SISMID 2021: Mathematical Models of Infectious Disease
Day 2 Breakout Session on Heterogeneity and Herd Immunity

Kevin C. Ma
kevinchenma@g.harvard.edu

1 Introduction

As we've seen in lecture, the dynamics of infectious diseases spread are influenced by population heterogeneity. This is especially true for herd immunity, which occurs when susceptible individuals in a population are indirectly protected from infection due to immunity in others. The herd immunity threshold (HIT) is the fraction of the population that is non-susceptible when an unmitigated epidemic reaches its peak. In a population with homogeneous mixing, the HIT is 1-1/R_0, where R_0 is the basic reproduction number; for example, this translates to an HIT of 67% using an R_0 of 3.

However, population homogeneity is an unrealistic assumption, and models incorporating heterogeneity in social exposure and infection susceptibility (defined as the probability of infection given exposure) generally result in lowered HITs [1–5]. The key idea behind these models is that subpopulations important for epidemic spread (i.e., those with substantially increased susceptibility or exposure) become infected – and thus develop immunity – early on in an epidemic's course. Herd immunity for the population overall is then achieved earlier because once these individuals are no longer susceptible to infection, further epidemic spread is slowed.

We will explore this effect using seroprevalence data on SARS-CoV-2 spread in New York City (NYC) and Long Island [6], focusing in particular on disparities in infection rates across racial and ethnic groups, and using models assuming proportionate or assortative mixing (see Slides 33-38 in Dr. Lipsitch's Heterogeneity lecture). Substantial racial and ethnic disparities in infection rates, hospitalizations, and deaths have been characterized across the US [7–13], but it is unclear how these heterogeneities in risk are expected to change over time, and what implications – if any – they have on overall epidemic dynamics. Seroprevalence studies characterize past exposure by identifying SARS-CoV-2 antibodies and are more reliable and unbiased than case data, which suffer from underreporting and other biases [14]. Identifying and building structured SEIR models using observed differences in infection rates across social strata is important for understanding how variation in ex-

posure or susceptibility and social disparities are interconnected. These models are also useful for designing interventions that can both reduce disparities and disrupt overall transmission by focusing efforts on groups most affected by high transmission rates [15, 16]. Note that the types of models we will explore today, however, are meant to be illustrative and intentionally oversimplified, and the numbers generated are not intended as best estimates for HITs in these locations. Among other limitations, we're only considering variation in risk along one social demographic, and we're fitting to a relatively limited data set (one cross-sectional seroprevalence data point).

2 Code overview

Let's start by loading in (or installing) the necessary packages (chunk 1):

```
# Load libraries needed for this computing session
# For missing packages, run install.packages("package")

library(tidyverse)
library(deSolve)
library(RColorBrewer)
library(reshape2)
```

and then loading in the datasets and parameters (chunk 2). Take a second to read through the example comment to make sure you understand the format of the census and serosurvey data:

```
# Set parameters and input data
# Parameters used throughout all models
g.parm = 1/4        # Mean infectious period of 4 days
r.parm = 1/3        # Mean latent period of 3 days
default.R0 = 3.0    # All analyses will be run assuming an R0 of 3


# Seroprevalence data are from Rosenberg et al. 2020 Table 2 and population census data are
    from 2018 ACS 1-year estimates Table B03002
# Ordering of vectors: non-Hispanic white, Hispanic or Latino, non-Hispanic Black /
    African-American, NH Asian, multiracial / other
```

```r
# pop.loc = total population at given location
# pop.p.loc = fraction of each population belonging to each demographic group from census
    data; proportion for "other" group is given by 1 minus sum of all other groups
# sero.p.loc = adjusted cumulative incidence at given location for each demographic group
# sero.tested.loc = number of people tested at given location for each demographic group

pop.nyc <- 8398748
pop.p.nyc <- c(0.319, 0.292, 0.217, 0.141, 0.032)
sero.p.nyc <- c(0.166, 0.330, 0.252, 0.145, 0.204)
sero.tested.nyc <- c(1758, 2103, 1392, 509, 184)
# Example: There are 8398748 people in NYC, of which 31.9% are non-Hispanic white. In the
    serosurvey, Rosenberg et al. tested 1758 non-Hispanic whites from NYC, of which 16.6%
    were seropositive (have SARS-CoV-2 antibodies, indicating they were infected at some
    point in the past).

pop.li <- 2839436
pop.p.li <- c(0.632, 0.186, 0.093, 0.068, 0.022)
sero.p.li <- c(0.087, 0.320, 0.158, 0.084, 0.207)
sero.tested.li <- c(1599, 301, 111, 50, 50)
```

Before we run code chunk 3, let's review some of the notation we're using for the SEIR model. Compared to the single-population, homogeneous models we've already seen, it's often convenient to use matrix notation to express a more general form of the SEIR model allowing for interactions between racial and ethnic grous (or other demographic groups of interest). The groups interact through a social contact matrix that governs the cross-talk between and within groups. The system of equations looks similar to the homogeneous case:

$$\frac{d\mathbf{S}}{dt} = -(\mathbf{BI}) \circ \mathbf{S} \tag{1}$$

$$\frac{d\mathbf{E}}{dt} = (\mathbf{BI}) \circ \mathbf{S} - r\mathbf{E} \tag{2}$$

$$\frac{d\mathbf{I}}{dt} = r\mathbf{E} - \gamma\mathbf{I} \tag{3}$$

$$\frac{d\mathbf{R}}{dt} = \gamma\mathbf{I} \tag{4}$$

but now $\mathbf{S}$, $\mathbf{E}$, $\mathbf{I}$, $\mathbf{R}$ are column vectors comprising the compartmental variables for each group (i.e.,

$\mathbf{S} = [S_0, ..., S_p]^T$ for p demographic groups), $\circ$ denotes element-wise multiplication, and $\mathbf{B}$ is a matrix. We let S_0 denote non-Hispanic whites, S_1 denote Hispanics or Latinos, S_2 denote non-Hispanic African-Americans, S_3 denote non-Hispanic Asians, and S_4 denote multiracial or other demographic groups, with similar ordering for elements in vectors $\mathbf{E}$ through $\mathbf{R}$. Given a mean incubation period and mean serial interval of 5 days as suggested by empirical studies [17, 18], we set the mean latent period $1/r$ to be 3 days to allow for pre-symptomatic transmission and the mean infectious period $1/\gamma$ to be 4 days to coincide with the observed serial interval.

Now we can run code chunk 3, which sets up the SEIR model outlined above:

```
# SEIR model with variable exposure for each demographic group via social mixing matrix;
    also allows for varying degrees of assortativity
# epsilon = 0 gives proportionate mixing and epsilon = 1 gives fully assortative mixing

exposure.SEIR <- function(t, x, parms){      # Set up function with three arguments
    with(as.list(c(t, x, parms)),{           # "with" lets us directly use the variables in
        parms
        S = c(S0, S1, S2, S3, S4)            # S vector comprises susceptible counts for
            all groups...
        E = c(E0, E1, E2, E3, E4)            # ... and similarly for the other compartments
            E through R
        I = c(I0, I1, I2, I3, I4)
        R = c(R0, R1, R2, R3, R4)
        beta <- (1-epsilon)*outer(activities, activities)/sum(pop*pop.p*activities) +
          epsilon*activities/(pop*pop.p)*diag(5)    # Transmission rates are dependent on...
                                                    # the contact rates of the two groups
                                                    # epsilon allows for assortative mixing
        beta <- beta/scaling.factor          # Scale beta so the R0 matches what we
            specified (default.R0 in the parameter section)

        dS <- -beta %*% I*S                  # System of rate equations in matrix
            formulation
        dE <- beta %*% I*S - r*E             # Note the matrix multiplication (%*%) ...
        dI <- r*E - g*I                      # and elementwise-multiplication (*) ...
        dR <- g*I                            # which matches the equations above.
        der <- c(dS, dE, dI, dR)             # Combine results into a single vector der
        list(der)                            # Return result as a list
```

```
    })
}
```

We'll now present the mathematics behind the proportionate mixing versus assortative mixing models here – feel free to jump into the code first and if you have time later, come back and read through this more closely. Note that the notation is a little different compared to Dr. Lipsitch's lecture notes, but the equations are all equivalent. The simplest variable exposure model we analyze will be the *proportionate mixing* model, which assumes that the contact rate for each pair of groups is proportional to the activity level of the two groups (i.e., total number of contacts per unit time for an individual of group i) [19]. Denoting a_i as the activity level for a member of group i and $\mathbf{a}$ as the $1 \times p$ vector of a_is, the ijth entry in the transmission matrix is given by:

$$\beta_{i \leftarrow j} = q \frac{a_i a_j}{\sum_k a_k N_k} \tag{5}$$

and the overall transmission matrix $\mathbf{B}$ can be written as:

$$\mathbf{B} = \frac{q}{\sum_k a_k N_k} \mathbf{a}\mathbf{a}^T \tag{6}$$

That's what's being calculated in this line of code in chunk 3 ($q = 1$ for all groups so that term is ignored):

```
outer(activities, activities)/sum(pop*pop.p*activities)
```

The key unknown parameters are the activity levels. More complex variable exposure models allow for *assortative mixing*: we can partition a fraction ϵ of contacts to be **exclusively within-group** and distribute the rest of the contacts according to proportionate mixing (with δ_{ij} being an indicator variable that is 1 when $i = j$ and 0 otherwise) [19]:

$$\beta_{i \leftarrow j} = (1 - \epsilon) q \frac{a_i a_j}{\sum_k a_k N_k} + \epsilon \delta_{ij} q \frac{a_i}{N_i} \tag{7}$$

$$\mathbf{B} = \frac{(1 - \epsilon) q}{\sum_k a_k N_k} \mathbf{a}\mathbf{a}^T + \epsilon \, q \, diag(\mathbf{a} \circ 1/\mathbf{N}) \tag{8}$$

That's what's being calculated in this line of code in chunk 3:

```
(1-epsilon)*outer(activities, activities)/sum(pop*pop.p*activities) +
         epsilon*activities/(pop*pop.p)*diag(5)
```

The main takeaway is to generally understand how these models differ – assortative mixing allows individuals from group i to preferentially make contacts with other individuals from group i, which is a more accurate depiction of many real-world interactions.

Now run code chunk 4. We won't step through all the lines here but they are commented in the code if you want to read through later (see also the Mathematical appendix for more information). There are two main things this code chunk function does: 1) it re-scales the transmission rates so the R_0 is equal to the R_0 we specified back in code chunk 2, and it also plots the next-generation matrix for each model. **The next-generation matrix gives the average number of infections in group i (row) caused by an infected individual in group j (column).** After that, run code chunk 5. This is a helper function that solves the model, calculates the final epidemic size and HIT, and plots the epidemic trajectory.

Finally, run code chunk 6 to load in the functions we need to fit these models using the seroprevalence data. In particular, we use the data to fit the activity level parameters (total number of contacts per unit time for an individual of group i) using maximum likelihood estimation. We will talk through this and other types of model fitting methods more in depth on Day 3, so for today we'll only briefly introduce the ideas. To conduct maximum likelihood, we first assume some probability distribution underlying how the data were generated or collected. In our case, we assumed that the seroprevalence data were collected via a binomial sampling process: at a given time point t_s representing the time of the serosurvey, the number of observed seropositive cases Y_i from group i is distributed $Bin(m_i, R_i/N_i)$, where m_i is the number of people tested from group i in the serosurvey and R_i/N_i is the fraction of recovered people from the SEIR model (i.e. the "true" seropositivity proportion). With this probability distribution in mind, we can then let R optimize and find the parameters that result in the highest likelihood (intuitively, the congruence between the data and the probability model). The main idea here is we're solving for parameters by **connecting the seroprevalence data to the number of people in the R compartment in the SEIR model.**

Great! We've run through all the functions necessary to actually begin estimating parameters, simulating models, and interpreting the results. Run code chunk 7 – we'll test everything out by fitting a model to the Long Island data assuming proportionate mixing between demographic groups. The first line will fit the model using maximum likelihood and estimate and print out the activity levels.

```
### Proportionate mixing
# epsilon = 0 runs a proportionate mixing model

# Fit the activity levels using maximum likelihood
activities.pmix.li <- fit.model(pop=pop.li, pop.p=pop.p.li, sero.p=sero.p.li,
    sero.tested=sero.tested.li,
                                epsilon=0, r.parm=r.parm, g.parm=g.parm)
```

The output should look something like this:

```
[1] "Normalized fitted values: 1 4.31 1.956 0.916 2.475"
```

The ordering of results by racial or ethnic group is non-Hispanic whites, Hispanics or Latinos, non-Hispanic Blacks or African-Americans, non-Hispanic Asians, and multiracial or other. The interpretation is that **under this proportionate mixing model**, Hispanics or Latinos have 4.31 times the activity level relative to non-Hispanic whites, non-Hispanic Blacks or African-Americans have 1.96 times the activity level relative to non-Hispanic whites, and so on. The next line will now take these activity level parameters and simulate the full epidemic, calculating the final epidemic size, the HIT, and the cumulative incidence of each group at the point that the HIT is reached. It will also generate two plots: the next-generation matrix plot, and an epidemic trajectory plot. In the first exercise, you'll discuss these results with your group and make sure that you all see the same numbers.

```
# Simulate the model with R0=3 and plot outputs / print results (HIT, final epidemic size)
results.pmix.li <- run.structured.model(R0.value=default.R0, pop=pop.li, pop.p=pop.p.li,
    epsilon=0,
                                    g.parm=g.parm, fitted.vars=activities.pmix.li)
print(results.pmix.li)

# For the NGM plot, Group A denotes non-Hispanic whites, B denotes Hispanics or Latinos,
# ... C denotes non-Hispanic African-Americans, D denotes non-Hispanic Asians,
# ... and E denotes multiracial or other demographic groups.
```

3 Exercises

A) Run all of code chunks 1-7: this will load all of the data and functions, conduct model fitting, and simulate the full (unmitigated) epidemic for Long Island assuming proportionate mixing. What is the overall estimated HIT and final epidemic size for the population? At the point that the HIT is reached, what is the cumulative incidence in each demographic group? What do you observe in the two figures (next-generation matrix and epidemic trajectory plots – N.B. instead of plotting all compartments, we're just plotting the R compartment for all demographic groups)? Discuss in groups your findings (make sure your answers match) and also whether or not you think the activity level estimates are reasonable. What assumptions underlie this model, and are they reasonable / unreasonable?

In this simplified model of an unmitigated epidemic in Long Island, the herd immunity threshold (HIT) is reached when 40% of the population is infected and the epidemic ends when 69% of the population is infected. When the HIT is reached, 29% of non-Hispanic whites, 77% of Hispanics or Latinos, 48% of non-Hispanic African-Americans, 27% of non-Hispanic Asians, and 57% of multiracial or other people are infected.

```
[1] "Normalized fitted values: 1 4.31 1.956 0.916 2.475"
$`Final epidemic size`
[1] 0.6932
$HIT
[1] 0.3977
$`Cumulative incidence by group at the HIT`
[1] 0.2861 0.7660 0.4828 0.2657 0.5658
```

The next-generation matrix gives the average number of infections in group i (row) caused by an infected individual in group j (column) (see Figure 1 below). Take column A for example (demographic groups are in the same order as before) – in the exponential growth phase of the epidemic, on average, 1 infected white person infects 0.41 white people, 0.52 Hispanics or Latinos, 0.12 Black people, 0.04 Asian people, and 0.04 multiracial or other individuals. This reflects a combination of the interactions between groups (assumed to be proportionate mixing), the activity levels, and the population size of each group.

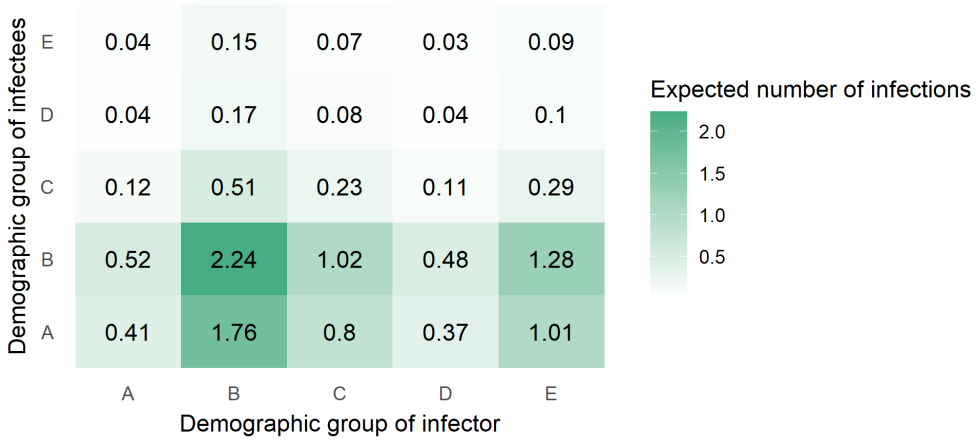

Figure 1

The epidemic trajectory plot (see Figure 2 below) shows the accumulation of recovered individuals (i.e., cumulative incidence) in each demographic group over time. Cumulative incidence is projected to be highest in several demographic groups (Hispanics or Latinos, non-Hispanic Black people, multiracial or other people) across the epidemic, which is in line with the cumulative incidence at the HIT numbers from above, and mirrors existing inequities in housing, education, healthcare, and beyond.

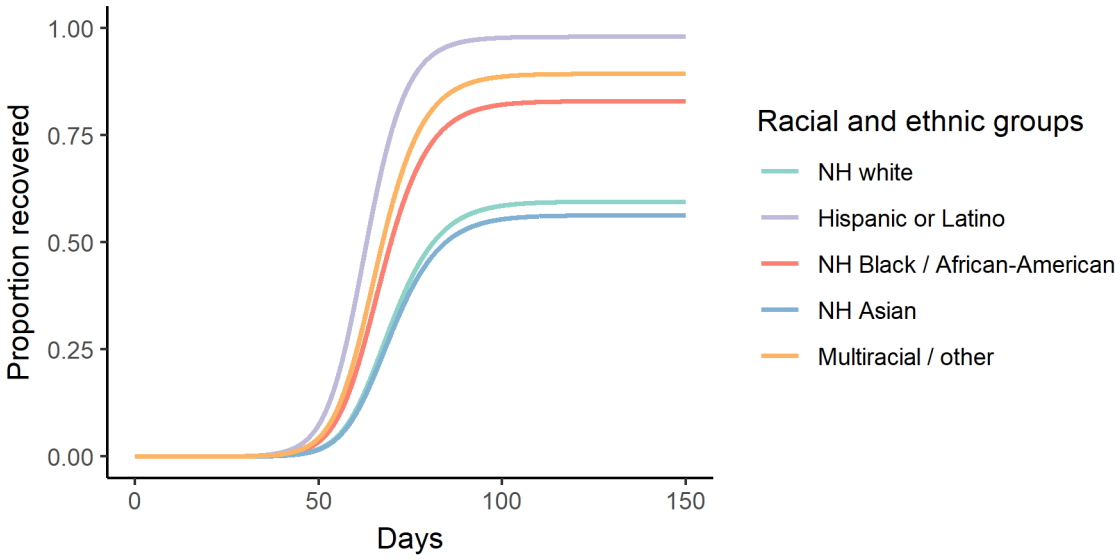

Figure 2

There are many assumptions underlying this model and its outputs, and returning to Dr. Martinez's model characterization in Lecture 1, we might consider this model to be fairly transparent and flexible, but with reduced accuracy and resolution compared to more complex models. Some assump-

tions include seropositivity implies complete immunity and that immunity does not wane. Additionally, variability in exposure is modeled as a function of only one demographic characteristic; models should ideally strive to also account for the effects of age (and other demographics) on susceptibility and exposure within strata of race and ethnicity.

B) Now go to code chunk 8, set ϵ to 0.5 in the arguments for both functions, and run the code chunk. This will conduct model fitting and simulation for Long Island assuming a high degree of assortative mixing ($\epsilon = 0.5$ i.e. half of contacts are exclusively within group, and the other half are proportionately mixed). Compare the activity levels, next-generation matrix plots, the epidemic trajectory plots, and the overall HIT values versus those in question 3A. What do you observe? Why do you think there is a difference?

Assuming a high degree of assortative mixing, in this simplified model of an unmitigated epidemic in Long Island, the herd immunity threshold (HIT) is reached when 50% of the population is infected and the epidemic ends when 83% of the population is infected. When the HIT is reached, 39% of non-Hispanic whites, 83% of Hispanics or Latinos, 61% of non-Hispanic African-Americans, 36% of non-Hispanic Asians, and 69% of multiracial or other people are infected. **These values are generally shifted higher compared to the proportionate mixing model.**

```
[1] "Normalized fitted values: 1 2.196 1.516 0.941 1.714"
$`Final epidemic size`
[1] 0.8268
$HIT
[1] 0.4955
$`Cumulative incidence by group at the HIT`
[1] 0.3875 0.8342 0.6099 0.3615 0.6871
```

Compared to the next-generation matrix (Figure 3) for proportionate mixing (top), the NGM for assortative mixing (bottom) shows much greater weight on the diagonals – this matches our intuition that assortative mixing models should increase the number of within-group contacts.

Figure 3

Now comparing the epidemic trajectories (Figure 4) for proportionate (left) and assortative (right) mixing, the final epidemic sizes for the highest-risk demographic groups only shifts slightly – the main change is in the final epidemic sizes for the lower-risk demographic groups. In other words, **under proportionate mixing, lower-risk demographic groups are protected from further infection due to built-up immunity in higher-risk demographic groups, but the magnitude of this indirect protection decreases as the proportion of exclusively within-group contacts increases and groups become more isolated.**

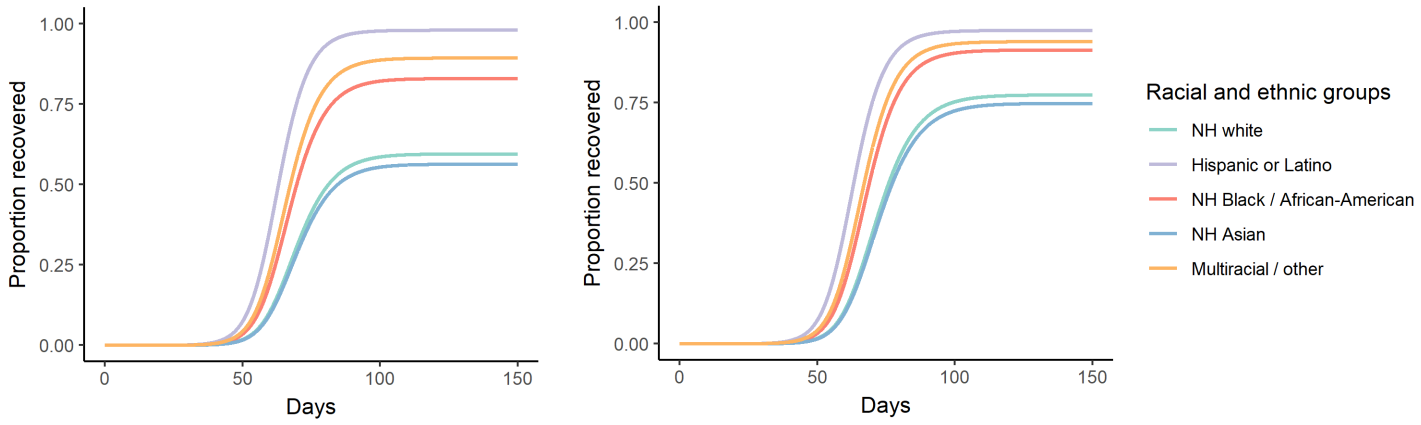

Figure 4

C) Now, create a new code chunk where you conduct model fitting and simulation for NYC. Do this both for a proportionate mixing model and an assortative mixing model ($\epsilon = 0.5$). How do your results compare to those of Long Island? Why do you think there is a difference (compare the seroprevalence data vectors for NYC and Long Island – e.g. calculate coefficient of variation)?

In this simplified model of an unmitigated epidemic in NYC, the results assuming proportionate mixing (top) and assortative mixing (bottom) are shown below. Compared to the results for Long Island, values are generally shifted higher, and the difference between proportionate and assortative mixing models is not as pronounced. The epidemic trajectories for an assortative mixing model (Figure 5) and proportionate mixing model (not shown) are also more similar across racial and ethnic groups. **The difference in seroprevalence across groups is more pronounced in Long Island compared to NYC (e.g. coefficient of variation for Long Island is 0.57 versus coefficient of variation for NYC is 0.33), and this increased heterogeneity could be responsible for the differences in model outputs we observe.**

```
### Proportionate mixing
```

```
[1] "Normalized fitted values: 1 2.249 1.615 0.863 1.279"
$`Final epidemic size`
[1] 0.8727

$HIT
[1] 0.5759
```

```
$`Cumulative incidence by group at the HIT`
[1] 0.4588 0.7487 0.6291 0.4114 0.5441

### Assortative mixing

[1] "Normalized fitted values: 1 1.561 1.322 0.909 1.161"
$`Final epidemic size`
[1] 0.9085

$HIT
[1] 0.6105

$`Cumulative incidence by group at the HIT`
[1] 0.5033 0.7661 0.6653 0.4545 0.5871
```

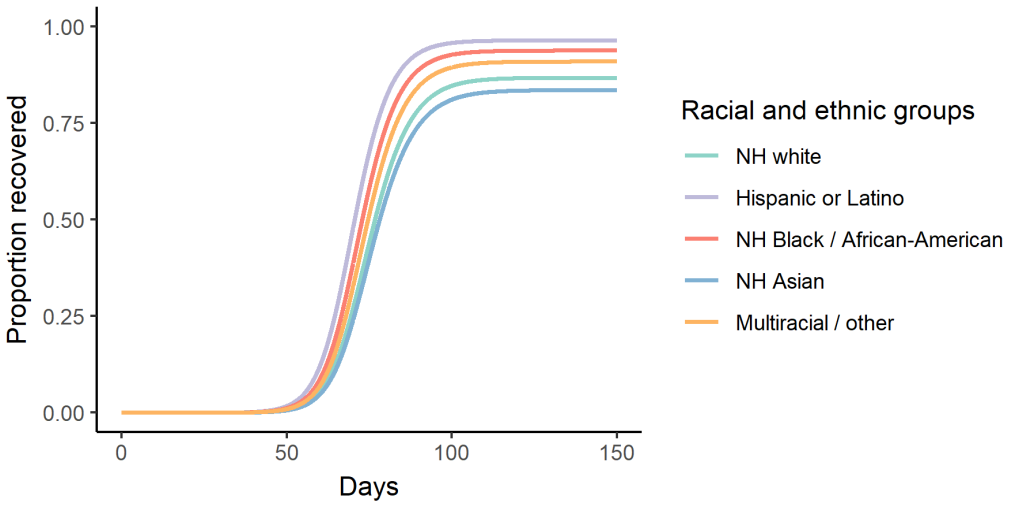

Figure 5

D) Test your ideas in 3C by creating your own vector of seroprevalence values. How does the HIT change when there is greater variation in seroprevalence across groups versus when there is little or even no variation? Compared to a homogeneous population, what do you think is the effect of modeling heterogeneity on HITs and final epidemic sizes?

To model a homogeneous population, we can simulate a data vector where seropositivity is equal in

all of the groups and run the usual functions. We see that now, the HIT is 0.67, exactly matching the canonical formula of 1-1/R0 from lecture. This value is higher than any of the variability in exposure models we've looked at in the previous questions.

```
activities.homogeneous <- fit.model(pop=pop.li, pop.p=pop.p.li, sero.p=rep(0.2,5),
    sero.tested=sero.tested.li,
                              epsilon=0, r.parm=r.parm, g.parm=g.parm)
results.homogeneous <- run.structured.model(R0.value=default.R0, pop=pop.li,
    pop.p=pop.p.li, epsilon=0,
                                    g.parm=g.parm, fitted.vars=activities.homogeneous)
print(results.homogeneous)

### Output
[1] "Normalized fitted values: 1 0.996 0.989 0.999 1"
$`Final epidemic size`
[1] 0.9414

$HIT
[1] 0.6666

$`Cumulative incidence by group at the HIT`
[1] 0.6677 0.6661 0.6637 0.6672 0.6675
```

We can also simulate a data vector where seropositivity is highly skewed. Consider a situation where seroprevalence is similar in 4 of the 5 groups but extremely high in the remaining group. The final epidemic size and HIT is now dramatically reduced – because infections in this high-risk group are the primary driver of the overall epidemic, once immunity has sufficiently built up, the overall population is robustly protected.

```
activities.heterogeneous <- fit.model(pop=pop.li, pop.p=pop.p.li,
    sero.p=c(0.05,0.8,0.05,0.05,0.05), sero.tested=sero.tested.li,
                              epsilon=0, r.parm=r.parm, g.parm=g.parm)
results.heterogeneous <- run.structured.model(R0.value=default.R0, pop=pop.li,
    pop.p=pop.p.li, epsilon=0,
                                    g.parm=g.parm, fitted.vars=activities.heterogeneous)
print(results.heterogeneous)
```

```
### Output
$`Final epidemic size`
[1] 0.2431

$HIT
[1] 0.1506
```

This suggests that both variability in the data and the choice of model affects HIT and final epidemic size estimates. All else equal (holding R0 constant and using the same types of models), increased variability in the seroprevalence data will generally reduce HITs and final epidemic sizes. Additionally, choosing to model dynamics using proportionate versus assortative mixing also matters – proportionate mixing models generally have reduced HITs and final sizes versus assortative mixing models. The best models to use will vary by situation, but generally assortative mixing models are more complex and realistic, and contain proportionate mixing as a subset (i.e. with ϵ set to 0).

E) Think through and discuss with your group how you would modify the code to incorporate the effect of non-pharmaceutical interventions (e.g. which parameters would you try to vary, for how long, by what magnitude, etc.). If you're stuck, see our Jupyter notebook (https://nbviewer.jupyter.org/github/kevincma/covid19-race-ethnicity-model/blob/main/covid19-race-ethnicity-models-notebook.ipynb) or paper (https://elifesciences.org/articles/66601) for ideas on how to implement these additions to the model.

This is an open-ended question with no right answer – in our paper, to account for the effects of social distancing and other non-pharmaceutical interventions (NPIs), we scaled the transmission rate by a factor α beginning when 5% cumulative incidence in the population was reached – representing an established and expanding epidemic – for a variable duration. We varied the parameters involved (duration of distancing and α) to see what effects including distancing had on model inferences. There have been many other approaches to modeling NPIs. Simple extensions to our method would allow for linear increases and decreases in NPI magnitude, since what we modeled was effectively just a stepwise function. More complex approaches could use mobility data as an outside data source to parameterize the strength of NPIs.

4 Appendix

We define contacts to be interactions between individuals that allow for transmission of SARS-CoV-2 with some non-zero probability. Following the convention for age-structured transmission models [20], we defined the $p \times p$ per capita social contact matrix $\mathbf{C}$ to consist of elements $c_{i \leftarrow j}$ at row i and column j, representing the per capita rate that individuals from group i are contacted by individuals of group j. Letting N_i be the total number of individuals in group i, the social contact matrix $\mathbf{M}$ consists of elements $m_{i \leftarrow j} = c_{i \leftarrow j} * N_i$, which represents the average number of individuals in group i encountered by an individual in group j. In the most general case, the susceptibility to infection can vary between groups, which can be modeled by allowing the probability of infection given contact with an infected individual to vary: $\mathbf{q} = [q_0, ..., q_p]^T$. Whether to build models that allow for variable susceptibility versus exposure will depend on the demographic groups being modeled. For instance, for some diseases, age-based variation in infection rates may be due to intrinsic biological differences in susceptibility by age to infection. In contrast, because observed disparities in infection rates in US cities are strongly attributable to differences in mobility and exposure, we focused primarily on building and analyzing the variable exposure models [21–23].

The transmission matrix $\mathbf{B}$ is then given by $(\mathbf{q}1^T) \circ \mathbf{C}$, where 1^T is a 1 by p vector of 1s:

$$\mathbf{B} = \begin{bmatrix} q_0 & \dots & q_0 & \dots & q_0 \\ \vdots & & \vdots & & \vdots \\ q_4 & \dots & q_4 & \dots & q_4 \end{bmatrix} \circ \begin{bmatrix} c_{0\leftarrow 0} & \dots & c_{0\leftarrow 2} & \dots & c_{0\leftarrow 4} \\ \vdots & & \vdots & & \vdots \\ c_{4\leftarrow 0} & \dots & c_{4\leftarrow 2} & \dots & c_{4\leftarrow 4} \end{bmatrix} \tag{9}$$

Given mean duration of infectiousness $1/\gamma$, the next-generation matrix $\mathbf{G}$, representing the average number of infections in group i caused by an infected individual in group j, is given by $(q1^T) \circ \mathbf{M}/\gamma = \mathbf{N} \circ \mathbf{B}/\gamma$. R_0 for the overall population described by this structured model was calculated by computing the dominant eigenvalue of matrix $\mathbf{G}$. The effective reproduction number R_t at time t was calculated by computing the dominant eigenvalue of $\mathbf{G_t} = (\mathbf{q}1^T) \circ \mathbf{M_t}/\gamma = \mathbf{S_t} \circ \mathbf{B}/\gamma$, where the elements in $\mathbf{M_t}$ are given by $c_{i \leftarrow j} * S_{i,t}$ and $S_{i,t}$ is the number of susceptible individuals in group i at time t. To hold R_0 values across model types constant when calculating HITs and final epidemic sizes, we re-scaled transmission matrices to have the same dominant eigenvalue.

Calculating the HIT and final epidemic size

We defined the herd immunity threshold (HIT) for all models as the fraction of non-susceptible peo-

ple when the effective reproduction number R_t first crosses 1. In the homogenous model, where $R_t = S_t\beta/\gamma$, the analytical solution for the HIT occurs when the fraction of non-susceptible individuals equals 1-1/R_0. In the structured models of heterogeneous populations, the HIT was calculated via simulation: we took the dominant eigenvalue of $\mathbf{G_t}$ at each timestep to calculate R_t, and identified the number of non-susceptible individuals when R_t first decreased below 1. Final epidemic sizes were calculated by simulation by running the epidemic out to 1 year.

References

[1] Hill, A. N. & Longini, I. M., Jr. The critical vaccination fraction for heterogeneous epidemic models. *Math. Biosci.* **181**, 85–106 (2003).

[2] Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849 (2020).

[3] Gomes, M. G. M. *et al.* Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. *medRxiv* (2020).

[4] Aguas, R. *et al.* Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. *medRxiv* 2020.07.23.20160762 (2020).

[5] Tkachenko, A. V. *et al.* Persistent heterogeneity not short-term overdispersion determines herd immunity to COVID-19 (2020). 2008.08142.

[6] Rosenberg, E. S. *et al.* Cumulative incidence and diagnosis of SARS-CoV-2 infection in new york. *Ann. Epidemiol.* **48**, 23–29.e4 (2020).

[7] Chamie, G. *et al.* SARS-CoV-2 community transmission disproportionately affects latinx population during Shelter-in-Place in san francisco. *Clin. Infect. Dis.* (2020).

[8] Moore, J. T. *et al.* Disparities in incidence of COVID-19 among underrepresented Racial/Ethnic groups in counties identified as hotspots during june 5-18, 2020 - 22 states, February-June 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1122–1126 (2020).

[9] Millett, G. A. *et al.* Assessing differential impacts of COVID-19 on black communities. *Ann. Epidemiol.* **47**, 37–44 (2020).

[10] Pan, D. *et al.* The impact of ethnicity on clinical outcomes in COVID-19: A systematic review. *EClinicalMedicine* **23**, 100404 (2020).

[11] Chen, J. T. & Krieger, N. Revealing the unequal burden of COVID-19 by income, race/ethnicity, and household crowding: US county vs ZIP code analyses. *Harvard Center for Population and Development Studies Working Paper Series* **19** (2020).

[12] Bassett, M. T., Chen, J. T. & Krieger, N. Variation in racial/ethnic disparities in COVID-19 mortality by age in the united states: A cross-sectional study. *PLoS Med.* **17**, e1003402 (2020).

[13] Hanage, W. P. *et al.* COVID-19: US federal accountability for entry, spread, and inequities-lessons for the future. *Eur. J. Epidemiol.* (2020).

[14] Metcalf, C. J. E., Viboud, C., Spiro, D. J. & Grenfell, B. T. Using serology with models to clarify the trajectory of the SARS-CoV-2 emerging outbreak. *Trends Immunol.* **41**, 849–851 (2020).

[15] Wallinga, J., van Boven, M. & Lipsitch, M. Optimizing infectious disease interventions during an emerging epidemic. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 923–928 (2010).

[16] Bubar, K. M. *et al.* Model-informed COVID-19 vaccine prioritization strategies by age and serostatus (2020).

[17] Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *Int. J. Infect. Dis.* **93**, 284–286 (2020).

[18] Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172**, 577–582 (2020).

[19] Hethcote, H. W. Modeling heterogeneous mixing in infectious disease dynamics. In *Models for Infectious Human Diseases: Their Structure and Relation to Data*, 215–238 (Cambridge University Press, 1996).

[20] Wallinga, J., Kassteele, J. v. d. & Hens, N. Contact patterns for contagious diseases. In Held, L., Hens, N., O'Neill, P. & Wallinga, J. (eds.) *Handbook of Infectious Disease Data Analysis*, 93–110 (Chapman and Hall/CRC, 2019), 1 edn.

[21] Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* (2020).

[22] Zelner, J. *et al.* Racial disparities in COVID-19 mortality are driven by unequal infection risks (2020).

[23] Kissler, S. M. *et al.* Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in new york city. *Nat. Commun.* **11**, 4674 (2020).