

Derivation #1: TSIR Models (1/2)

(1) Let Y_t denote # infected at time t. Denote $Y_0 = y_0$.

(2) Suppose that each initial infected independently generates a tree of new infections according to a *linear birth process* over $(0, t)$:

$$\Pr\Big(\text{new infection in } (t, t+dt)\Big) = \lambda dt + o(dt)$$

Derivation #1: TSIR Models (1/2)

(1) Let Y_t denote # infected at time t. Denote $Y_0 = y_0$.

(2) Suppose that each initial infected independently generates a tree of new infections according to a *linear birth process* over $(0, t)$:

$$\Pr\Big(\text{new infection in } (t, t+dt)\Big) = \lambda dt + o(dt)$$

(3) Kendall (1949) show that the number of new infections from tree i is T_i (where a "success" = the single bernoulli event of *not* infecting someone):

$$(T_1, \ldots, T_{y_0}) \overset{\text{iid}}{\sim} \text{Geom}\Big(\underbrace{e^{-\lambda t}}_{\text{success prob}} \Big)$$

Derivation #1: TSIR Models (1/2)

(1) Let Y_t denote # infected at time t. Denote $Y_0 = y_0$.

(2) Suppose that each initial infected independently generates a tree of new infections according to a *linear birth process* over $(0, t)$:

$$\Pr\Big(\text{new infection in } (t, t+dt)\Big) = \lambda dt + o(dt)$$

(3) Kendall (1949) show that the number of new infections from tree i is T_i (where a "success" = the single bernoulli event of *not* infecting someone):

$$(T_1, \ldots, T_{y_0}) \overset{\text{iid}}{\sim} \text{Geom}\Big(\underbrace{e^{-\lambda t}}_{\text{success prob}} \Big)$$

(4) Following the properties of the Negative Binomial distribution,

$$(\underbrace{Y_t - y_0}_{\text{\# failures}}) = \Big(\textstyle\sum_{i=1}^{y_0} T_i\Big) \sim \text{NB}\Big(\underbrace{y_0}_{\text{\# successes}}, \underbrace{e^{-\lambda t}}_{\text{success prob}} \Big) \sim \text{NB}\Big(\underbrace{y_0(e^{\lambda t} - 1)}_{\text{mean}}, \underbrace{y_0}_{\text{dispersion}} \Big)$$

$$\underbrace{Y_t}_{\text{\# trials}} \sim \text{NB}\Big(\underbrace{y_0}_{\text{\# successes}}, \underbrace{e^{-\lambda t}}_{\text{success prob}} \Big) \sim \text{NB}\Big(\underbrace{y_0 e^{\lambda t}}_{\text{mean}}, \underbrace{y_0}_{\text{dispersion}} \Big)$$

Derivation #1: TSIR Models (2/2)

(5) Generalizing this formulation from the interval $(0, t)$ to our discretized intervals $(t-1, t)$, we can write

$$(Y_t - y_{t-1} \mid Y_{t-1} = y_{t-1}) \sim \mathrm{NB}\Big(y_{t-1}(e^{\lambda} - 1), y_{t-1}\Big)$$

and, if we make the simplifying assumption that prevalence = incidence (i.e. all infections recover before the next time period), then just

$$(Y_t \mid Y_{t-1} = y_{t-1}) \sim \mathrm{NB}\Big(y_{t-1}(e^{\lambda} - 1), y_{t-1}\Big)$$

Derivation #1: TSIR Models (2/2)

(5) Generalizing this formulation from the interval $(0, t)$ to our discretized intervals $(t-1, t)$, we can write

$$(Y_t - y_{t-1} \mid Y_{t-1} = y_{t-1}) \sim \mathrm{NB}\Big(y_{t-1}(e^{\lambda} - 1), y_{t-1}\Big)$$

and, if we make the simplifying assumption that prevalence = incidence (i.e. all infections recover before the next time period), then just

$$(Y_t \mid Y_{t-1} = y_{t-1}) \sim \mathrm{NB}\Big(y_{t-1}(e^{\lambda} - 1), y_{t-1}\Big)$$

(6) At this point, there's room for creativity in describing the hazard rate λ. Based on mass-action, we could take, e.g., $\lambda = \frac{\beta x_{t-1}}{N}$. so that

$$\begin{aligned} \mathrm{E}\left(Y_t \mid Y_{t-1} = y_{t-1}, X_{t-1} = x_{t-1}\right) &= y_{t-1} e^{\frac{\beta x_{t-1}}{N}} - y_{t-1} \\ &\approx \frac{\beta x_{t-1} y_{t-1}}{N} - y_{t-1} \qquad \text{for small } \beta x_{t-1}/N \end{aligned}$$

Derivation #2: Epidemic-Endemic / hhh4 Models

(1) Assume prevalence = incidence as in TSIR (fixed, unit-length infection time)

Derivation #2: Epidemic-Endemic / hhh4 Models

(1) Assume prevalence $=$ incidence as in TSIR (fixed, unit-length infection time)

(2) Assume basic survival model for susceptibles: constant hazard for each time period so that the time until a susceptible is infected is exponential:

$$\Pr\Big(\text{infection in } (t-1,t] \;\Big|\; \text{no infection by } t-1\Big) = 1 - e^{-\lambda_t}$$

Derivation #2: Epidemic-Endemic / hhh4 Models

(1) Assume prevalence = incidence as in TSIR (fixed, unit-length infection time)

(2) Assume basic survival model for susceptibles: constant hazard for each time period so that the time until a susceptible is infected is exponential:

$$\Pr\Big(\text{infection in } (t-1,t] \;\Big|\; \text{no infection by } t-1\Big) = 1 - e^{-\lambda_t}$$

(3) Treating each susceptible as an independent Bernoulli trial, the number of new cases at time t becomes

$$\begin{aligned}(Y_t \mid X_{t-1} = x_{t-1}) &\sim \text{Binom}\left(x_{t-1}, 1 - e^{-\lambda_t}\right) \\ &\dot{\sim} \text{Pois}\left(x_{t-1}\lambda_t\right) \qquad \text{when } x_{t-1} \text{ large and } \lambda_t \text{ small}\end{aligned}$$

Derivation #2: Epidemic-Endemic / hhh4 Models

(1) Assume prevalence = incidence as in TSIR (fixed, unit-length infection time)

(2) Assume basic survival model for susceptibles: constant hazard for each time period so that the time until a susceptible is infected is exponential:

$$\Pr\Big(\text{infection in } (t-1,t] \;\Big|\; \text{no infection by } t-1\Big) = 1 - e^{-\lambda_t}$$

(3) Treating each susceptible as an independent Bernoulli trial, the number of new cases at time t becomes

$$\begin{aligned}(Y_t \mid X_{t-1} = x_{t-1}) &\sim \text{Binom}\left(x_{t-1}, 1 - e^{-\lambda_t}\right) \\ &\dot{\sim} \text{Pois}\left(x_{t-1}\lambda_t\right) \qquad \text{when } x_{t-1} \text{ large and } \lambda_t \text{ small}\end{aligned}$$

(4) Again, there's room for creativity in describing the hazard rate λ. Based on mass-action, we arrive at

$$(Y_t \mid Y_{t-1} = y_{t-1}, X_{t-1} = x_{t-1},) \dot{\sim} \text{Pois}\left(\frac{\beta x_{t-1} y_{t-1}}{N}\right)$$

Our Research Goals

- Develop a spatio-temporal model for the joint distribution of

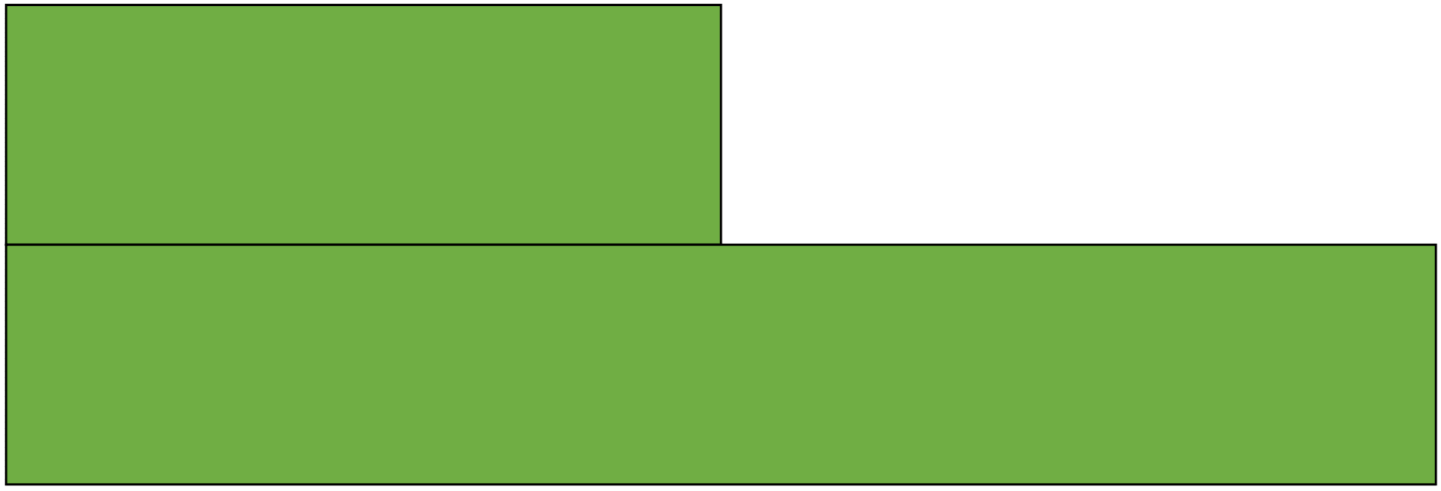
$$X_{tgij} = \text{number of cases from } (t, g, i) \text{ observed with race } j$$
$$M_{tgi} = \text{number of cases from } (t, g, i) \text{ missing race}$$

  for every time period t, geographic unit g, stratum i, and race j.

- Conduct a simulation study to demonstrate the model's advantage over SIR-like models that account for missingness in other ways (e.g. imputation or complete-case analysis)
- Apply the model to real-world data (in our case, COVID-19 incidence in Michigan, disaggregated by race/ethnicity, age group, and sex)

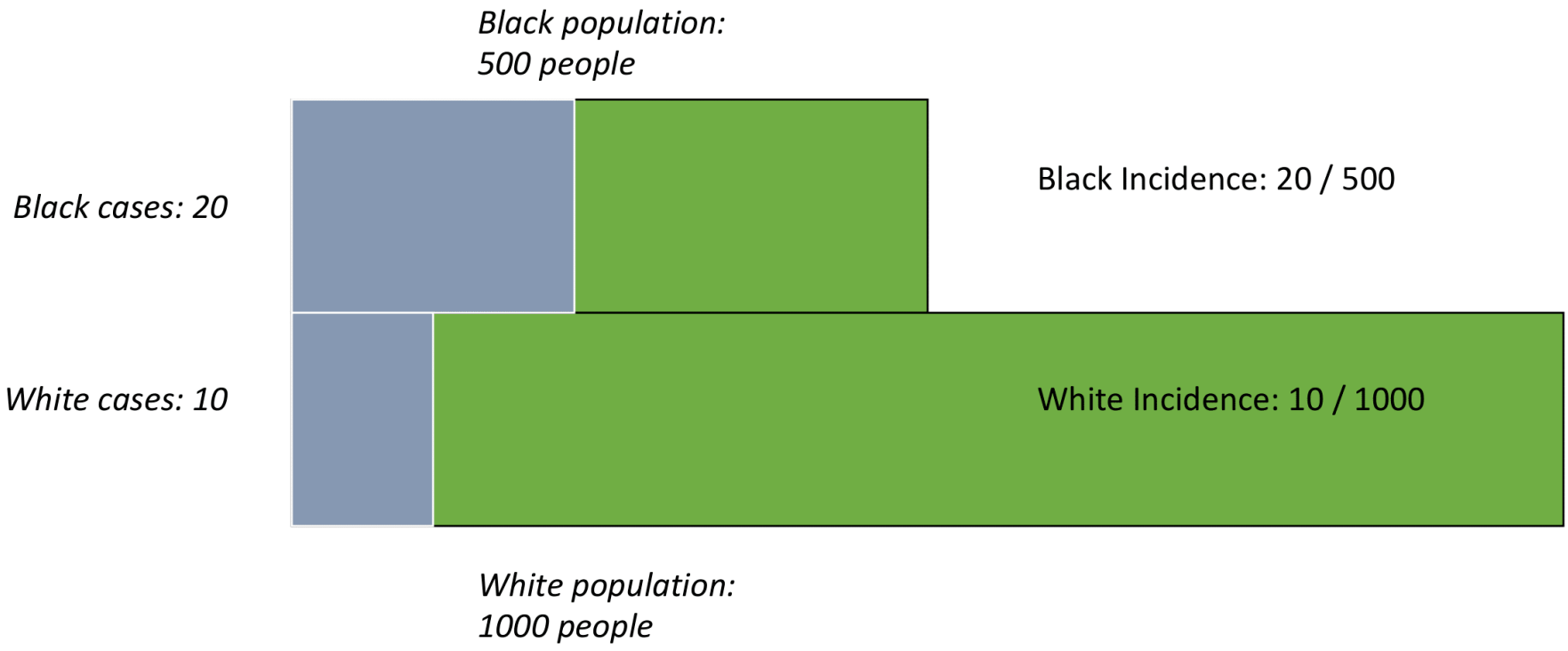
Missingness in the Simplest Case

Black population:
500 people

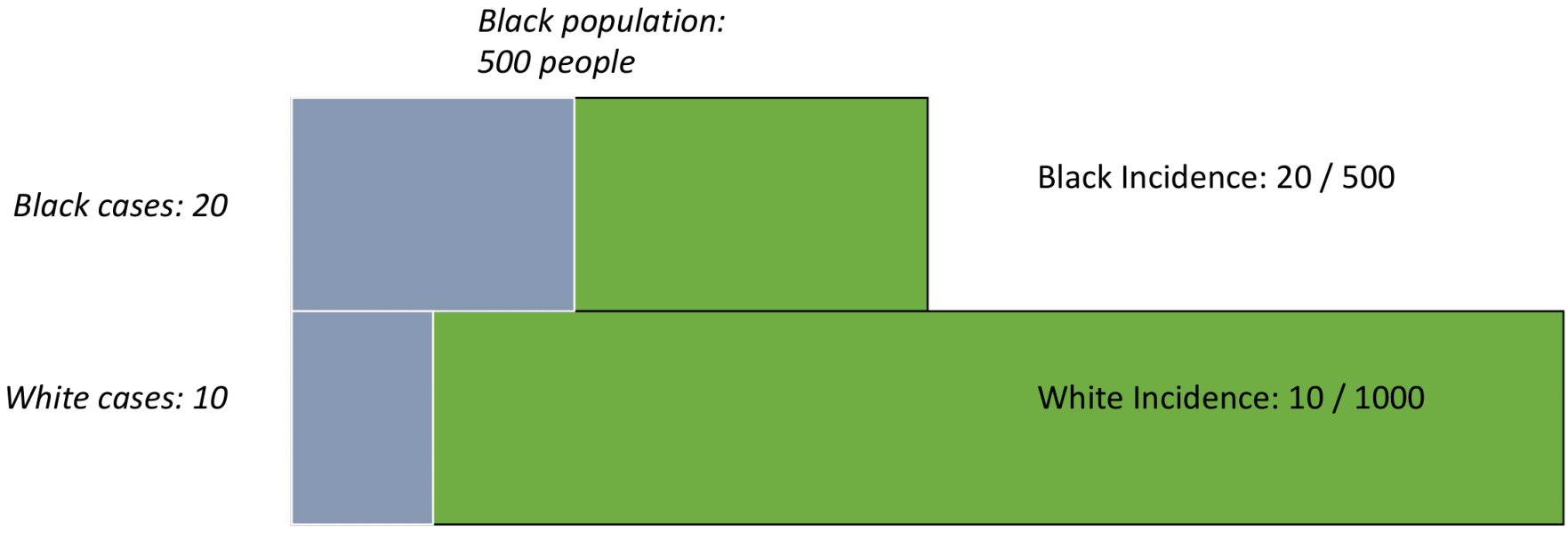


White population:
1000 people

Missingness in the Simplest Case



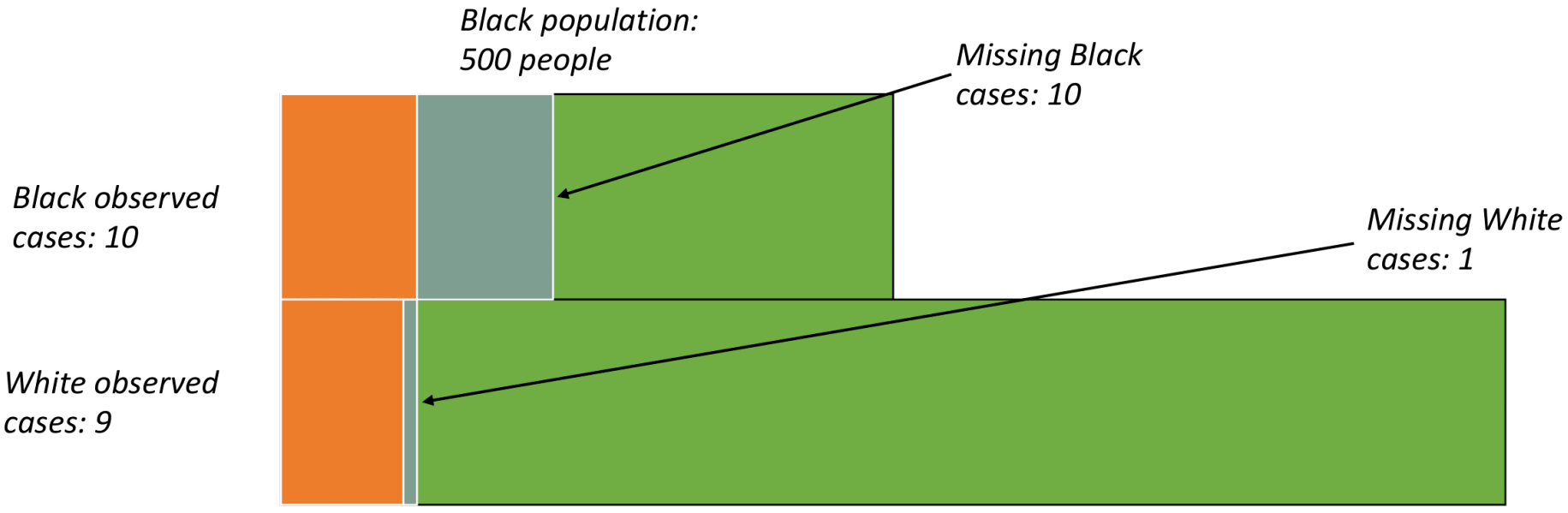
Missingness in the Simplest Case



True risk of Black residents relative to White residents:

20 / 500 / (10 / 1000) = 4

Missingness in the Simplest Case



Observed risk of Black residents relative to White residents:

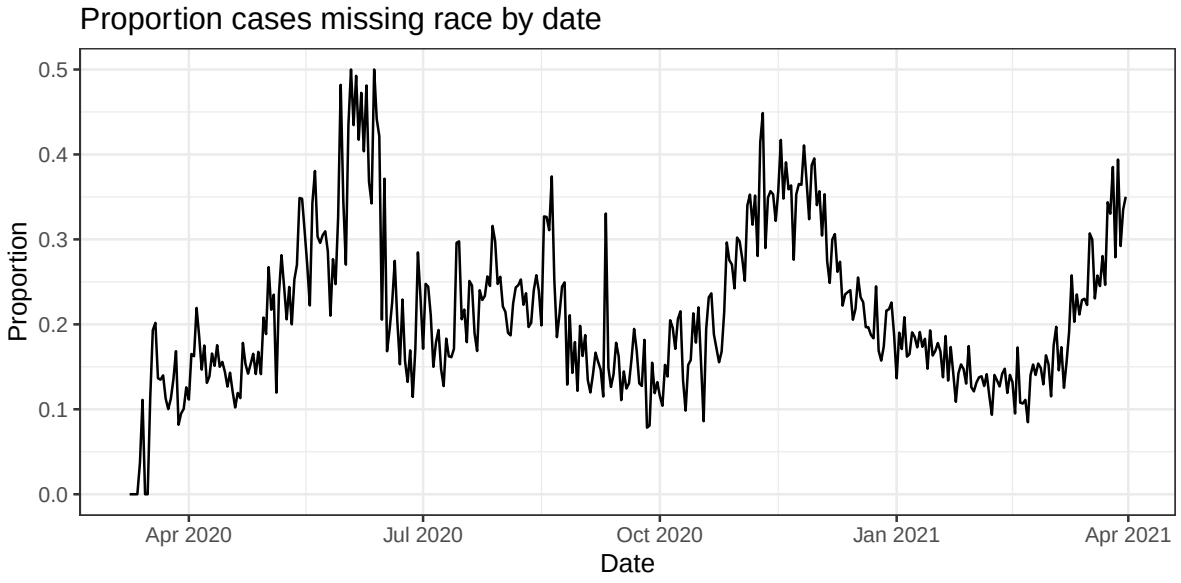
$10 / 500 / (9 / 1000) \cong 2$

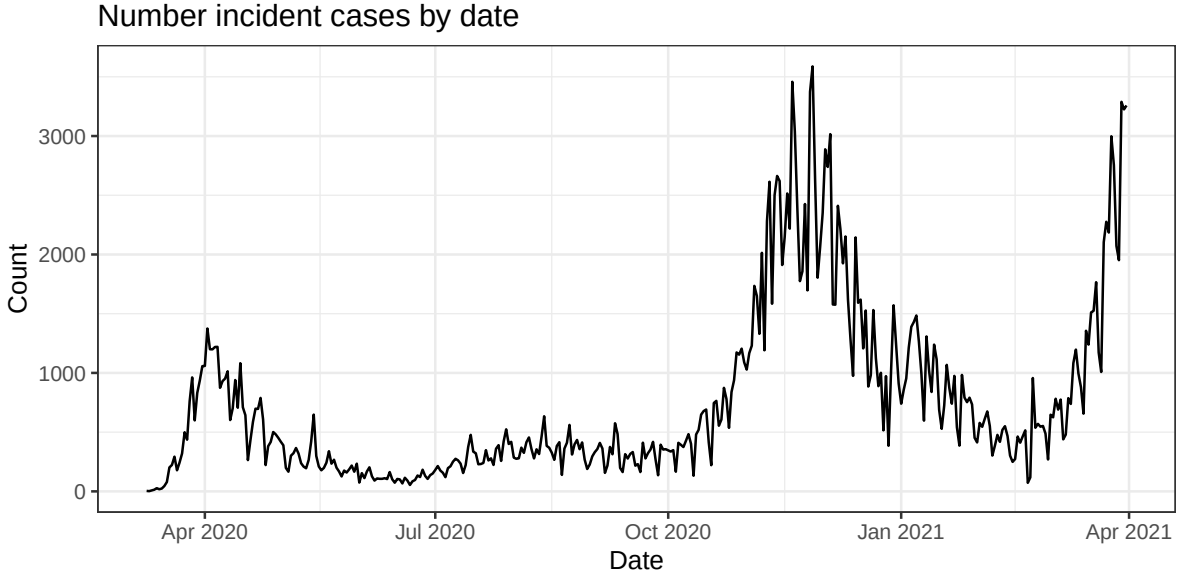
Missingness in the Simplest Case

Observed risk of Black residents relative to White residents:

$10 / 500 / (9 / 1000) \cong 2$

Missingness in Real Life





Overview: Missingness Process Models

- At time t, for individual k from race j, let
  - Y_{tjk} denote presence/absence of infection (binary)
  - M_{tjk} denote missingness (binary)

  so that Y_{tjk} is observed only when $M_{tjk} = 1$

- A model for data subject to missingness is just a specification of $f(\mathbf{y}, \mathbf{m})$, which is done through some kind of decomposition:

Selection Factorization: $f(\mathbf{y}, \mathbf{m}) = f_1(\mathbf{y}) f_2(\mathbf{m}|\mathbf{y})$

Pattern-Mixture Factorization: $f(\mathbf{y}, \mathbf{m}) = g_1(\mathbf{m}) g_2(\mathbf{y}|\mathbf{m})$

Random Effects: $f(\mathbf{y}, \mathbf{m}) = \int h_1(\mathbf{y}|\mathbf{u}) h_2(\mathbf{m}|\mathbf{u}) h_3(\mathbf{u}) d\mathbf{u}$

Derivation #3: Missingness in a TSIR-like Model

(1) Model *true* incidence-by-race with a discrete-time 1st-order Markov model, i.e. $(Y_{tgij}|\mathbf{Y}_{(t-1)}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij}\right)$

Derivation #3: Missingness in a TSIR-like Model

(1) Model *true* incidence-by-race with a discrete-time 1st-order Markov model, i.e. $(Y_{tgij}|\mathbf{Y}_{(t-1)}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij}\right)$

(2) Model *observed* incidence-by-race as independent draws from the population of infected, i.e. $(X_{tgij}|Y_{tgij}) \sim \text{Binom}\left(Y_{tgij}, p_{tgj}\right)$

Derivation #3: Missingness in a TSIR-like Model

(1) Model *true* incidence-by-race with a discrete-time 1st-order Markov model, i.e. $(Y_{tgij}|\mathbf{Y}_{(t-1)}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij}\right)$

(2) Model *observed* incidence-by-race as independent draws from the population of infected, i.e. $(X_{tgij}|Y_{tgij}) \sim \text{Binom}\left(Y_{tgij}, p_{tgj}\right)$

(3) Marginalize over Y_{tgij} to obtain the *observational* model:

$$\left\{ \begin{array}{l} X_{tgij} \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgij}\right) \\ M_{tgi} = Y_{tgi\bullet} - \sum_{j=1}^{J} X_{tgij} \\ \qquad\quad \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1 - p_{tgij})\right) \end{array} \right\}$$

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$
$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$
$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgj}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{j} + \beta^{\text{AR}}_{g}$$
$$\log\left(\lambda^{\text{NE}}_{tgj}\right) = \mu^{\text{NE}} + \alpha^{\text{NE}}_{j} + \beta^{\text{NE}}_{g}$$
$$\log\left(\lambda^{\text{EN}}_{tgj}\right) = \mu^{\text{EN}} + \alpha^{\text{EN}}_{j} + \beta^{\text{EN}}_{g} + \gamma^{\text{EN}} t + \delta^{\text{EN}} \sin(\tfrac{t}{52} 2\pi) + \varepsilon^{\text{EN}} \cos(\tfrac{t}{52} 2\pi)$$
$$\text{logit}\left(p_{tgj}\right) = \mu^{(\text{p})} + \alpha^{(\text{p})}_{j} + \beta^{(\text{p})}_{g}$$

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$
$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$
$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgj}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{j} + \beta^{\text{AR}}_{g}$$
$$\log\left(\lambda^{\text{NE}}_{tgj}\right) = \mu^{\text{NE}} + \alpha^{\text{NE}}_{j} + \beta^{\text{NE}}_{g}$$

The Data (Part 1)

$(t, g, i, j) = (\textit{time},\ \textit{location},\ \textit{stratum},\ \textit{race})$

X_{tgij} = number of cases from (t, g, i) observed with race j

M_{tgi} = number of cases from (t, g, i) missing race

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$

$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$

$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgj}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{j} + \beta^{\text{AR}}_{g}$$

$$\log($$

$$\log($$

$$\text{logit}($$

The Data (Part 2)

$Y_{tgi\bullet}$ = *true* number of cases from (t, g, i)

$Y_{tg\bullet\bullet}$ = *true* number of cases from (t, g)

$\mathbf{Y}_{tgi\bullet} = (Y_{tg1\bullet}, Y_{tg2\bullet}, ..., Y_{tgI\bullet})$

Note: $Y_{tgi\bullet}, Y_{tg\bullet\bullet}$ are observed, but Y_{tgij} aren't

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$

$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$

$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{t\cdot j}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{\cdot} + \beta^{\text{AR}}$$

$$\log\left(\lambda^{\text{NE}}_{t\cdot j}\right)$$

$$\log\left(\lambda^{\text{EN}}_{t\cdot j}\right)$$

$$\text{logit}\left(p_{\cdot}\right)$$

The Data (Part 3)

$w_{gg'}$ = distance weight between locations g and g'

$$= \begin{cases} \frac{1}{(\#\text{ neighbors})_g} & \text{if } g \text{ adjacent to } g' \\ 0 & \text{if } g \textit{ not} \text{ adjacent to } g' (\text{or } g = g') \end{cases}$$

E_{gij} = population count from (g, i, j)

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$

$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$

$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgi}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{i} + \beta^{\text{AR}}_{g}$$

The Disease Process (Competing Risks Framework)

$\lambda^{\text{TOT}}_{tgij}$ = hazard rate from all sources

$\lambda^{\text{AR}}_{tgj}$ = hazard rate of self-area/"autoregressive" infections

$\lambda^{\text{NE}}_{tgj}$ = hazard rate of neighboring-area infections

$\lambda^{\text{EN}}_{tgj}$ = hazard rate of background/"environmental" infection

EE-Like Observational Model Specification

$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$
$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$
$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgj}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{j} + \beta^{\text{AR}}_{g}$$
$$\log\left(\lambda^{\text{NE}}_{tgj}\right) = \mu^{\text{NE}} + \alpha^{\text{NE}}_{j} + \beta^{\text{NE}}_{g}$$
$$\log\left(\lambda^{\text{EN}}_{tgj}\right) = \mu^{\text{EN}} + \alpha^{\text{EN}}_{j} + \beta^{\text{EN}}_{g} + \gamma^{\text{EN}} t + \delta^{\text{EN}} \sin(\tfrac{t}{52} 2\pi) + \varepsilon^{\text{EN}} \cos(\tfrac{t}{52} 2\pi)$$

The Missingness Process

p_{tgj} = probability that a case from (t, g, j) reports their race

EE-Like Observational Model Specification

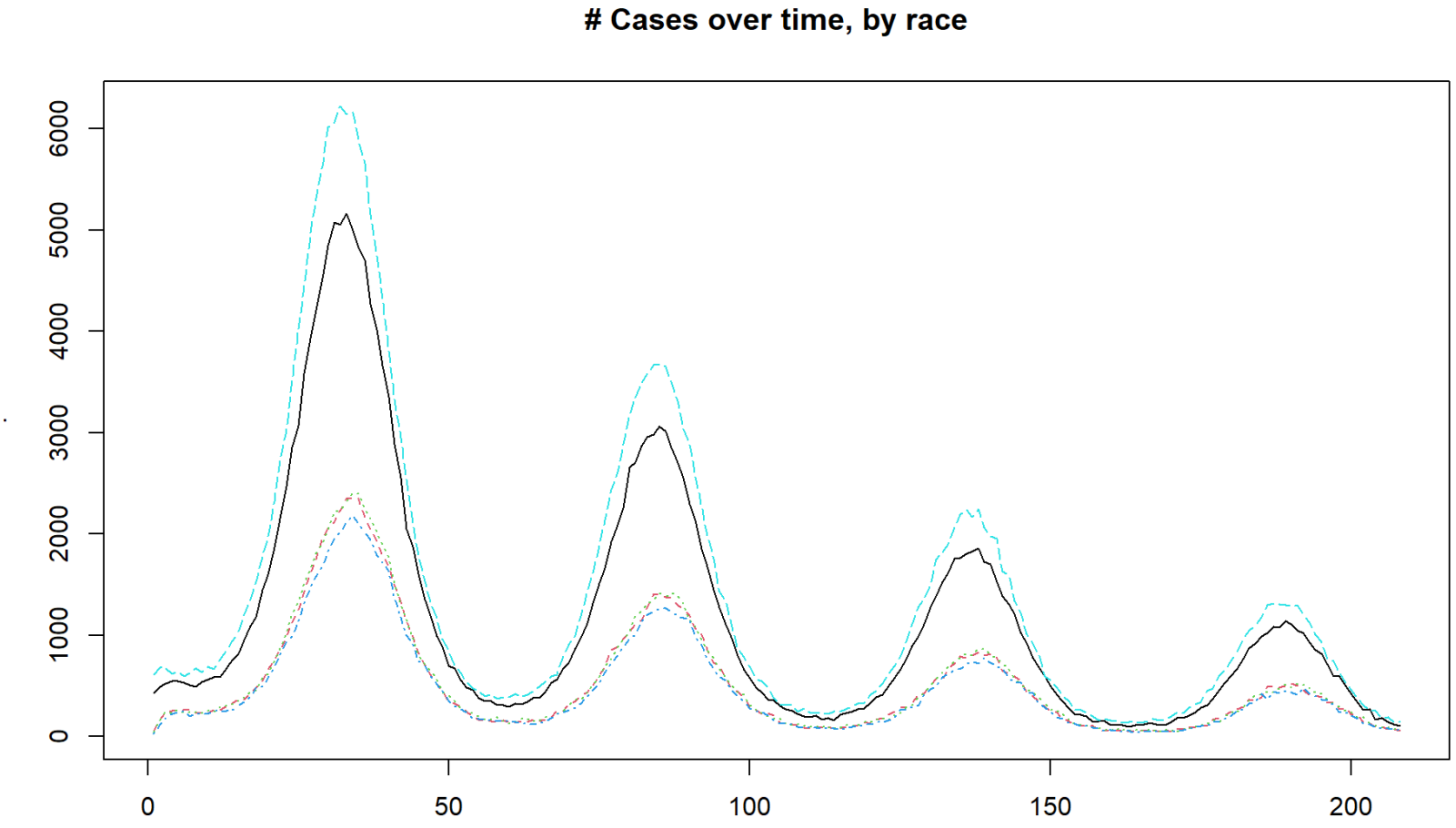
$$(X_{tgij}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\lambda^{\text{TOT}}_{tgij} p_{tgj}\right)$$
$$(M_{tgi}|\mathbf{Y}_{(t-1)gi\bullet}) \sim \text{Pois}\left(\sum_{j=1}^{J} \lambda^{\text{TOT}}_{tgij}(1-p_{tgj})\right)$$
$$\lambda^{\text{TOT}}_{tgij} = \lambda^{\text{AR}}_{tgj} Y_{(t-1)gi\bullet} + \lambda^{\text{NE}}_{tgj} \sum_{g'=1}^{G} w_{gg'} Y_{(t-1)g'\bullet\bullet} + \lambda^{\text{EN}}_{tgj} E_{gij}$$

$$\log\left(\lambda^{\text{AR}}_{tgj}\right) = \mu^{\text{AR}} + \alpha^{\text{AR}}_{j} + \beta^{\text{AR}}_{g}$$
$$\log\left(\lambda^{\text{NE}}_{tgj}\right) = \mu^{\text{NE}} + \alpha^{\text{NE}}_{j} + \beta^{\text{NE}}_{g}$$
$$\log\left(\lambda^{\text{EN}}_{tgj}\right) = \mu^{\text{EN}} + \alpha^{\text{EN}}_{j} + \beta^{\text{EN}}_{g} + \gamma^{\text{EN}} t + \delta^{\text{EN}} \sin(\tfrac{t}{52} 2\pi) + \varepsilon^{\text{EN}} \cos(\tfrac{t}{52} 2\pi)$$
$$\text{logit}\left(p_{tgj}\right) = \mu^{(\text{p})} + \alpha^{(\text{p})}_{j} + \beta^{(\text{p})}_{g}$$

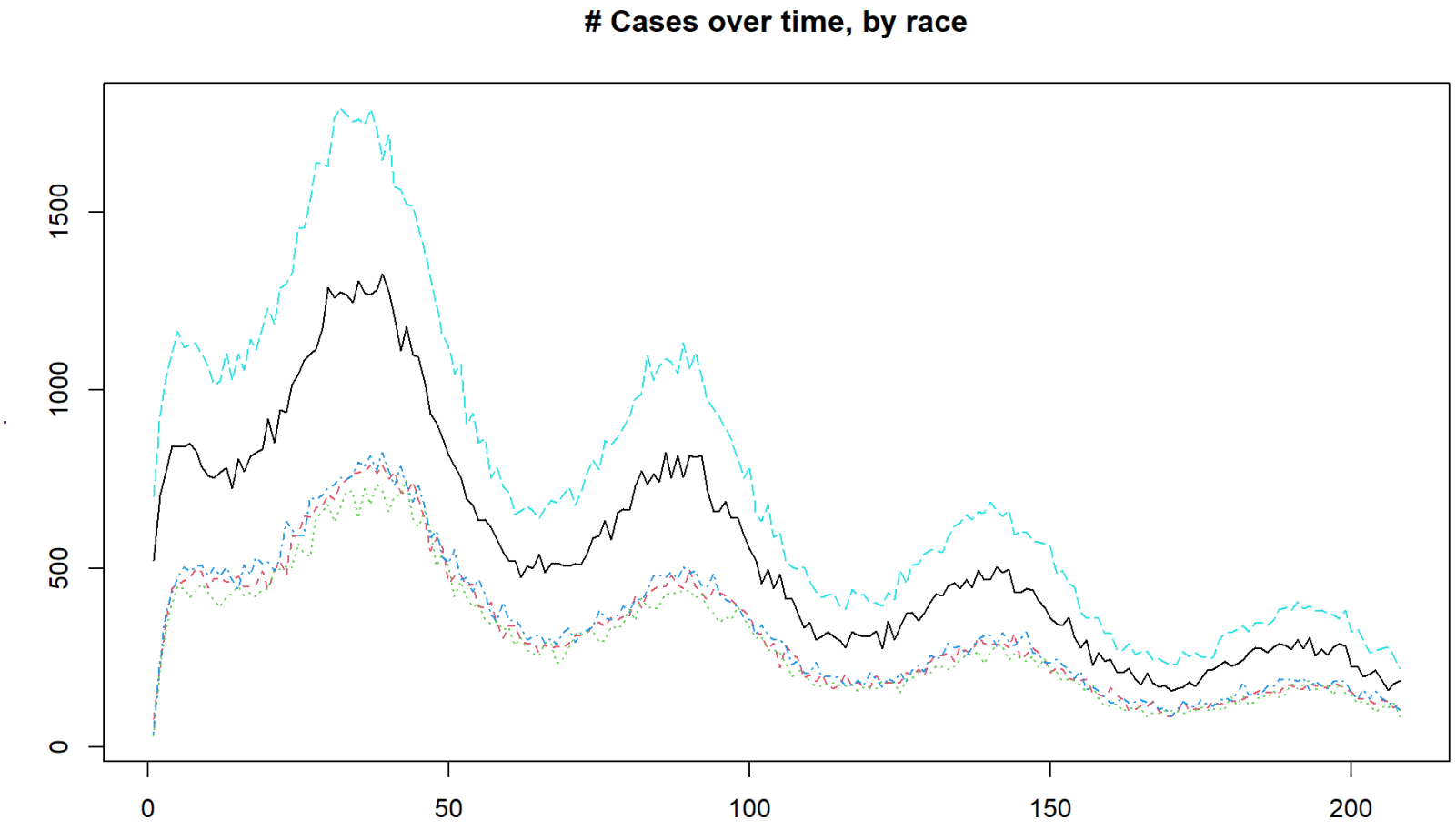
Simulation Study Overview (Work-In-Progress)

- Data generated according to the Latent Model in R
- Model fitting performed via HMC with Stan
- Scenarios to Consider:
  - Similar EE-type model, ignoring missingness entirely
  - Similar EE-type model, imputation via statistical model
  - Similar EE-type model, imputation via ML or MICE
  - TSIR models (i.e. Neg. Bin. likelihood) for each of the above

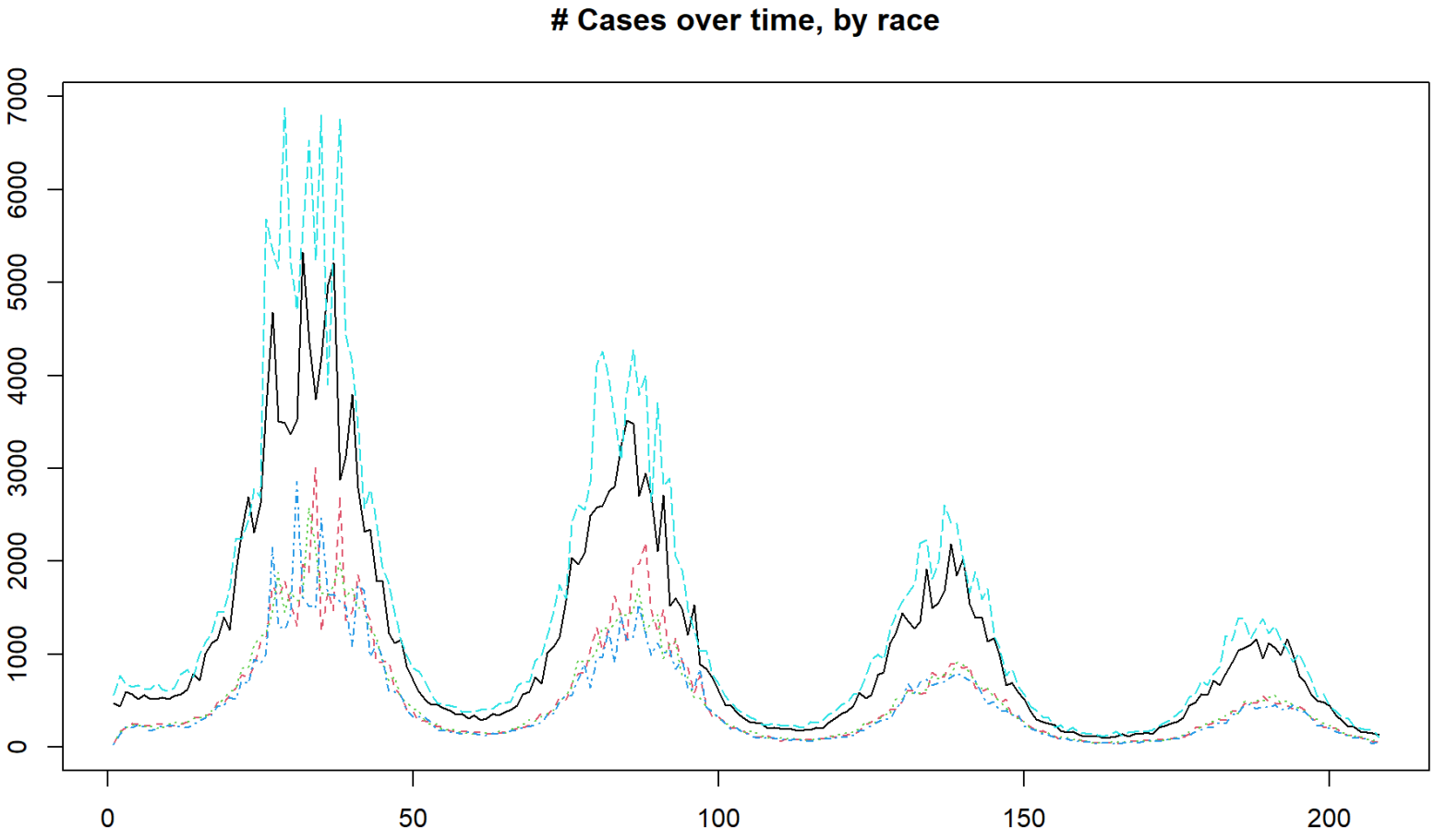
Simulated Data (Poisson)



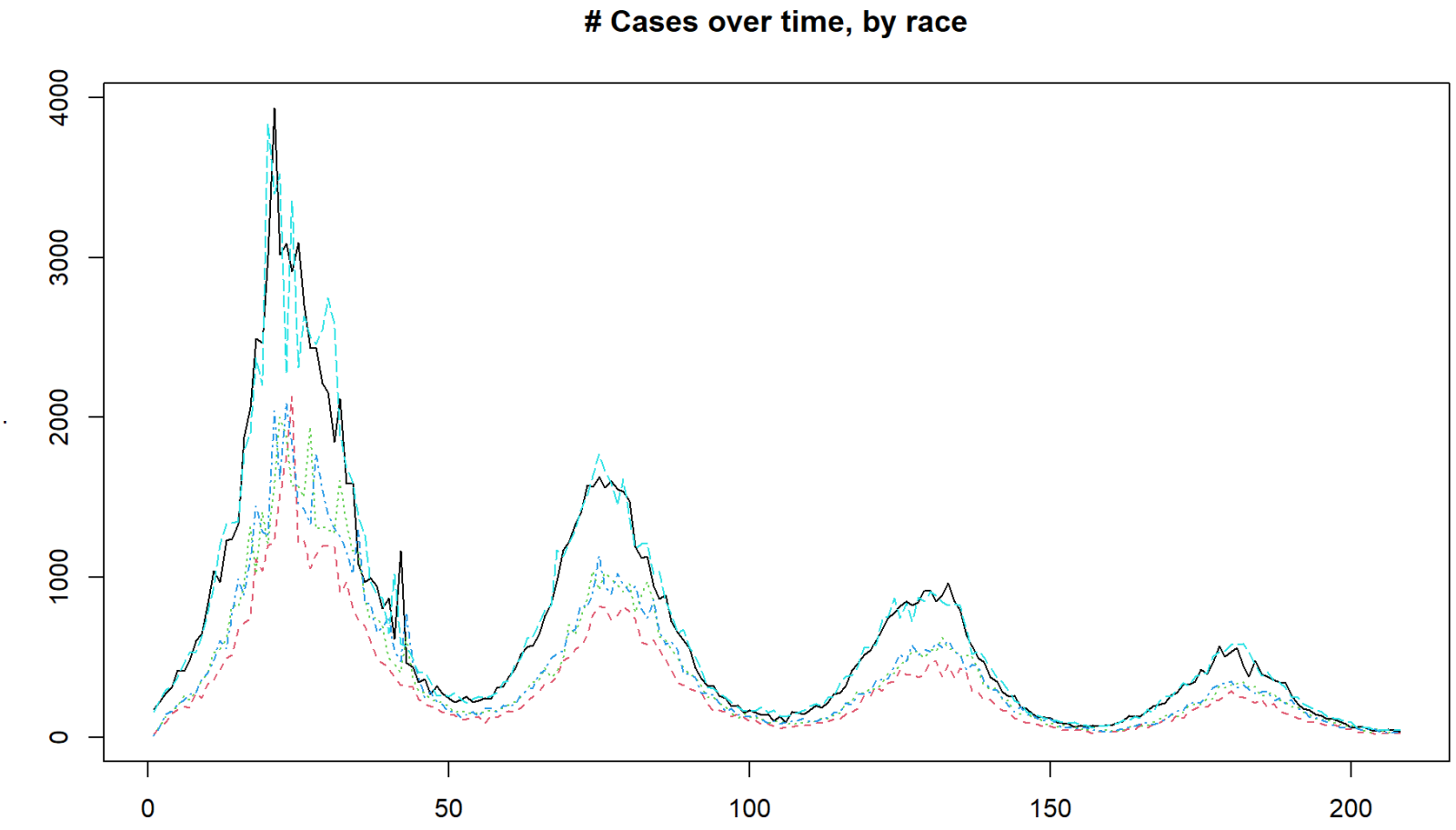
Simulated Data (Poisson)



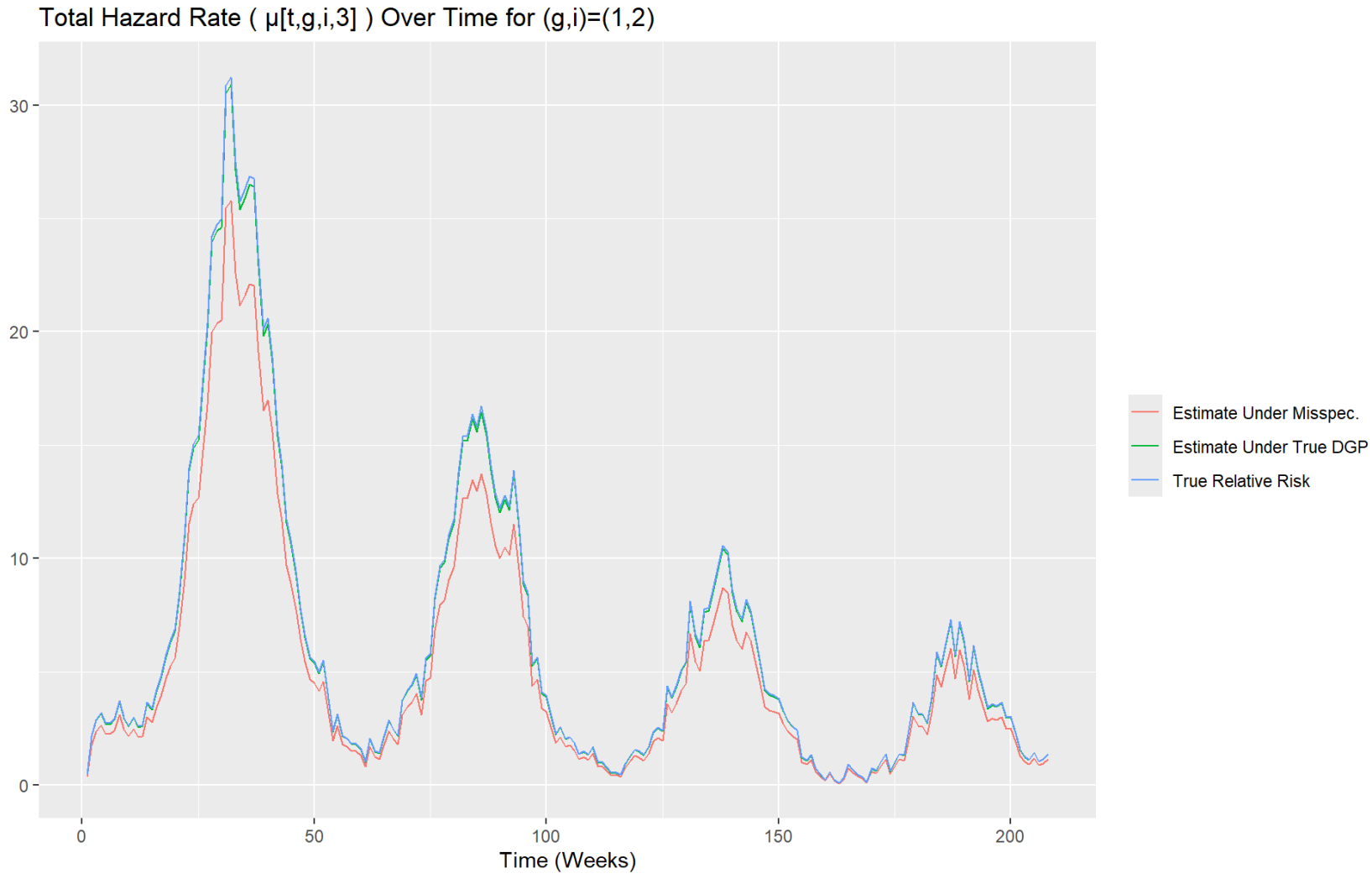
Simulated Data (Negative Binomial)



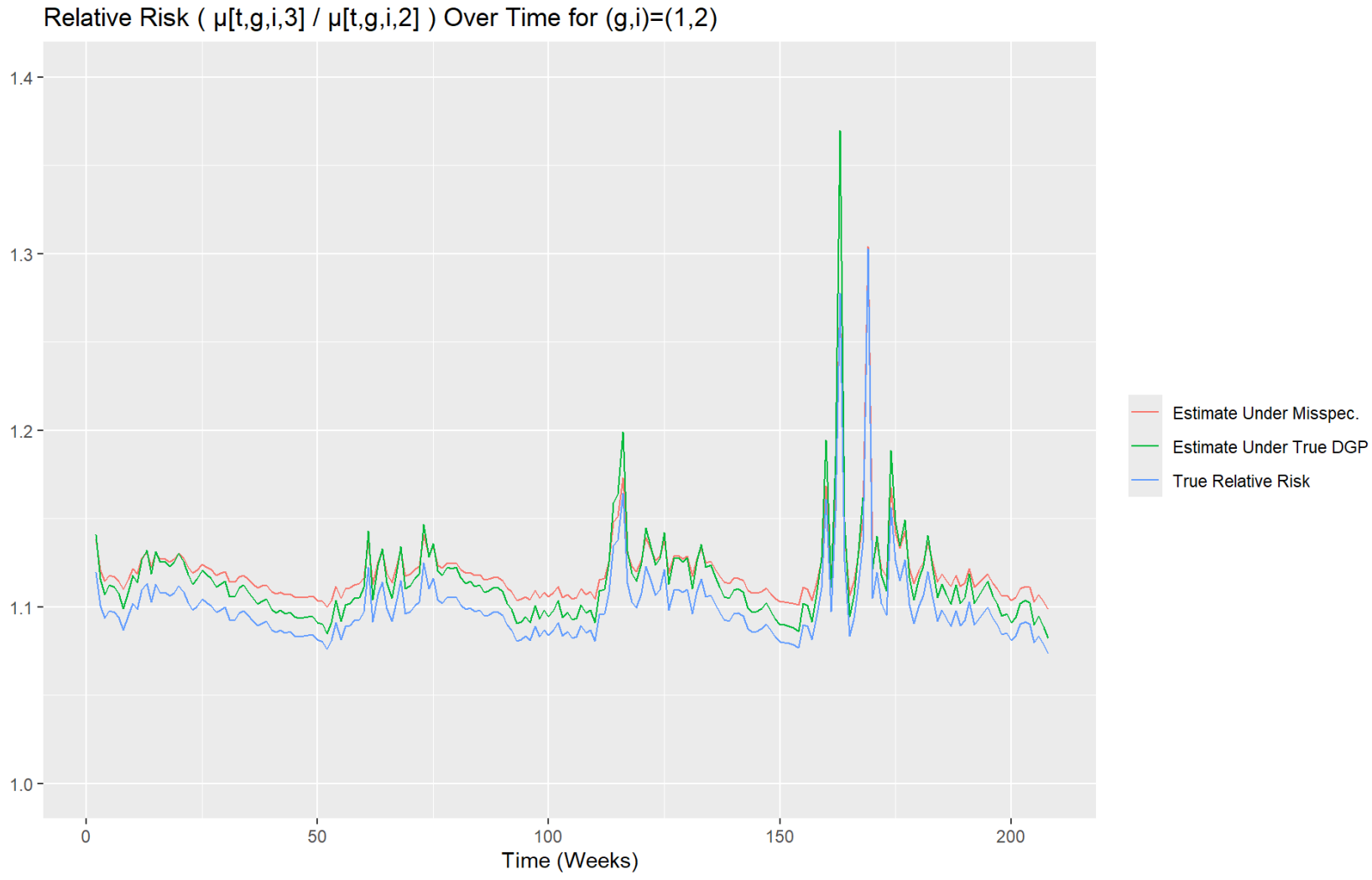
Simulated Data (Negative Binomial)



Preliminary Results (Idealized Scenario)

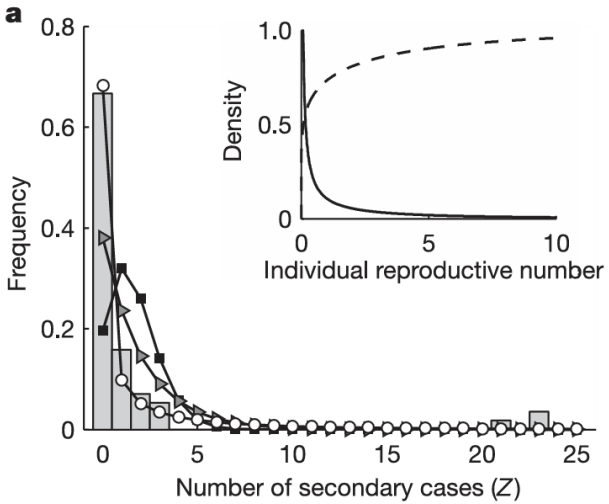


Preliminary Results (Idealized Scenario)



Motivation for Zelner Contact-Heterogeneity Model

- Observation: Individual-level contact-tracing data seems to suggest the average number of secondary-infections caused by an individual (i.e. R_0) exhibits individual-level heterogeneity (see Lloyd-Smith et al. 2005)



- Observation: If we wish to stratify a disease model by geography and demography, there is no obvious way in the aforementioned frameworks to specify how new infections are doled-out *across* strata

Derivation #4: Zelner Contact-Heterogeneity Model

(1) Suppose the (latent) individual R_0 of infected i in geography g during the interval $(t-1, t)$ is given by

$$r_{tgi} \overset{\text{iid}}{\sim} \text{Gamma}\Big(\underbrace{\tfrac{R_0}{\theta}}_{\text{shape}}, \underbrace{\theta}_{\text{scale}} \Big)$$

(2) The total (latent) infectiousness at time t for geography g becomes

$$(r_{tg\bullet} \mid Y_{tg} = y_{tg}) = \Big(\textstyle\sum_{i=1}^{y_{tg}} r_{ti}\Big) \sim \text{Gamma}\Big(\tfrac{R_0}{\theta} y_{tg}, \theta\Big)$$

(3) Assume homogeneous mixing, and that all of the latent infectiousness is deposited in the single time period after infection. Then the force of infection is

$$\lambda_{tg} = \zeta \frac{r_{(t-1)g}}{n_g} + (1-\zeta) \sum_{g' \neq g} \frac{r_{(t-1)g'}}{N - n_{g'}}$$

(4) Following a process almost identical to the EE model framework:

$$(Y_{tg} \mid \mathbf{r}_{(t-1)}) \sim \text{Pois}\left(n_g \lambda_{tg}\right)$$

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component
- Demonstrating efficacy will revolve around comparing relative risk estimates between different models and missing-data techniques

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component
- Demonstrating efficacy will revolve around comparing relative risk estimates between different models and missing-data techniques
- Future work includes

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component
- Demonstrating efficacy will revolve around comparing relative risk estimates between different models and missing-data techniques
- Future work includes
  - Determining parameterizations that dole-out infectiousness among strata in meaningful ways

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component
- Demonstrating efficacy will revolve around comparing relative risk estimates between different models and missing-data techniques
- Future work includes
  - Determining parameterizations that dole-out infectiousness among strata in meaningful ways
  - Determining conditions for local and global identifiability

Recap

- Going forward, achieving health equity will require more sophisticated methods for handling MNAR data
- Our attempt at this is to augment the EE-like and TSIR-like models of disease incidence with a selection-model component
- Demonstrating efficacy will revolve around comparing relative risk estimates between different models and missing-data techniques
- Future work includes
  - Determining parameterizations that dole-out infectiousness among strata in meaningful ways
  - Determining conditions for local and global identifiability
  - Determining validity of rare-disease assumption and consequences of violation

References I

Bartlett, M. S. (1956). "Deterministic and Stochastic Models for Recurrent Epidemics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health*. Vol. 3.4. University of California Press, pp. 81–110. (Visited on 11/03/2024).

Bauer, Cici and Jon Wakefield (2018). "Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in China". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.5, pp. 1379–1398. ISSN: 1467-9876. DOI: 10.1111/rssc.12284. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12284.

Bjørnstad, Ottar N., Bärbel F. Finkenstädt, and Bryan T. Grenfell (2002). "Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model". In: *Ecological Monographs* 72.2, pp. 169–184. ISSN: 1557-7015. DOI: 10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2.

References II

Held, Leonhard, Michael Höhle, and Mathias Hofmann (2005). "A statistical framework for the analysis of multivariate infectious disease surveillance counts". In: *Statistical Modelling* 5.3, pp. 187–199. ISSN: 1471-082X. DOI: 10.1191/1471082X05st098oa. URL: https://doi.org/10.1191/1471082X05st098oa.

Held, Leonhard and Michaela Paul (2012). "Modeling seasonality in space-time infectious disease surveillance data". In: *Biometrical Journal* 54.6, pp. 824–843. ISSN: 1521-4036. DOI: 10.1002/bimj.201200037. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201200037.

Kendall, David G. (1949). "Stochastic Processes and Population Growth". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 11.2, pp. 230–264. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1949.tb00032.x. (Visited on 11/03/2024).

References III

Lloyd-Smith, J. O. et al. (2005). "Superspreading and the effect of individual variation on disease emergence". In: *Nature* 438.7066, pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature04153. URL: https://doi.org/10.1038/nature04153.

Trangucci, Rob, Yang Chen, and Jon Zelner (2023). "Modeling racial/ethnic differences in COVID-19 incidence with covariates subject to nonrandom missingness". In: *The Annals of Applied Statistics* 17.4, pp. 2723–2758. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/22-AOAS1711. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-4/Modeling-racial-ethnic-differences-in-COVID-19-incidence-with-covariates/10.1214/22-AOAS1711.full (visited on 11/04/2024).

Wakefield, Jon, Tracy Qi Dong, and Vladimir N. Minin (2019). "Spatio-Temporal Analysis of Surveillance Data". In: *Handbook of Infectious Disease Data Analysis*. 1st. Chapman and Hall/CRC, pp. 455–475. ISBN: 978-1-315-22291-2.

References IV

Zelner, Jon et al. (2020). "Understanding the Importance of Contact Heterogeneity and Variable Infectiousness in the Dynamics of a Large Norovirus Outbreak". In: *Clinical Infectious Diseases* 70.3, pp. 493–500. ISSN: 1058-4838. DOI: 10.1093/cid/ciz220.