

MIT Open Access Articles

Implications of heterogeneous SIR models for analyses of COVID-19

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Ellison, G. Implications of heterogeneous SIR models for analyses of COVID-19. *Rev Econ Design* 28, 651–687 (2024).

As Published: <https://doi.org/10.1007/s10058-024-00355-z>

Publisher: Springer Berlin Heidelberg

Persistent URL: <https://hdl.handle.net/1721.1/159167>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-ShareAlike



Implications of heterogeneous SIR models for analyses of COVID-19

This Accepted Manuscript (AM) is a PDF file of the manuscript accepted for publication after peer review, when applicable, but does not reflect post-acceptance improvements, or any corrections. Use of this AM is subject to the publisher's embargo period and AM terms of use. Under no circumstances may this AM be shared or distributed under a Creative Commons or other form of open access license, nor may it be reformatted or enhanced, whether by the Author or third parties. By using this AM (for example, by accessing or downloading) you agree to abide by Springer Nature's terms of use for AM versions of subscription articles: <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>

The Version of Record (VOR) of this article, as published and maintained by the publisher, is available online at: <https://doi.org/10.1007/s10058-024-00355-z>. The VOR is the version of the article after copy-editing and typesetting, and connected to open research data, open protocols, and open code where available. Any supplementary information can be found on the journal website, connected to the VOR.

For research integrity purposes it is best practice to cite the published Version of Record (VOR), where available (for example, see ICMJE's guidelines on overlapping publications). Where users do not have access to the VOR, any citation must clearly indicate that the reference is to an Accepted Manuscript (AM) version.

Implications of Heterogeneous SIR Models for Analyses of COVID-19*

Glenn Ellison[†]

April 2024

Abstract

This paper provides a quick survey of results on the classic SIR model and variants allowing for heterogeneity in contact rates. It notes that calibrating the classic model to data generated by a heterogeneous model can lead to forecasts that are biased in several ways and to understatement of the forecast uncertainty. Among the biases are that we may underestimate how quickly herd immunity might be reached, underestimate differences across regions, and have biased estimates of the impact of endogenous and policy-driven social distancing.

1 Introduction

The economic literature on the COVID-19 epidemic developed at a remarkable pace. Many economics papers have built on the classic Susceptible-Infectious-Recovered (SIR) model to study how the epidemic may progress and how it may be affected by various policies.¹ In this paper I review some results from the epidemiological literature on an SIR extension that economists have mostly not yet adopted, incorporating heterogeneity in the activity rates of different subpopulations, and note ways in which analyses building on classic SIR models can potentially yield misleading views.

The classic SIR model of Kermack and McKendrick (1927) has been a foundational model in epidemiology for nearly a century. It illustrates basic tradeoffs, including that epidemics spread or die out depending on whether a key parameter R_0 is greater than or less than one, and provides a simple framework that can be easily built on. Subsequent work in epidemiological theory has extended the model in various ways, and modern epidemiological forecasts typically work with models that are more flexible in a number of dimensions.² In this paper I focus on some theoretical extensions developed in the 1980's and 1990's that seem quite relevant to the

*I thank Daron Acemoglu, Chris Avery, Victor Chernozhukov, Adam Clark, Jonathan Dushoff, Sara Fisher Ellison, Jim Stock, Ivan Werning, David McAdams and an anonymous for helpful conversations and comments and Chris Ackerman and Bryan Kim for research assistance.

[†]Department of Economics, Massachusetts Institute of Technology, Cambridge MA 02139 and NBER, e-mail: gellison@mit.edu

¹See among others Acemoglu et al. (2021), Alvarez, Argente, and Lippi (2021), Baqaee et al. (2020), Eichenbaum, Rebelo, and Trabandt (2021), Farboodi, Jarosch, and Shimer (2021), Fernández-Villaverde and Jones (2022), Jones, Philippon, and Venkateswaran (2021), and Rowthorn and Toxvaerd (2020).

²See, for example, Champredon et al. (2018), Viboud et al. (2018), and Unwin et al. (2020).

COVID-19 epidemic. Specifically, I discuss two classic models that focus on heterogeneity in the frequency with which different individuals engage in interactions that risk spreading the disease. Heterogeneity in interactions arose naturally from differences in how individuals lived prior to the pandemic and from differences in how they reacted to it. For example, those living with many others, frequenting bars and nightclubs, or working in healthcare facilities would be expected to have many more risky interactions than those who were retired or worked from home, rarely went out, and wore high-quality masks in public places. While the first waves of the pandemic are long past, the model remains relevant for thinking about how new variant-driven waves will spread and how mitigation measures might affect their course.

Section 2 reviews the classic SIR model and extensions. Each extension discussed is a multipopulation SIR model that supposes that the subpopulations differ in their “activity” levels. As with the classic SIR model, the differential equations describing the rates at which members of each subpopulation transition from the susceptible to the infectious state can be motivated by a process in which agents are randomly matched in continuous time with each interaction between susceptible and infectious agents potentially leading to a new infection. One version assumes “uniform” matching in which the probability that any two agents are randomly matched is proportional to the product of their activity levels. The other assumes “homophilic” matching in which agents are more likely to interact with others in their own subpopulation. Each model behaves much like the classic SIR model. Small infections initially grow at an exponential rate if a composite parameter analogous to R_0 is greater than one and new infections slow (and eventually die out) after the fraction with acquired immunity passes a “herd immunity” threshold. The composite R_0 and the herd immunity thresholds depend on the characteristics of the various subpopulations, and I review results from prior work to illustrate important principles about how epidemics spread in heterogeneous populations.

Sections 3 and 4 then draw out implications of these models for analyses of the initial COVID-19 epidemic, subsequent variant waves, and other epidemics. Section 3 emphasizes that thinking about heterogeneity in contact patterns suggests that making predictions about the course of any epidemic wave and the impacts of relaxing restrictive policies is inherently difficult. Heterogeneous models have more parameters that need to be calibrated. Long run outcomes can be sensitive to activity levels of the less active, and it is difficult to calibrate these parameters early in an epidemic when there are few cases in less-active communities. This is particularly true when one contemplates removing restrictions and thereby increasing activity among the currently inactive. Predictions based on classic SIR models that do not allow for heterogeneity may be overconfident.

Section 4 then focuses on ways in which conclusions drawn from applying homogeneous SIR models to a world that may be like a heterogeneous SIR model can be misleading. One observation is that homogeneous SIR models may substantially overstate the fraction of the population that must be infected in order to achieve

herd immunity. Intuitively, if a small high-contact group plays a central role in spreading the disease, then incidence will be much higher in this group, and once many in this group have acquired immunity the epidemic may die out. Less obvious but still relevant effects are present with less extreme heterogeneity. A second related observation is that (targeted) lockdown policies can be more cost effective in heterogeneous populations. There can be substantial gains either from taking permanent measures to reduce spread among the highly active or from temporarily locking down less active groups to minimize overshooting of herd immunity thresholds. The differences in dynamics also imply that time-series estimates of policy impacts may be biased. In each case, effects depend both on the magnitude of the heterogeneity that it present and on the degree of homophily in matching. The discussions attempt to bring out comparative statics and plausible magnitudes of effects.

The analytic results in Section 2 concern how the addition of heterogeneity in contact rates affect the dynamics of the classic SIR model. The discussions in Sections 3 and 4 focus on practical implications of these results. Four years of experience has by now taught us that factors absent from these models are critically important for the COVID pandemic. The pandemic did not spread as fast as models predicted in part because people chose to reduce activity levels. The end that SIR models predicted did not arrive because COVID spawned many new variants that at least partially evaded immunity. And scientific advances made the rapid development of vaccines feasible. While it would be interesting to analyze heterogeneity in the context of more complex models with some of these features, I have left the analysis focused on the simplest models for multiple reasons. The primary reason is that the models are tractable and bring insights that should be more generally applicable. For example, understanding when a population has herd immunity in a standard heterogeneous SIR model is important both for thinking about whether it is possible for behavior to go back to normal in a behavioral SIR model, and about vaccine prioritization and the take up that will be necessary for a vaccine program to halt an epidemic. The results are also directly relevant to thinking about the early stages of a new epidemic (or variant) in which there is uncertainty about model parameters and people have not yet modified their behavior, and for thinking about distancing fatigue.

The final section of the paper discusses some practical implications of the results. The message that it is difficult to assess the severity of any new variant-driven wave until the variant is quite widespread is troublesome. But the models suggest fairly easy ways in which economists could extend their models and also point to data opportunities that might reduce the critical uncertainties. The messages that controlling an epidemic may not be nearly as hard as it appears in some models and that herd immunity might be reached much sooner than one would naïvely predict also give room for optimism.

This paper is related to a number of others in epidemiology and economics. The discussion of heterogeneous SIR models is a review of a literature in epidemiology that dates back to the late 1980s and mid 1990s, with

the particular formulations drawing heavily on Dushoff and Levin (1995). Empirical epidemiologists have also for quite some time been interested in multipopulation SIR models to examine interactions between age-based (important for other reasons for childhood diseases) and other groups, e.g. health care workers in the Ebola epidemic, who play an important role in transmission.³ Two post-COVID epidemiology papers made observations similar to the observation in section 4.1 (also reported in Ellison (2020)) that herd immunity thresholds can be substantially lower in heterogeneous SIR models than in homogeneous SIR models.⁴ Gomes et al. (2020) graphs the herd immunity threshold as a function of the coefficient of variation in contact rates in a heterogeneous SEIR model, noting estimates of the coefficients of variation that have been previously reported for other diseases. Britton, Ball, and Trapman (2020) give herd immunity thresholds for an 18-group model calibrated to estimated interactions across 6 age groups with assumed low-activity and high-activity individuals assumed to have activity levels that are half and twice the average activity levels and discuss partial lockdown policies that hold the infection to this level. There are also post-COVID epidemiology papers motivating examining heterogeneous transmission. Miller et al. (2020) examine transmission in Israel using full genome sequence and conclude that there are “high levels of transmission heterogeneity . . . with between 1-10% of infected individuals resulting in 80% of secondary infections. Worobey et al. (2020) concludes that early imported cases formerly thought to have triggered epidemics in Washington and Italy appear to not be related to the subsequent epidemics there, suggesting that the communities in which they occurred had low enough R_0 so that the epidemics they started died out.

The primary motivation for the paper is the large recent literature in economics mentioned in the first paragraph that provides policy advice using models building on SIR models. Most closely related within this group are several papers, including Acemoglu et al. (2021), Baqaee et al. (2020), Favero, Ichino, and Rustichini (2020), Rampini (2020), and Javadi, Quercioli, and Smith (2024a), that use calibrated multipopulation SIR models to examine the impact of COVID-19 mitigation policies, and in the case of Acemoglu et al. (2021) to identify optimal policies from a broad class. These papers use age-defined group structures to illustrate the substantial gains from age-targeted policies due to how dramatically death rates vary with age. They do not focus on the impact of contact heterogeneity, nor do most of the calibrations include within-age-group heterogeneity, which is presumably much larger than cross-age group heterogeneity, but three of them do include some heterogeneity in contact rates. Baqaee et al. (2020) calibrate a five by five matrix of age group to age group contact rates using both general contact survey data and a workplace proximity survey to reflect differences in occupational mixes across age groups. Acemoglu et al. (2021) use uniform mixing in their main analyses,

³See, for example, Britton (1998), Demiris and O’Neill (2005), Lloyd-Smith et al. (2005), and Champredon et al. (2018).

⁴Avery, Bossert, et al. (2020a) and Avery, Bossert, et al. (2020b) informally discuss the potential relevance of transmission heterogeneity.

but also calibrate a three by three age group contact matrix to data from another contact survey. The groups in Favero, Ichino, and Rustichini (2020) are age \times activity based with medium- and high-activity individuals assumed to be 12% and 18% more active than the low-activity group.

As mentioned earlier, the literature has moved well beyond the simple SIR-based models in a number of dimensions, and in a state-of-the-art analysis one would want to include elements from several. Goodkin-Gold et al. (2022) augment a homogeneous SIR model to develop an analysis of endogenous vaccine take up and its effects, while Avery, Chen, and McAdams (2024) combines endogenous vaccination and endogenous distancing in a SIRS model with reinfection. Vaccines clearly change how one would frame any policy discussion, although as Glenester et al. (2024) note, even with the rapid pace at which new mRNA vaccines can be developed they are generally being brought to market too late to avoid large costs from infection and/or suppression. The availability of diagnostic tests also clearly impacts the information available to agents and their resulting incentives, as studied by Javadi, Quercioli, and Smith (2024b), Droste, Atkeson, et al. (2024), and Troger (2024). The slowdown of economic (and other) activity even before initial COVID wave restrictions were put in place provided clear evidence that endogenous distancing was important. McAdams (2021) provides an insightful review of the burgeoning literature which includes the theoretical analyses of McAdams (2020) and Toxvaerd (2020) showing that dynamics can be very different under distancing, policy-oriented papers like Eichenbaum, Rebelo, and Trabandt (2021) and Farboodi, Jarosch, and Shimer (2021) noting that this can make optimal mitigation policies very different from what they would be in a pure SIR model, and empirical analyses including Atkeson, Kopecky, and Zha (2021) and Droste and Stock (2021) estimating behavioral parameters and noting that they appear to differ across time and place.

2 Heterogeneous SIR Models

In this section I'll quickly review the standard SIR model and then spend more time on two heterogeneous versions drawing on previous results.

2.1 The standard homogeneous SIR model

A number of recent economic analyses of the COVID-19 epidemic build on a standard homogeneous SIR model. Consider a continuum population of unit mass. Assume that at each time t each member of the population is in one of three states: Susceptible, Infectious, or Recovered. Write $S(t)$, $I(t)$, and $R(t)$ for the fractions in each

state at time t . Assume that the dynamics of these fractions are:

$$\begin{aligned}\dot{I}(t) &= S(t)I(t)R_0\gamma - \gamma I(t) \\ \dot{R}(t) &= \gamma I(t) \\ \dot{S}(t) &= -S(t)I(t)R_0\gamma\end{aligned}$$

One way to motivate the model is to suppose that agents are being uniformly randomly matched in continuous time. Each agent meets another with probability $R_0\gamma dt$ in a dt time interval. A susceptible agent matched with an infectious agent becomes infectious. Agents transition from the Infectious state to the Recovered state at Poisson rate γ . These transitions reflect both true recoveries and deaths from the disease. The theoretical results derived below do not depend on the share of infections that end in death (under the implicit assumption that interactions that would have occurred with someone who died are not replaced by another interaction), but the death rate is obviously incredibly important to welfare analyses and also plays an important role in empirical work because deaths are more easily observable than infections.

Note that this is a very primitive model with some assumptions that would be unrealistic to apply when a deadly disease is known to be rapidly spreading. In particular, the model assumes that people never change their behavior in a way that reduces transmission, regardless of how many people are currently infectious. But understanding the model's dynamics nonetheless provides important insights even for such situations. And the model is a good one to apply literally in various situations, e.g. very early in an epidemic when people do not yet know a virus is spreading or how to slow its spread, when the effects are not so severe as to trigger behavioral changes, or if most in a population have sufficient lockdown fatigue to have given up trying to alter their behavior.

The parameter R_0 can be thought of as the expected number of people that a newly infected person will directly infect when everyone is susceptible. It turns out to be the most important parameter to understand in order to think about the behavior of the model. When considering the spread of a new disease variant to which some are not susceptible due to vaccination or prior exposure to an earlier related strain, the proportion $S(0)$ who are susceptible when the virus arrives is also a critical determinant of what will happen. The important fact relevant to this situation is:

- If $R_0 > 1$, then the equilibrium $(S, I, R) = (1, 0, 0)$ is locally unstable. Adding a small number of infected agents leads to initially exponential growth of I . Equilibria with $I = 0$ are locally stable if $R_0 < 1$. A small infection dies out.⁵

⁵See Budish (2024) for insights on the importance of this dichotomy.

The intuition for the above result is simple. We can write $\dot{I}(t)$ as $\dot{I}(t) = \gamma(S(t)R_0 - 1)I(t)$. When $S(t) \approx 1$ this is approximately $\dot{I}(t) = \gamma(R_0 - 1)I(t)$, which has solution $I(t) = e^{\gamma(R_0-1)t}I(0)$. Hence, a small initial infection will initially grow at an exponential rate if $R_0 - 1$ is positive and shrink at an exponential rate if it is negative.

“Herd immunity” is an important often-discussed concept in SIR-style models. What people mean by the term is not always clear, so I’ll give it a precise meaning here. Define the “herd immunity region” H in the SIR model to be a subset of the possible values of the initial susceptibility for which the disease-free state is a locally-stable steady state, i.e.

$$H = \{S \in [0, 1] \mid (S, 0, 1 - S) \text{ is a locally stable steady state of the model.}\}$$

Whether a population has herd-immunity protection against a new virus or variant, i.e. whether $S(0) \in H$, is obviously incredibly important for thinking about what will happen after that virus’s arrival. But the concept is also very helpful for thinking about an ongoing epidemic in which we have the option of instituting policies that change the dynamics. If $S(t) \in H$, then we can end the epidemic almost immediately incurring only short run costs by imposing a super-strict lockdown for a week or two that will prevent most transmissions in that period and thereby reduce I to a very low level. When the lockdown ends and the dynamics of the system again describe its evolution, the system will quickly converge to the zero-infection steady state.⁶ Conversely, if $S(t) \notin H$, then no short-run policy will stop the epidemic. We can slow the spread with temporary restrictions (or if people temporarily reduce activity on their own), but once restrictions are removed and people go back to their former behaviors, the disease will again spread in a second wave.⁷ Hence, unless restrictions are permanent, the system must reach the herd immunity region. Note that the herd immunity region can be reached via vaccination as well as via immunity-conferring infection when vaccines that reduce susceptibility exist.

An important fact about herd immunity in the SIR model is:

- The herd immunity region of the SIR model is $H = [0, \bar{S}]$ where $\bar{S} \equiv 1/R_0$.

Suppose that the original COVID variant had $R_0 \approx 2.3$ in the US population.⁸ This would give $H \approx [0, 0.43]$, i.e. we would not be able to keep the fraction infected below 57% unless behavior/transmission remain altered until a vaccine arrives.

⁶Note that in saying this we use the fact that the fraction susceptible only goes down over time in the SIR model, and if $S(t) \in H$, then $S' \in H$ for all $S' < S(t)$.

⁷See Rachel (2024).

⁸This is the growth rate assumed by Ferguson et al. (2020) and Acemoglu et al. (2021) in their models motivated by the spread of the initial COVID strain. It is also consistent with some of the more sophisticated estimates of growth rates such as that of Miller et al. (2020).

Another important concept in SIR models is “overshooting.” People tend to be even more imprecise about what they mean by this so we’ll again give a formal definition. Write $C(t) = I(t) + R(t)$ for the cumulative fraction of the population that has been infected by time t and $C(\infty) \equiv \lim_{t \rightarrow \infty} C(t)$ for the fraction who are ever infected. Define the extent of overshooting in the course of the epidemic by $O \equiv C(\infty) - C(t_h)$ where $t_h = \min\{t | S(t) \in H\}$ is the time at which the herd immunity region is reached. In the standard SIR model $O = C(\infty) - (1 - \bar{S})$ because the fraction infected or recovered when we first reach herd immunity is $1 - \bar{S}$. Why is there overshooting? When the herd immunity threshold is first reached in the course of an epidemic we have $\dot{I}(\bar{S}, I, R) = 0$. This means that the fraction infectious is (locally) constant, with new infections occurring as fast as people are recovering. If the herd immunity threshold is reached at a point when I is large, then I will remain high for a while, and many more than $C(t_h)$ people will eventually be infected. In the standard SIR model, one can find $C(\infty)$ by solving a simple equation:

$$1 - C(\infty) = e^{-R_0 C(\infty)}$$

Intuition for this equation is that if a total of $C(\infty)$ people are infected and each interacts with an average of R_0 others while infected, then the number of times each person is on the receiving end of such an interaction will be Poisson distributed with parameter $R_0 C(\infty)$. The probability of escaping infection is therefore the probability that such a Poisson variable takes on a value of zero. It is also equal to $1 - C(\infty)$. Overshooting in the SIR model can be quite large. For example, in the SIR model with $R_0 = 2.3$, we have $C(\infty) \approx 0.87$ so $O \approx 0.30$. If it were feasible to impose a very strict lockdown at the moment when herd immunity is first reached, the benefits would be enormous.

Note that while the comments above about the unavoidability of reaching the herd immunity threshold will apply very generally even if people endogenously change their behaviors in the height of a pandemic, the formula characterizing the magnitude of overshooting is just about the pure SIR model. If temporary policies or behavioral changes reduce the rate of new infections around the time when herd immunity is reached, then the amount of overshooting that occurs would be lower.

Another useful observation about the model is:

- Define the growth rate of the infectious population by $g(t) = \frac{d}{dt} \log(I(t))$. Then, $g(t) = \gamma(R_0 S(t) - 1)$.

In the initial phase of a new epidemic when $S(t) \approx 1$, this fact says that the growth rate of the infectious population is approximately $\gamma(R_0 - 1)$. One can think of this as a cumulative growth rate of $R_0 - 1$ over the $1/\gamma$ average duration of an infection. Investigations of whether COVID restrictions were “flattening the curve” often graphed the log of cumulative infections, i.e. $\log(C(t))$, versus time. This curve will be approximately

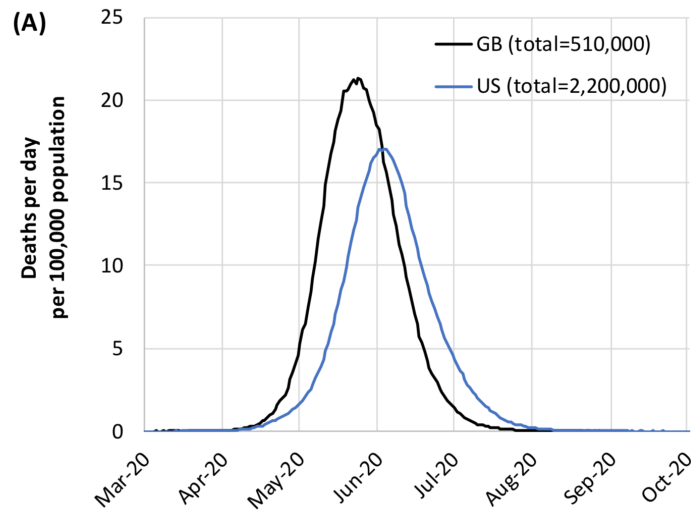


Figure 1: Figure reproduced from Ferguson (2020) Figure 1A: “Unmitigated epidemic scenarios for GB and the US. (A) Projected deaths per day per 100,000 population in GB and US.”

linear with slope $\gamma(R_0 - 1)$ as long as the ever-infected fraction of the population remains small, e.g. when the US had had 10 million cases. Attempts to infer R_0 from such curves are common given the desire to assess where the herd immunity threshold might be.

One other relevant feature of SIR models is:

- For many values of R_0 the time-path of new infections (and of deaths) has a shape that is fairly symmetric about its peak and looks somewhat like a normal density.

For example, figure 1 below reproduces Figure 1A from Ferguson et al. (2020) illustrating the predictions of an SIR-like model for Great Britain and the US.

Epidemiologists commonly work with extensions of the SIR model. Among the standard additions are an additional state E of agents who are infected but not yet infectious, more flexible recovery processes that allow non-exponential infectious durations, an explicit death state, and population inflows/outflows. Some economic models also incorporate some of these elements. To simplify the discussion I will not incorporate any of these features here, but similar conclusions should apply. Economists have also emphasized that the SIR’s model of no behavioral responses is quite unrealistic when we are in the midst of a COVID-like pandemic and this can have very important effects.⁹ The implications of heterogeneous interaction patterns could be different in such models. I will only occasionally note how conclusions might change there.

⁹See McAdams (2021) for a survey of this literature.

2.2 A heterogeneous SIR model with uniform matching

In practice, some individuals are more interactive than others. For example, supermarket cashiers will be in the vicinity of many more people in a typical day than will retirees. Epidemiologists have developed SIR extensions that allow for such heterogeneity.¹⁰

A nicely tractable version is motivated by population-wide matching in a population consisting of N equally sized subpopulations indexed by $i = 1, 2, \dots, N$. Suppose that members of group i are randomly matched with probability $R_{0i}\gamma dt$ in each dt time interval. Order the populations so that $R_{01} > R_{02} > \dots > R_{0N}$. Assume that the matchings are independent across time and uniform in the sense that the probability that a matched agent from group i meets a group j agent is $R_{0j}/\sum_k R_{0k}$, and the group j agent that is met is representative of group j in terms of having the subpopulation-mean probability of being in each infection state when the interaction occurs. The state of the population at time t is now described by a triple of vectors rather than a triple of scalars, $(S(t), I(t), R(t))$, with $S_i(t)$, $I_i(t)$, and $R_i(t)$ giving, respectively, the fraction of agents in group i who are susceptible, infectious, and recovered at time t .

Again, suppose that any matching between a susceptible and an infectious agent results in the susceptible agent becoming infectious. With the same recovery process as before, this motivates analyzing a system of differential equations:

$$\begin{aligned}\dot{I}_i(t) &= S_i(t) \sum_j \beta_{ij} I_j(t) - \gamma I_i(t) \\ \dot{S}_i(t) &= -S_i(t) \sum_j \beta_{ij} I_j(t) \\ \dot{R}_i(t) &= \gamma I_i(t)\end{aligned}$$

with $\beta_{ij} \equiv \gamma R_{0i} \frac{R_{0j}}{\sum_k R_{0k}}$. With the assumption that each subpopulation size remains constant, the state is fully described by $S(t)$ and $I(t)$ and we will usually omit $R(t)$ from the state vector.

When analyzing this model, most of the explicit characterizations that one can give are about behavior when the fraction of infectious agents is small. But it remains true that the characterizations of how the system behaves in such states have broader implications. And one can combine the formal characterizations with simulations of the model's behavior in other states to bring out other insights.

As in the previous subsection, I will first discuss the local stability of disease-free states. For any vector S^0 giving the fraction of susceptibles in each group, the disease free state $(S, I) = (S^0, 0)$ is a steady state. To

¹⁰See Andreasen and Christiansen (1989), May and Anderson (1989), Diekmann, Heesterbeek, and Metz (1990), Dushoff and Levin (1995), Jacquez, Simon, and Koopman (1995), Hethcote (2000), and Van den Driessche and Watmough (2002). The exposition below draws heavily on Dushoff and Levin (1995).

analyze the stability of such steady states and the behavior of the system in a neighborhood thereof, we linearize the system around the steady state. Note also that all derivatives $\frac{\partial \dot{I}_i}{\partial S_j}$ are equal to zero when evaluated at a state with $I = 0$. Hence, the behavior of I in a neighborhood of $(S^0, 0)$ in the full $2N$ -dimensional system has the same first-order approximation as that of I in the N -dimensional system

$$\dot{I} = A^{S^0} I,$$

where A^{S^0} is the partial derivative matrix with ij th element

$$a_{ij} = \left. \frac{\partial \dot{I}_i}{\partial I_j} \right|_{(S^0, 0)} = \begin{cases} S_i^0 \beta_{ij} - \gamma & \text{if } j = i \\ S_i^0 \beta_{ij} & \text{if } j \neq i. \end{cases}$$

In particular, the equilibrium is locally stable if all eigenvalues of this matrix have negative real parts, and unstable if any eigenvalue has a positive real part.

The A^{S^0} matrix has positive off-diagonal elements, so the eigenvalue with the largest real part is real, and corresponds to a strictly positive eigenvector. This eigenvector gives the relative prevalence of the infected across groups for which the total number of infected grows most rapidly. The special structure of this matrix allows one to easily find this eigenvector. It is $v_1 = (S_1^0 R_{01}, \dots, S_N^0 R_{0N})$, i.e. prevalence is proportional to the product of the susceptible fraction and the contact rate. The eigenvalue corresponding to this eigenvector is

$$\lambda_1 = \gamma \left(\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} - 1 \right).$$

Two important implications of this are:

1. The equilibrium $(S^0, 0)$ is locally stable if $\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} < 1$ and locally unstable if $\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} > 1$.
2. For small δ , the growth rate of the log of the total infectious population at the state $(S^0, \delta v_1)$ is approximately $\gamma \left(\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} - 1 \right)$.

Note that if we start from any state with a very small fraction δ infected, then the initial cases will initially grow at different rates in the different groups in a way that makes the distribution of cases across groups aligned with the principal eigenvector v_1 .¹¹ Hence, provided that this alignment has already occurred by the time the epidemic starts to be measured,

- The early growth of a heterogeneous-SIR epidemic with activity vector R_0 will resemble the early growth of a homogeneous-SIR epidemic with parameter $\bar{R}_0 \equiv \frac{\sum_i R_{0i}^2}{\sum_i R_{0i}}$.

¹¹Suppose the initial population infected is δv with $v = \sum a_i v_i$ where the v_i are the eigenvectors of A with eigenvalues λ_i . In a neighborhood of this point we will have $I(t) \approx \sum a_i e^{\lambda_i t} v_i$, which becomes aligned with v_1 .

This formula should be intuitive. Early in the pandemic, members of each subpopulation will pick up infections with a probability that is proportional to their subpopulation's activity level. The expected number of new infections that an infectious member of population i spawns is R_{0i} . We can thus compute the expected number of people that a newly infected agent will infect via a standard conditional expectation calculation, conditioning on the population i to which the newly infected agent belongs. This implies the expectation of the follow-on infections caused by each newly infected agent is $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} R_{0i}$.

The rewriting of \bar{R}_0 in the above expression makes clear that growth rates depend on a weighted average of group-level R_{0i} 's, with the weights being proportional to the activity level in each group. This weighted average can be substantially higher than the unweighted average. A practical implication is that the early growth rate of a heterogeneous-SIR epidemic can be substantially higher than the growth rate would be in the standard SIR model with R_0 equal to the unweighted average of the R_{0i} . Intuitively, those who are doing a lot of bumping into others have themselves been bumped into a lot recently, and hence are more likely than a random agent to be infected.

To understand the extent to which heterogeneity elevates initial growth rates in an epidemic beyond what they would be in an SIR model with a meeting rate equal to an unweighted average of the subpopulation-specific rates, a second rewriting of the relationship is useful:

$$\bar{R}_0 = \frac{\sum_i R_{0i}^2}{\sum_i R_{0i}} = \frac{E(R_{0i}^2)}{E(R_{0i})} = E(R_{0i}) + \frac{\text{Var}(R_{0i})}{E(R_{0i})}.$$

This equality indicates that growth rate is the sum of the unweighted average of the R_{0i} and the ratio of the variance of the R_{0i} across groups to the mean. The latter can easily be quite important quantitatively.

A striking implication of the importance of cross-group variance is that it is possible (for completely plausible parameter values) that reducing the R_{0i} in every group can *increase* \bar{R}_0 . A simple numerical example that illustrates the effect is a comparison of five-subpopulation models with parameters (3, 1.5, 1.5, 1.5, 1.5) and (3, 0.5, 0.5, 0.5). Most people's first thought would be that the epidemic will obviously grow faster in the first population. But plugging into the formula we see that both have $\bar{R}_0 = 2$. Why does this happen? Holding fixed the distribution of new infections across groups (in a population with $S \approx \bar{1}$) each new infection obviously causes more follow-on infections in the former model. But a critical offsetting effect is that early infections are more concentrated in the high-activity population in the latter model because a higher fraction, 60% vs. 33%, of the interactions of someone infected are with a high-activity agent.¹² It's easy to come up with modifications of this example where all relations are strict, e.g. \bar{R}_0 is larger when the activity levels are (4, 0.5, 0.5, 0.5, 0.5)

¹²Kremer (1996) discusses SI models with endogenous behavior (motivated by HIV) in which a similar mechanism can lead to counter-intuitive comparative statics.

than when the activity levels are (4.5, 1.5, 1.5, 1.5, 1.5).

Note that the intuition given above for why transmission rates are high early in the pandemic reverses later on in a pandemic. Late in a pandemic, growth rates slow in the traditional SIR model solely because some of those whom the newly infected meet are no longer susceptible. In this heterogeneous SIR model, two additional factors slow the growth. First, because many members of the high activity groups were infected long ago, a randomly selected newly-infected person is more likely to be from a moderately active group. This can make the expected number of potentially-disease-transmitting interactions that a newly infected person has below the unweighted average of the R_{0i} . Second, the fact that those whom the newly infected person meets are disproportionately high-activity people means that a larger fraction of those whom a newly infectious person meets are not susceptible. Both of these factors contribute to the slowing of the pandemic spread.

We can extend the concept of herd immunity we defined earlier to the heterogenous SIR model. I do so by defining the “herd immunity region” with almost exactly the same equation, i.e.

$$H = \{S \in [0, 1]^N | (S, 0, 1 - S) \text{ is a locally stable steady state of the model.}\}$$

The one difference is that because S is a vector, the herd immunity region is a subset of $[0, 1]^N$ rather than a subset of $[0, 1]$. The herd-behavior region is again very useful for thinking about what could possibly be accomplished with policies/behavior changes that are only temporary. Once a system has reached the herd-immunity region, a very strict brief lockdown would essentially end the pandemic. But if the system has not yet reached the herd-immunity region, then such a policy cannot stop it: if a small pocket of infection remained and interactions returned to the pre-pandemic norm, then the number of infected would again soon grow at an approximately exponential rate.

When I discuss “overshooting” in heterogeneous SIR models, I will do so thinking about a definition analogous to that used in the standard SIR model, $O \equiv C(\infty) - C(t_h)$ where $t_h = \min\{t | S(t) \in H\}$. The one change here is that I redefine $C(t) \equiv \frac{1}{N} \sum_{i=1}^N (I_i(t) + R_i(t))$ as the fraction of the full population to ever have been infected.¹³ The overshooting O represents the fraction of the population who become infected after the herd immunity regions has been reached (at which point the pandemic could have been stopped by a temporary lockdown). However, note that $C(t_h)$ can no longer be thought of as the minimum number who must be infected if all restrictions/behavioral changes end by some finite T . Temporary policies could potentially alter the path by which the epidemic reaches the herd immunity region, affecting which point of the herd immunity regions is first reached, and thereby affecting $C(t_h)$.

To give a first illustration of how the paths of homogeneous and heterogeneous SIR epidemics may differ

¹³Recall that I have assumed that the N subpopulations are equal-sized, which makes the simple arithmetic average appropriate.

away from a small neighborhood of the zero-infection state, Figure 2 graphs the time path of the new daily infections for two models: a standard homogenous SIR model with $R_0 \approx 2.3$, and a heterogeneous SIR model with five equally-sized subpopulations having activity rates 3.5, 1.5, 1, 0.5, and 0.5. The choice to illustrate a homogenous model with $R_0 \approx 2.3$ again reflects that this was thought to be a plausible parameter for the original COVID strain in the US. The choice to compare it with this particular heterogeneous SIR model reflects three considerations: (1) The model has $\bar{R}_0 \approx 2.3$, so the initial growth rates in the two models will be identical;¹⁴ (2) The coefficient of variation of the cross-group differences, 0.8, roughly matches the variation in reported contacts in the BBC Pandemic Project data, so the model is plausibly an approximation to real-world activity heterogeneity; and (3) The model has a limited number of groups and only involves whole numbers and halves, facilitating describing and remembering it. One take-away from this figure is that for reasonable parameter values the heterogeneous SIR model also produces daily infection numbers that resemble a normal density. The initial dynamics of the two models are hard to tell apart in a graph at this scale. This reflects the result that the initial growth of an epidemic in the heterogeneous population model will resemble that in a homogeneous population model with parameter \bar{R}_0 .¹⁵ The resemblance holds not just in the earliest days of the pandemic when the fraction infected is minuscule. In this illustration, the two epidemics are hard to distinguish visually 50 days into the simulation, a time when the spread would correspond to 1.5 million new infections per day in the US. Beyond this point, the two paths do soon become quite different, with the factors that we noted were relevant later in a pandemic leading to a more rapid slowdown in the heterogeneous population model. We will say more about differences in the herd immunity regions and differences in the course of the pandemic in Section 4.

2.3 A heterogeneous SIR model with homophily

While supermarket cashiers may interact with a fairly representative sample of the population, some other more and less active groups disproportionately interact with others in their own group. For example, those who frequent crowded bars and restaurants and those living in COVID-cautious senior housing communities may mostly interact with others like themselves. Heterogeneous SIR models with homophilic matching are more difficult to analyze, but epidemiologists have also derived insightful characterizations of some such models.¹⁶ I will present some results here. Again, the formal results will be about the stability of the zero-infection steady state, but understanding this stability gives broader insights and other insights can be gained from simulations.

¹⁴To make them identical, the homogenous model graphed has $R_0 = \frac{16}{7}$ which is the \bar{R}_0 of the heterogeneous model.

¹⁵As we noted, growth slows sooner in the heterogeneous model. To make this less visually apparent in the figure, I started the heterogeneous population model with a higher initial infectious rate, 0.0001 vs. 0.00005.

¹⁶These models are also sometimes referred to sometimes as models with “preferred mixing” or “like-with-like preference”.

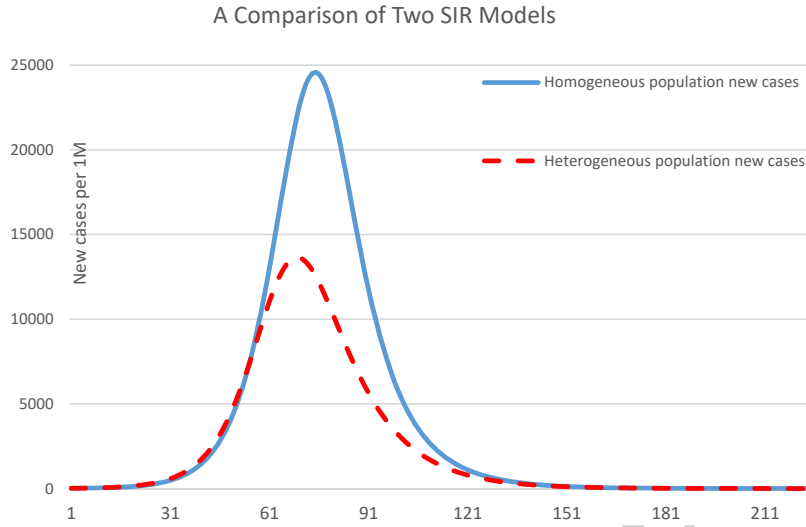


Figure 2: New daily cases in two models: The solid blue line is a homogenous SIR model with $R_0 \approx 2.3$. The dashed red line is a heterogeneous SIR model uniform matching and $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$.

To motivate one such model, consider an N group model as in the previous subsection, but suppose that when an agent from group i is randomly matched the probability that the person with whom they are matched is in group j is

$$p_{ij} = \begin{cases} h + (1-h) \frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j = i \\ (1-h) \frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j \neq i. \end{cases}$$

Such a matching process leads to an SIR model nearly identical to that in the previous subsection with

$$\dot{I}_i(t) = S_i(t) \sum_j \beta_{ij}^h I_j(t) - \gamma I_i(t),$$

where

$$\beta_{ij}^h = \begin{cases} \gamma R_{0i} (h + (1-h) \frac{R_{0j}}{\sum_k R_{0k}}) & \text{if } j = i \\ \gamma R_{0i} (1-h) \frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j \neq i. \end{cases}$$

Once again, any $(S^0, 0)$ is a steady state of the system and we can analyze the stability of this steady state by looking at a linearized N -dimensional system:

$$\dot{I} = A^{S^0 h} I,$$

where $A^{S^0 h}$ is the partial derivative matrix with ij th element

$$a_{ij}^h = \left. \frac{\partial \dot{I}_i}{\partial I_j} \right|_{(S^0, 0)} = \begin{cases} S_i^0 \beta_{ij}^h - \gamma & \text{if } j = i \\ S_i^0 \beta_{ij}^h & \text{if } j \neq i. \end{cases}$$

The off-diagonal elements of this matrix are again positive, so the eigenvector with the largest real part is again unique and corresponds to a positive eigenvector. It is no longer easy to give an explicit formula for the eigenvalue, but as noted by Diekmann, Heesterbeek, and Metz (1990) and Dushoff and Levin (1995) we can give explicit necessary and sufficient conditions for $(S, 0)$ to be a locally stable steady state.¹⁷

1. Within-group stability condition: If $S_i^0 h R_{0i} > 1$ for any i , then $(S^0, 0)$ is unstable.
2. Across-group stability condition: If $S_i^0 h R_{0i} < 1$ for all i , then $(S^0, 0)$ is unstable if $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{1}{1-hS_i^0 R_{0i}} (S_i^0 R_{0i} - 1) > 0$ and stable if $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{1}{1-hS_i^0 R_{0i}} (S_i^0 R_{0i} - 1) < 0$.

The within-group stability condition should be very intuitive. The number of infectious in subpopulation i would increase even without any cross-group transmission when $S_i^0 h R_{0i} > 1$ because a newly infected member of group i will interact with (more than) $h R_{0i}$ of group i while infectious, and hence spawn at least $S_i^0 h R_{0i} > 1$ new infections in the group. Hence, the number of new infections in this group must also be increasing in the model of this subsection, which also has some cross-group transmission.

Giving intuition for the across-group stability condition is more difficult than it was for the analogous condition for the uniform matching model because the matrix calculation is more complex. But the formula itself is not hard to parse and gives insights. Note first that when $h = 0$ the across-group stability condition simplifies to a version of the expression we gave earlier for the uniform model: $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} (S_i^0 R_{0i} - 1) < 0$. For a disease-free equilibrium to be stable in a model with $h > 0$, it must satisfy the additional constraints that $S_i^0 h R_{0i} < 1$ for all i as well as the modified inequality $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{1}{1-hS_i^0 R_{0i}} (S_i^0 R_{0i} - 1) < 0$. Note that the summation in this inequality differs from the summation for $h = 0$ in that we multiply the i th term by $\frac{1}{1-hS_i^0 R_{0i}}$. These multiplicative factors are positive for all terms, and they are larger for the terms with $S_i^0 R_{0i}$ larger. Hence, we can think of the sum as proportional to a reweighting of the $h = 0$ sum that puts greater weight on the terms with $S_i^0 R_{0i}$ large and less weight on the terms with $S_i^0 R_{0i}$ small. As a result, if the model with $h = 0$ is unstable, then the model with $h > 0$ is unstable as well.¹⁸

The argument above extends easily to a full monotonicity theorem:

- If a disease-free equilibrium is unstable for some value of h , then it is unstable for any value $h' > h$.

¹⁷See page 34 of Dushoff and Levin (1995).

¹⁸Mathematically, this is a classic Chebyshev inequality argument: if a_i and b_i are monotone increasing, then $\sum_i a_i b_i > \frac{1}{N} \sum_i a_i \sum_i b_i$.

Hence, a clear intuition one can take away from this model is that homophilic matching is an obstacle to the stability of disease free states.

The definition of the herd immunity region we gave in the last subsection applies equally well to this model with no modification. I will write H^h for the herd immunity region in this model given homophily parameter h , so the herd-immunity region for the uniform-matching model in the previous subsection is H^0 . The monotonicity result implies the herd-immunity region shrinks as h increases:

$$0 < h < h' \implies H^{h'} \subset H^h \subset H^0$$

The nesting of the herd-immunity regions implies that the minimum number that must be infected given only temporary policy/behavioral changes is larger when h is larger. It does not, however, necessarily imply that the fraction infected will actually be higher when h is larger, i.e. we cannot say (and indeed it's not always true) that $C^{h'}(\infty) \geq C^h(\infty) \geq C^0(\infty)$. There are two reasons for this: the point within the herd-immunity region that is first reached is affected by h ; and so is the extent of overshooting.

To illustrate how homophily affects the paths of heterogeneous SIR epidemics, Figure 3 graphs the time path of the new daily infections for several heterogeneous SIR model with five equally-sized subpopulations having activity rates 3.5, 1.5, 1, 0.5, and 0.5. The red dashed line is the model with uniform matching, i.e. $h = 0$. The successively darker blue lines show epidemics with h set to 0.3, 0.6, 0.9, and 0.99. The initial effects of adding substantial amounts of homophily are small qualitatively: the epidemics with $h = 0.3$ and $h = 0.6$ have a somewhat more rapid growth, but they have bell-shaped epidemic paths of about the same magnitude that simply peak a little earlier. When we increase h all the way to 0.9 the shape of the epidemic path does change: it is no longer symmetric and we see a slower decline. When $h = 0.99$ the epidemic path is no longer unimodal, and instead consists of an initial wave that rises and burns itself out in the high-activity subpopulation well before a second wave peaks (mostly in the second most active subpopulation).

Note also that it is not homophily on its own that is important to determining the dynamics of an epidemic, it is only the interaction of homophily and heterogeneity that has effects. Consider a model in which the R_{0i} are identical across groups. If we initially infect the same fraction of each subpopulation, then the dynamics of the homophilic multipopulation model are identical to those of the $h = 0$ model. Even if we were to initially introduce the infection in just one subpopulation, then as long as the initial infection is small enough (relative to how close h is to one), the infectious rate will equalize across subpopulations before levels of prevalence are substantial. Hence, all of the effects of homophily discussed above should be understood as the effects of the combination of homophily and contact heterogeneity. The effects of homophily shown in Figure 3 would be

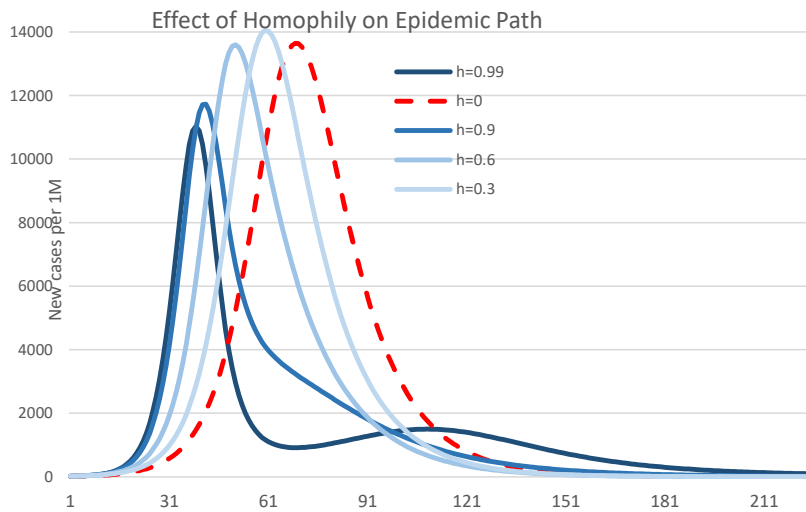


Figure 3: New daily cases with varying levels of homophily. The dashed red line is a heterogeneous SIR model uniform ($h = 0$) matching and $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$. The darkening blue lines show the effects of increasing h to 0.3, 0.6, 0.9, and 0.99.

more pronounced if the R_{0i} had been more different across groups, and would not exist if the R_{0i} were all the same.

3 Challenges Inherent in Analyzing Heterogeneous Population Epidemics

In this section and the one that follows, I turn to the task of drawing out implications of the above models for providing policy advice in an epidemic. This section stresses a cautionary implication: it can be difficult to provide policy advice in the early stages of an epidemic when we think that transmission resembles that of the heterogeneous-population SIR model. In particular, it is challenging to estimate activity rates in less active populations, and the impacts that policies will have can be sensitive to these hard-to-estimate parameters.

3.1 Difficulty in calibrating models

Early in the COVID-19 pandemic, several authors noted that it was difficult to calibrate critical parameters of even the homogeneous SIR model in the initial phase of an epidemic.¹⁹ This challenge arises again whenever new variant strains emerge. Early in an epidemic, we may not have reliable data on anything but deaths. The

¹⁹See Atkeson (2020), Fernández-Villaverde and Jones (2022), Korolev (2021), and Stock (2020).

fact that deaths in the model increase at an exponential rate makes it fairly easy to estimate the infection growth rate, which in the model is $\gamma(R_0 - 1)$. But different combinations of γ and R_0 consistent with this growth rate would lead to dramatically different future paths of the epidemic. One way to separately identify γ and R_0 is with an estimate of the death rate, because that lets us also match the parameters to the level of $I(t)$. But when many cases go unreported it is hard to calibrate the death rate. This identification problem goes away when we get other pieces of information that let us estimate the death rate. Some potential sources for this are random serology tests to estimate the fraction that have ever been infected, and fatality data from locations, e.g. South Korea or the Diamond Princess, where we think almost all cases have been identified. Another source of identification that will come somewhat later is seeing an epidemic peak, which is informative about $\bar{S} = 1/R_0$.²⁰

As more information became available, many economists analyzed SIR models with calibrated R_0 and death rate parameters. The state of the art, in fact, fairly soon moved well beyond this with several papers analyzing calibrated multipopulation SIR models that allow death rates and contacts to vary by age group.²¹ Dividing the population into age-based groups, a common practice in epidemiology when discussing childhood diseases, was an important extension when discussing COVID-19 because death rates varied so much with patient age. The contact rate calibrations in these papers relied on three survey datasets. POLYMOD (Mossong et al. (2020)) and the BBC Pandemic Project (Klepac et al. (2020)) are survey datasets which asked respondents to list those with whom they had contact in the previous 24 hours. And employment website O*Net had asked workers in a large number of occupations to report how physically close to others they worked on a 5 point scale. One can also capture some changes in activity over time using movement data available from firms with phone-tracking capabilities.

Although several papers incorporated different mean contact rates across groups, it was less common to allow for idiosyncratic variation in contact rates by breaking a population (or each age group or other cell) into subpopulations that differ in activity levels.²² The fact that predictions can be dramatically affected by heterogeneity in R_0 suggests that conclusions might have differed substantially if attempts had been made to reflect the heterogeneity that surely exists within age groups and other subpopulations. Two reasons why economic researchers did not do so are probably that solving complex dynamic problems is harder with more state variables and that the additional parameters capturing within-group heterogeneity would have needed to be calibrated.

²⁰Fernández-Villaverde and Jones (2022) note that more complex SIR models fit under a variety of assumptions about accessory parameters make very similar predictions about the future course of epidemics in locations where epidemics have peaked.

²¹See Acemoglu et al. (2021), Baqaee et al. (2020), and Favero, Ichino, and Rustichini (2020).

²²Favero, Ichino, and Rustichini (2020) is a notable exception, but it chose to only incorporate fairly small differences across its subgroups.

How might one calibrate these extra parameters? A first thought might be to calibrate the model to the path of the epidemic to-date. The initial growth rate of an epidemic plus an estimate of the death rate should let us estimate the composite parameter \bar{R}_0 mentioned earlier. However, in a heterogeneous population SIR model, there is a weak identification problem when one tries to get more than this: it can be very difficult to obtain estimates of the activity rates in the less-active populations even after there has been substantial spread of the infection. Intuitively, when there is substantial heterogeneity in the R_{0i} , there will be a substantial number of infections when the epidemic surges in the highest R_{0i} subpopulations. At that point, there may still be few infections in many of the less-active groups, particularly if matching is homophilic. This can make it very difficult to estimate activity parameters for the low infection groups from aggregate infection data.

A second approach to calibrating the extra parameters might be to use data on the variance of reported contacts in contact surveys. Although the surveys mentioned above have been used to estimate the relative prevalence of different age-group to age-group contacts, they seem less compelling as a source for estimating contact heterogeneity. For one thing, the way that contacts were defined, e.g. in the BBC survey contacts were defined as those whom one had physically touched or had a face-to-face conversation of at least three words with, leaves out many contacts that may be important in spreading COVID-19: singing near someone in a choir practice, standing near someone in a crowded bar, riding on the same subway train, being served by a cruise ship waiter, etc. The obvious heterogeneities in the frequency of such unrecorded contacts may mostly cancel out when one computes means for a large group, but we would definitely want to capture them to calibrate a model of contact heterogeneity. Another limitation of the main contact surveys is that they record contacts on a single day. Hence, recorded cross-subject variation confounds differences in cross-sectional means and time-series variation.

3.2 Difficulty in predicting future epidemic paths

The fact that some parameters of the heterogeneous SIR model are difficult to calibrate would not be troubling if the hard-to-estimate parameters of the model did not affect model predictions that we care about. This would hold if we only cared about early epidemic growth rates. Unfortunately, this does not hold for the heterogeneous SIR model given that we care very much about things like total fatalities. One reason is that activity levels in the relatively low activity groups can have a substantial impact on the long run course of the epidemic. As an illustration, Figure 4 graphs new daily cases for two heterogeneous SIR models.²³ The parameters of the two models were chosen so that new cases take off at about the same time, rise to a peak at about the same rate, and

²³Both models have ten equally-sized subpopulations with $h = 0.7$. The population with the long-lasting epidemic has $R_0 = (5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1, 1, 1)$. The population with the shorter-lived epidemic has $R_0 = (5.4, 2.6, 0.6, 0.4, \dots, 0.4)$ and a lower fraction initially infected.

peak at about the same level. Despite the nearly identical behavior up to the point when the epidemics peak, however, the epidemics proceed very differently on the way back down. By the time they die out, one epidemic has infected more than twice as many people as the other, 58% vs. 28% of the population. The fraction who will eventually be infected if policies and activity levels continue to be exactly what they are is obviously highly policy-relevant. This example illustrates that it can be very difficult to predict this even when an epidemic is sufficiently far along as to have already reached its peak, and even when policies and behaviors are simple constants that do not change during the course of the epidemic.

Intuitively, the way in which the example was constructed is that the two models each have fairly homophilic matching ($h = 0.7$) and feature a highly-active subpopulation in which the epidemic peaks before many in the less active subpopulations have been extensively infected. The models differ in the activity levels of the less-active. In one population, corresponding to the dashed red line, seven of the ten subpopulations have $R_{0i} = 0.4$. The epidemic never really takes off in these groups and this results in the fairly rapid decline in infection rates once the epidemic has burned through the highly active groups. In the other population, corresponding to the solid blue simulation, nine of the ten groups have R_{0i} equal to 1.5 or 1.0. The infections coming out of the most active group set off a spread in these groups that goes on for several months. This produces an asymmetric peak with a decline that is much more gradual than the run up. Most of the total infections occur post-peak.

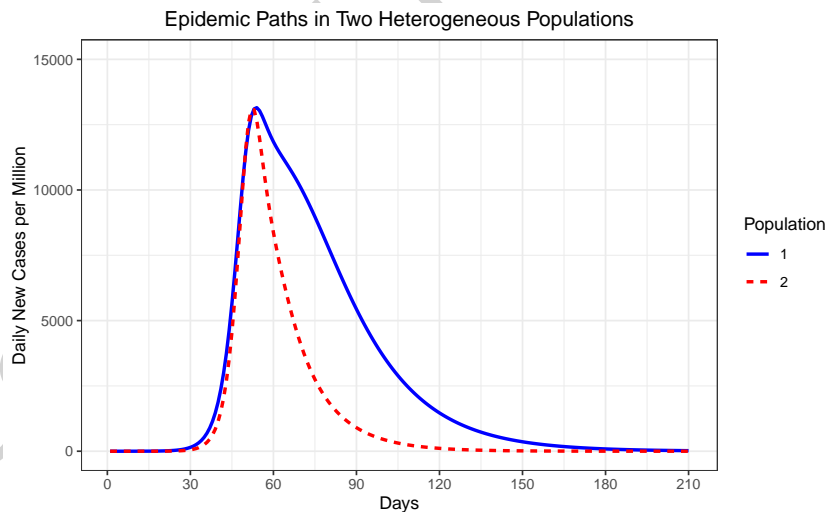


Figure 4: Example illustrating the difficulty in predicting the long-run course of a heterogeneous-population epidemic given the path of infection rates up to the point when the infection peaks. New daily cases are graphed for two ten population heterogeneous SIR models with $h = 0.7$. Model 1 has $R_0 = (5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1, 1, 1)$. Model 2 has $R_0 = (5.4, 2.6, 0.6, 0.4, \dots, 0.4)$

Analyses that do not consider the possibility that there may be heterogeneity in contact rates can report confidence intervals that are much too narrow for this reason. One clear example is that homogeneous SIR models predict that epidemics rise and fall in a nearly symmetric manner. Hence, once a peak has been observed, a homogeneous-SIR-based model will make highly confident predictions about the future course of the epidemic. For example, the April 29, 2020 update of the widely discussed IHME model gave its confidence interval for August 1, 2020 Massachusetts COVID-19 deaths as just 0 to 2, whereas actual deaths on that date turned out to be 17.

3.3 Difficulty in predicting policy impacts

As the removal of government restrictions that had limited activity became salient, a number of economic analyses modeled the effects that relaxations of restrictions might have.²⁴ Thinking about heterogeneous models, however, suggests that it will be challenging to confidently make such predictions. Uncertainty about activity levels in low-activity populations can become even more important when considering policies that relax social distancing. Intuitively, if we are far from the herd-immunity region (given the new policy), then the relaxation will set off a substantial second wave. If we are already close to or in the herd-immunity region, then the second wave will be smaller or nonexistent. Where we are relative to the herd immunity threshold depends on the full set of R_{0i} , including the hard-to-estimate activity levels in populations that have seen few infections while activity is tightly restricted.

Figure 5 provides a numerical illustration. It shows the time paths that an epidemic would follow under the same nonconstant policy path in two five-group heterogeneous SIR populations. The policy involves a severe lockdown, reducing activity levels, i.e. each group i 's R_{0i} , by 65%, imposed gradually over a two-week period just as the epidemic is taking off, and a partial relaxation about a month later that allows activity levels to return to 70% of their pre-lockdown values. The left panel plots new daily cases. The right panel plots cumulative cases to date. The vertical lines mark the dates when the initial lockdown starts its phase in and the date on which it is relaxed. The epidemics rise at very similar rates in the two populations prior to the lockdown. They have similar declines once the initial severe lockdown is imposed. Indeed, in the right panel it is very hard to see any difference in the courses of the two epidemics up through the date at which the relaxation occurs.

Despite this similarity in the initial run up and through the lockdown, the two epidemics follow very different paths following the relaxation. As in the previous example, this reflects that the parameters were chosen so that activity levels in the less active groups differ. In one population, whose outcomes correspond to the solid blue line, the relatively low activity populations have $R_{0i} = 1.5$. When we relax distancing rules, a large second

²⁴See, for example, Baqaee et al. (2020).

wave takes off in these groups, infecting nearly three times as many people as had the first wave. In the other population, corresponding to the dotted red line, the low-activity populations have $R_{0i} = 0.7$ and this makes the second wave much smaller. Difficulty in distinguishing the blue from the red population at the point when the relaxation is occurring will make it difficult to predict which future course we should anticipate.

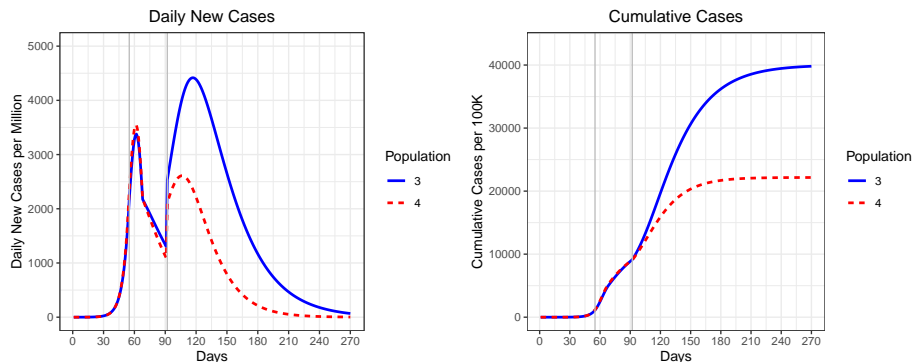


Figure 5: Example of epidemics that diverge after the partial loosening of a policy that had sharply limited interaction. The figure graphs new daily cases and cumulative cases for heterogeneous SIR models with $h = 0.7$ under a policy intervention involving a severe lockdown and a partial relaxation. Model 3 has $R_0 = (3.63, 1.5, \dots, 1.5)$. Model 4 has $R_0 = (3.55, 0.7, \dots, 0.7)$.

Note that the situation considered in this subsection differs from those discussed in the previous subsections in that it is occurring a month or two later. The fact that prevalence in less-active groups would still be fairly low would make it challenging to estimate/calibrate to transmission rates in those groups from data on the course of infections. But by that time we might have started to get a better sense of the types of interactions that are driving transmission, and perhaps could then use external data sources like cell-phone location data or contact surveys, to hazard guesses about the degree of heterogeneity that is present. The example above illustrates that the estimates of the degree of heterogeneity would have to be fairly precise to be useful. In one of the two scenarios that behave quite differently, the less active have an activity level that is roughly 22% below the population average, and in the other it is roughly 45% below.

4 Potential Biases From Ignoring Heterogeneity

Many economic analyses of the COVID-19 epidemic build on the simpler homogeneous SIR model, even though heterogeneous models seem more natural. Although calibrating models with heterogeneity in contact rates would have been more difficult, the direction of the parameter bias involved in assuming that there is no

idiosyncratic variation is clear: we are assuming less heterogeneity than exists. This section notes several ways in which ignoring or understating heterogeneity in contact rates would be expected to propagate into biases in the conclusions that analyses reach.

4.1 Overstatement of the damage incurred in reaching herd immunity

While the original COVID strain had a level of transmissibility that made it feasible (at least in some countries) to keep it in check for months or years, this was very costly. Life could not go back to normal until societies achieved herd immunity. Hence, forecasting how many cases would have to occur for herd immunity to be reached solely via immunity-conferring infection was critical to the assessment of any policy that tried to avoid this endpoint.²⁵ And knowing where we were relative to herd immunity would be important for designing and implementing policies to limit overshooting, whether by slowing the spread throughout or using strict lockdowns once herd immunity was reached. Probably the most influential early paper on this topic was Ferguson et al. (2020) which explained that the two types of mitigation strategies one could consider “differ in whether they aim to reduce the reproduction number, R , to below 1 (suppression) – and thus cause case numbers to decline – or to merely slow spread by reducing R , but not to below 1.” Estimating where the herd immunity region lay was also important for assessing the feasibility of (and later how to) achieve herd immunity via vaccination.

The Ferguson et al. (2020) paper was so influential in part because of the shocking numbers it presented. It presented simulations of a standard homogeneous SIR model estimating that uncontained spread as in that model would result in 2.2 million US deaths. Greenstone and Nigam (2020) noted that standard value-of-life calculations would indicate that the social costs in this scenario are enormous. Two critiques of these calculations are that deaths would not have been as high as the models suggest even without a government response due to endogenous social distancing, and that fatality rates may have been lower. Indeed, Ferguson et al. (2020) recognized the former in describing the simulations as what happens “In the (unlikely) absence of any control measures or spontaneous changes in individual behavior.” There is also another potentially important reason why they may have overstated the costs of reaching herd immunity that they did not mention: models with heterogeneous activity suggest that herd immunity thresholds were lower than naïve calculations based on homogeneous SIR models suggested, and that the peak infection rates that occur on this path will not be as great (and therefore as overwhelming to hospitals) as the homogeneous SIR model suggests.

In the homogeneous SIR model, herd immunity is reached when $S = 1/R_0$, implying that the fraction of the population infected on the path to herd immunity must be at least $1 - 1/R_0$. If the system is instead described by the heterogeneous SIR model with uniform matching, then the naïve estimation of an R_0 parameter may

²⁵Also critical to such calculations is an assessment of the excess deaths that may occur due to exceeding hospital capacity.

lead us to misestimate the herd immunity threshold as

$$\hat{S} = \frac{1}{\bar{R}_0} = \frac{\sum_i R_{0i}}{\sum_i R_{0i}^2}.$$

This is indeed the threshold at which herd immunity is reached if the susceptible fraction is equal in all groups, but we can reach herd immunity with fewer infected by concentrating infections in the more active groups. And infections will concentrate in the more active groups in a heterogeneous SIR epidemic.

If $R_{0i} > 1$ for all i , then one state that obviously achieves herd immunity is to set $S_i = 1/R_{0i}$ for all i . The fraction susceptible is $\frac{1}{N} \sum_i 1/R_{0i}$. That this is always greater than \hat{S} can be seen via an elegant two-step argument comparing both expressions to the reciprocal of the arithmetic mean of the R_{0i} ,

$$\frac{1}{N} \sum_i 1/R_{0i} = \frac{1}{1/\left(\frac{1}{N} \sum_i \frac{1}{R_{0i}}\right)} \geq \frac{1}{\frac{1}{N} \sum_i R_{0i}} \geq \frac{\frac{1}{N} \sum_i R_{0i}}{\frac{1}{N} \sum_i R_{0i}^2} = \hat{S},$$

with the two inequalities coming from the two parts of the root mean square-arithmetic-harmonic mean inequality. More important than the elegance is that the difference can be quite large in practical terms. For example, in the five-population example loosely calibrated to the original COVID strain, $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$, the naïve homogeneous SIR calculation gives $\hat{S} = 7/16 \approx 0.44$, suggesting that 56% of the population must be infected before herd immunity is reached. However, the $S_i = 1/R_{0i}$ state is in the herd immunity region and has just 21% of the population infected.

Recall also that the simulations of the heterogeneous SIR model that we presented in Figure 2 indicate that the peak rate of infection in an unconstrained heterogenous SIR epidemic may only be about half as high as the peak in a homogeneous SIR epidemic with the same early rate of spread. Both probably overstate the truth given that many people will self-isolate well before disease prevalence gets to the peak levels achieved in either model, but still the failure to include heterogeneity in the SIR model is another reason why concerns about the peak were overstated. And with a lower peak there would be fewer deaths due to hospital overcrowding.

Thinking more generally about the herd immunity region, the maximum fraction that can remain uninfected at any population state in the herd immunity region H is calculated by solving

$$\begin{aligned} & \max_{S_1, \dots, S_N} \sum_i S_i \\ \text{s.t.} & \sum_i \frac{R_{0i}}{\sum_k R_{0k}} S_i R_{0i} \leq 1. \end{aligned}$$

The linearity of the objective function and constraint make clear that the optimal solution involves concentrating

the infections in the highest activity groups, i.e. S_i equal to zero in the highest-activity groups, S_i equal to one in the lowest activity groups, with S_i perhaps at an intermediate level in one marginal group to make the constraint hold with equality.²⁶ This will require even fewer infections than the $S_i = 1/R_{0i}$ state. In the five-population example above, we can achieve herd immunity with just 14% of the population infected by fully concentrating infections in the highest-activity subpopulation. Note also that the “most active” subpopulation is probably much younger than the population as a whole. Hence, the reduction in deaths if one were to achieve herd immunity in this manner would be even larger than the proportional reduction in infections. While the example is clearly very loosely calibrated, the fact that the true level of infection needed to reach herd immunity is just one-fourth of what a naïve homogeneous-SIR based calculation indicates that contact heterogeneity is potentially a very important consideration.

For another example that may provide additional intuition, consider the spread of an epidemic in a less-developed country that lacks adequate personal protective equipment for its health care workers. In such an environment, transmissions from COVID-infected patients to health care workers to patients who are in hospitals for other reasons could play a major role in disease transmission. Suppose that this transmission resembled that in a ten-group uniform matching model with $R_0 = (6, 1, \dots, 1)$. The most-active group in this model could represent the health care workers. In the early stages of the epidemic any non-health care worker who is infected will infect on average one other: 0.4 health care workers and 0.6 non-health care workers. An infected health care worker will in turn infect six others, again with 40%-60% split between health care workers and others. If a homogeneous SIR model is fit to early growth of such an epidemic one would estimate $\hat{R}_0 = \bar{R}_0 = (6^2 + 1^2 + \dots + 1^2)/(6 + 1 + \dots + 1) = 45/15 = 3$ and infer that herd immunity will not be reached until two-thirds of the population is infected. In fact, herd immunity can be reached much more easily. The key is to stop the within-hospital transmission. If five-sixths of the health care workers are immune, then each new infection will lead to just one other. The health care workers are just one-tenth of the population, so we can reach herd immunity with just 8.3% of the population having been infected.

A similar comment would apply to thinking about how and whether the country could achieve herd immunity via a vaccination campaign. We do not need to vaccinate two-thirds of the population to reach herd immunity – we only need enough doses to reach 8.3% of the population if we use the doses to vaccinate the health-care workers who are playing such a big role in disease transmission. Understanding that this is what the herd immunity region looks like could very well affect recommendations about whether limited doses should be used to directly protect the most vulnerable population or to try to reach herd immunity.

²⁶Acemoglu et al. (2021) also include a discussion of targeting which point in the herd immunity region the system reaches. When death rates differ across different groups, a calculation that seeks to minimize deaths will still have a linear objective function, and hence still be maximized by concentrating cases in some groups.

In the model with homophilic matching we achieve herd immunity if we reach any state S_1, \dots, S_N such that $hS_i R_{0i} < 1$ for all i and such that

$$\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{S_i R_{0i} - 1}{1 - hS_i R_{0i}} \leq 0.$$

Here, the point in the herd-immunity region with the lowest total number of infected again involves having a lower fraction susceptible in the more active groups. But the solution will typically not be to fully concentrate the infected. Although the initial change in the constraint from reducing S_i away from one is greatest in the most active group, the marginal benefit of reducing the fraction susceptible decreases as the fraction susceptible in a group is reduced, which may make the solution interior in multiple populations.

Achieving herd immunity with homophilic matching is more difficult than achieving herd immunity with uniform matching. This follows directly from the result about the nesting of the herd immunity regions we noted in section 2.3: $H^h \subset H^0$. The minimum fraction of the population that must have been infected to achieve herd immunity is therefore monotonically increasing in h . Finding the minimum threshold is very easy in the $h = 1$ case: the model is essentially a set of separate homogeneous SIR models so the solution is simply to set $S_i = \text{Min}(1, 1/R_{0i})$ in each subpopulation. In our five-population example with $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$ this involves infecting $2/7$ of those in subpopulation 1, $1/3$ of those in subpopulation 2, and no others, which is the 21% of the total population mentioned earlier. For intermediate h one needs to solve the maximization problem described above, but we know the threshold increases continuously from 14% to 21% as h goes from 0 to 1. For $h = 0.5$ it is 15.5%.

An important factor to keep in mind when thinking about implications of results on herd immunity in heterogeneous SIR models is that overshooting is also important if we are not able to precisely estimate when herd immunity is about to be reached and/or it is not politically feasible to impose a very strict lockdown at that time. Elaborating on what was said in Section 2, in a heterogeneous SIR model uncontrolled spread infects more people than are infected in the minimal herd immunity state for two reasons: overshooting, and because the path of the infection does not concentrate infections in the high-activity population to as great a degree as does a path that enters the herd immunity region at the minimal-infection point. These additional sources of excess infections can be extremely potent in the pure heterogeneous SIR model. For example, whereas the five-population uniform-matching SIR model with $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$ (which has $\bar{R}_0 = 16/7$) can be in the herd-immunity region with as little as 14% of the population infected, an infection starting from a small evenly distributed mass of infected will not reach the herd immunity region until 33% of the population is infected, and overshooting will result in 54% eventually being infected.

I noted above that the minimal-infection herd immunity point entails more infections when matching is more homophilic. However, overshooting can be less extreme in homophilic models, and when epidemics spread in an uncontrolled manner this can more than offset the difference in the herd immunity thresholds. For example, with the same R_0 vector as above, the fraction $C(\infty)$ eventually infected is 46% with $h = 0.5$, 42% with $h = 0.75$, and just 31% with $h = 1$. Intuitively, in the homophilic model the epidemic spreads through the high R_{0i} communities more quickly and completely, but those in low R_{0i} communities fare better because their infectious rates never get as high. We are approaching the herd immunity region closer to its minimal-infection point, and overshooting can be reduced.

4.2 Overestimation of the difficulty of controlling an epidemic

While heterogeneous population models suggest that reaching herd immunity need not involve nearly as many infections as homogeneous SIR models suggest, they also suggest that avoiding herd immunity via selective lockdown policies may not be as difficult as homogeneous SIR models suggest.

Several papers have discussed optimal policies using frameworks in which transmission rates constant at time t can be reduced to $R_0(1 - x_t)$ by “locking down” a fraction x_t of the population.²⁷ This can reduce the fraction infected before a vaccine is developed and reduce excess deaths from exceeding hospital capacity. In a homogeneous SIR model, lockdown policies that keep the population from ever reaching herd immunity incur large economic costs because the fraction infected will grow unless x_0 is large enough so that $(1 - x_0)S(t)$ remains below the herd immunity threshold, and this involves locking down a large fraction of the population in the early part of the pandemic. Acemoglu et al. (2021) emphasize that this creates a grim choice between large economic losses and many deaths if policies cannot be age-targeted. In Alvarez, Argente, and Lippi (2021), the optimal policy does involve a temporary lockdown to lower peak infection rate in the benchmark case where hospital overcrowding substantially increases death rates. But in an alternate simulation where death rates are not sensitive to crowding, the optimal policy is a corner solution with no reduction in activity: the reduction in overshooting deaths on its own is not sufficient to warrant incurring the lockdown costs. In Eichenbaum, Rebelo, and Trabandt (2021) the main simulations are conducted with $R_0 = 1.5$, perhaps to avoid the conclusion that any control will be too costly.

The lower herd immunity thresholds of homogeneous models can also make control options seem less grim, and hence might have altered the conclusions of these papers. One set of policies that look better are activity-targeted permanent lockdowns. Such policies could keep the fraction infected from ever expanding while locking down a much smaller fraction of the population. For example, in the $R_0 = (3, 1.5, 1.0, 0.5, 0.5)$ example discussed

²⁷See Acemoglu et al. (2021), Alvarez, Argente, and Lippi (2021), and Rowthorn and Toxvaerd (2020).

in the herd immunity section, the problem of determining the minimum fraction the population that must be permanently locked down to keep the epidemic from ever expanding is mathematically equivalent to earlier calculation of the minimal herd-immunity threshold. Hence, the epidemic can be stopped by permanently locking down 14% of the population in the uniform matching case or at most 21% of the population in the homophilic model. This would be much less costly than locking down 57% of the population, which is what most policy papers assumed would have been necessary.

Temporary lockdowns that don't prevent a population from reaching herd immunity can also be appealing in heterogeneous population models because they can guide the system toward a more desirable part of the herd-immunity region and reduce overshooting. Here, the natural targeting approach would be the opposite of what one would do for a permanent lockdown, e.g. one could lock down members of the least-active groups once prevalence in those groups was nontrivial, keep them locked down as the infection spreads through the most-active populations, and then release them from lockdown once the population is close to herd immunity.

Figure 6 provides a numerical example illustrating that the benefits of activity-targeted temporary lockdowns can be quite large in a heterogeneous population SIR model. The solid light-blue series and the dashed light-red series are the homogeneous and heterogeneous SIR models we showed in Figure 2: a homogeneous model with $R_0 = 16/7 \approx 2.3$, and a five group uniform matching model with $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$. The infection in the heterogeneous population reaches herd immunity sooner and many fewer people are eventually infected than in the homogeneous population. But the lower damage absent a lockdown does not mean that the incremental benefit from a temporary lockdown is dramatically lower. The darker dashed red line in the figure illustrates the lower level of infections that occur in the heterogeneous population model when one implements temporary activity-targeted lockdown: we reduce activity by 20% in the highest activity group and by 60% in all of the other groups for a 60 day period. This guides the system toward a point in the herd-immunity region with fewer total infections, and it reduces overshooting by limiting the number who are infectious when the herd-immunity region is reached. In the numeric example, fraction $C(\infty)$ who are ever infected is reduced from 54% to 37%.

In the homogeneous SIR model, there is no notion of entering the herd immunity region at a better point, but temporary lockdowns still reduce $C(\infty)$ by reducing overshooting. In the textbook homogeneous SIR model, this effect can be large. Imposing the same lockdown as in the above simulation at about the same time would hardly accomplish anything in the homogeneous SIR model: it would limit infections for the 60 day period, but infections would surge when the restrictions are removed and infect almost as many in the long run.²⁸ This is illustrated by the light blue dotted line in the figure. A later lockdown, however, would be effective

²⁸We treat the policy as keeping 80% of the population under the more restrictive policy and 20% under a less restrictive policy for the 60 day period, so the dynamics are described by a heterogeneous SIR model in that 60 day period.

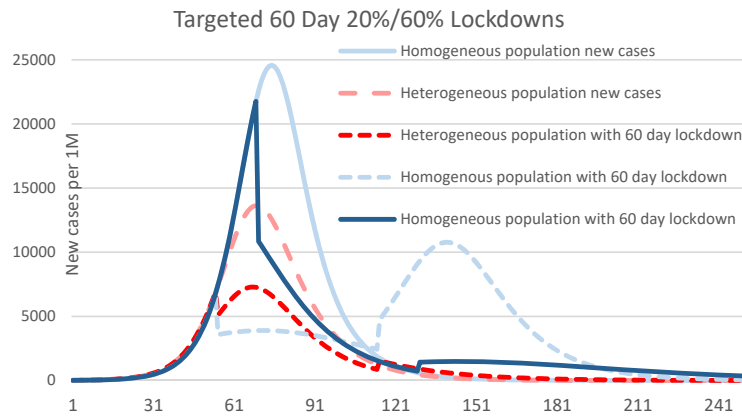


Figure 6: Effect of a temporary targeted shutdown in a heterogeneous population. The figure graphs new daily cases in five models. The lighter dashed red line is a heterogeneous SIR model with uniform matching and $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$. The darker dashed red line graphs cases for the same population with a temporary 60-day lockdown imposed during the peak. The light solid blue line is a homogeneous SIR model with $R_0 = 16/7$ and no lockdown. The other two blue lines show the effects of 60-day lockdowns imposed at different dates.

at reducing overshooting in the homogeneous SIR model. The dark blue line shows the effect of imposing the 60 day lockdown timed so that it will end around when herd immunity is reached. It reduces total infections from 87% to 67%. In practice, however, this argument for temporary lockdowns may be less plausible than that involving guiding where the system enters the herd-immunity region: the reduction in total infections is so large because an extraordinary number of people are infectious when herd immunity is reached. Such extreme infectious rates seem implausible in a world where individuals can choose to isolate themselves.

4.3 Overestimation of the impact of social distancing policies and endogenous behavioral responses

Cumulative US COVID-19 deaths grew roughly exponentially throughout March 2020, passing 5000 on April 1st. Growth subsequently slowed dramatically. Many noted that (1) government-mandated policies and (2) endogenous individual reactions would be expected to contribute to the change.²⁹ Understanding the causal impact of each factor was critical to forecasting the impact of loosening government restrictions. A third effect that would also be responsible for a portion of the slowdown got much less attention: in an SIR model the growth rate $\gamma(R_0 S(t) - 1)$ decreases as $S(t)$ declines. It is understandable that people ignored this: it would

²⁹See Baqaee et al. (2020), Farboodi, Jarosch, and Shimer (2021), Fernández-Villaverde and Jones (2022), Jones, Philippon, and Venkateswaran (2021), and Kudlyak, Smith, and Wilson (2020). Epidemiological estimates of changes in growth rates include Miller et al. (2020) and Unwin et al. (2020).

have been small in homogeneous SIR models calibrated to the US experience up to the point where case growth first slowed because $S(t)$ was still close to one.³⁰ In heterogeneous SIR models, however, this third effect can be nontrivial even in the early stages of an epidemic, particularly if matching is homophilic, and hence one can overestimate the combined effect of the first two effects.

The third effect is easiest to quantify in the uniform matching model. If the susceptible fraction has been reduced to S and the infection rate is still small, the growth of the infection will resemble that of a homogeneous SIR model with parameter $\bar{R}_0(S) = \sum_i \frac{R_{0i}}{\sum_k R_{0k}} S_i R_{0i}$. Writing \bar{S} for the average fraction susceptible, the dominant eigenvector implies that the relative frequencies in the early infected population will be roughly proportional to their activity levels, so $S_i \approx 1 - N \frac{R_{0i}}{\sum_k R_{0k}} (1 - \bar{S})$. Differentiating with respect to \bar{S} we find $\frac{d\bar{R}_0(\bar{S})}{d\bar{S}} = N \sum_i w_i^2 R_{0i}$, where we have written $w_i \equiv \frac{R_{0i}}{\sum_k R_{0k}}$ for the fraction of early infections which are in population i . Focusing just on the effect due to reductions in the most active group, we have

$$\frac{d \log \bar{R}_0(\bar{S})}{d\bar{S}} = \frac{N \sum_i w_i^2 R_{0i}}{\sum_i w_i R_{0i}} \geq \frac{N w_1^2 R_{01}}{\sum_i w_i R_{0i}} = w_1 \frac{w_1 R_{01} / \sum_i w_i R_{0i}}{1/N}.$$

Note that the first term in the product is population 1's share of early infections and the second is the ratio of population 1's contribution to \bar{R}_0 to its share of the total population. In extreme examples where almost all early infections are in one small subpopulation, this effect can be very large. For example, if $w_1 \approx 1$ in a model in which population 1 is just $1/N$ of the total population, we have $\frac{d \log \bar{R}_0(\bar{S})}{d\bar{S}} \approx N$, i.e. the apparent \bar{R} will have been reduced by about $N\%$ by the time 1% of the population has been infected. (In a homogeneous SIR model, the reduction in the apparent R_0 would be 1% .) The effect is smaller in uniform-matching models with less extreme heterogeneity in R_{0i} . For example, in the $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$ example I have used frequently, $w_1 = \frac{1}{2}$, $\frac{w_1 R_{01}}{\sum_i w_i R_{0i}} \approx \frac{3}{4}$, and $1/N = 0.2$, so $\frac{d \log \bar{R}_0(\bar{S})}{d\bar{S}} \approx 2$. This suggests that the apparent \bar{R}_0 will have been reduced by about 10% when cumulative infections have reached 5% . This is larger than the 5% prediction of a homogenous SIR model, but is not a dramatic difference.

The reduction in the apparent \bar{R}_0 can be much larger in models with homophilic matching because infections and loss of susceptibility are both more concentrated in the highest activity groups. For a simple illustration, think of a model with $h \approx 1$. Here, the power of exponential growth means that if we start from a tiny fraction infected in each group we will soon have almost all of the infected in the highest-activity group. As a result, we can perceive the growth process early in the epidemic to be close to R_{01} growth. If the infection peaks and declines in population 1 before it reaches a substantial size in population 2, the apparent growth rate can temporarily fall to well below one even though the epidemic is still in its early stages. Growth will then rise back

³⁰A study in Sweden indicated that despite their embrace of herd immunity, the fraction in Stockholm with antibodies was just 7% at the time.

to look like R_{02} growth in a second wave, and so on. For the set of activity parameters we used in constructing Figure 3, such nonmonotonic growth rates only occur when h is very, very close to one, but the fairly rapid early decline in the apparent growth rate as the epidemic burns itself out in the highest-activity population is a feature that persists well away from the $h = 1$ limit. If we take $h = 0.7$ in that example, growth that looks like $\bar{R}_0 \approx 3$ growth early in the epidemic will slow to what looks like $\bar{R}_0 \approx 2.5$ growth by the time 5% of the population has been infected. Almost 20% of the highest-activity population is no longer susceptible at this point, and this substantially reduces the epidemic growth rate.

The slowdown of an epidemic continues as it approaches and passes the herd immunity threshold. Hence, viewing an epidemic in light of a homogeneous SIR model can both lead one to mistakenly conclude that initial behavior changes were more effective than they were at slowing the epidemic and that later relaxations of policy caused less acceleration than they did.

4.4 Underestimation of heterogeneity in R_0 across regions

The SIR parameter R_0 reflects both the contagiousness of a disease and the frequency and closeness of interactions in a population. It seems natural that R_0 should be larger in some countries or states than in others. For example, we might expect it to be larger in more densely populated and highly urbanized Belgium than in Sweden. But few economic analyses incorporate heterogeneity in R_0 across regions. This presumably reflects at least in part that the early epidemiological literature did not provide clear evidence of cross-country or cross-region differences. For example, Flaxman et al. (2020) provided estimates for 11 European countries from the period before lockdowns went into effect, and the 50% credible interval for Sweden (roughly 3.7–4.3) overlaps with the 50% credible intervals for 9 of the other 10 countries including Belgium.³¹

The limited heterogeneity in reported R_0 parameters could reflect what is estimated when one applies a homogeneous SIR model to a heterogeneous world with homophilic matching. As an illustration, suppose that the differences between two countries lie not in differences between activity vectors $(R_{01}, R_{02}, \dots, R_{0N})$, but in the fact that country a has a higher fraction of its population in the high activity groups than does country b . For example, it may be that both countries have working class subpopulations living in crowded urban housing and riding public transportation to jobs where they work in close proximity to others and rural populations with much lower contact rates, with the primary cross-country difference being in the relative fractions in each group. In an extreme homophilic model with $h \approx 1$, an estimation of R_0 would yield identical estimates of $\hat{R}_0 = R_{01}$ in both countries, regardless of whether important differences in the population compositions were

³¹Unwin et al. (2020) estimates a more flexible model with more recent data and reports much more substantial heterogeneity across US states, as do Fernández-Villaverde and Jones (2022). There is also a substantial range in early estimates of the rate at which COVID-19 spread in China.

present.

Again, these differences persist well away from the $h = 1$ limit. For example, with $h = 0.5$ a model with five equal sized populations with $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$ will resemble $\bar{R}_0 \approx 2.75$ growth early in the epidemic, whereas a model with the same R_0 vector, but in which the three most active populations are each 10% of the population rather than 20% will resemble $\hat{R}_0 = 2.5$ growth. Homogeneous SIR epidemics with $R_0 = 2.75$ and $R_0 = 2.5$ follow similar paths – herd immunity is reached when 64% are infected in one model vs. 60% in the other and with overshooting the epidemics eventually infect 93% and 90%. It would be natural to not bother to incorporate differences of this magnitude in an economic analysis based on a homogeneous SIR model. But the two heterogeneous models follow quite different paths with one eventually infecting 49% of the population and the other eventually infecting 29%. Accounting for the potential impacts of cross-country differences seems much more important with this perspective.

4.5 Misestimation of when epidemics start

A number of early papers fitting SIR models produced estimates of when the COVID pandemic started. In addition to satisfying intellectual curiosity, one motivation for such an exercise is that it may provide evidence on the size of the asymptomatic population. Features of the heterogeneous SIR model suggest that it will be very difficult to produce reliable estimates via such a method. Specifically, while I emphasized earlier that a heterogeneous SIR model will appear to grow at a rapid pace \bar{R}_0 from quite early on, this is not true at the very, very beginning. The infection only starts to grow at a rate related to the largest eigenvalue once the distribution of infections across the populations is aligned with the principal eigenvector. Before this occurs, the growth rate can be much lower or a little higher depending on whether the initial infections are in a low- or high-activity population. This makes early growth rates unpredictable, and makes inferences about when an epidemic started very imprecise.

One of the most influential and inaccurate early papers on the COVID-19 epidemic may have misled in part for this reason. Lourenço et al. (2020) calibrated an SIR model to estimate the fraction of the UK and Italy populations who had already been infected as of March 19, 2020. In the three primary scenarios included in their Figure 1, they estimated that the start of the UK epidemic occurred about 30 days before the first reported death, and then projected forward to estimate that between 36% and 67% of the UK population had already been infected by that date. One reason for the inaccuracy is that the analysis assumed that the death rate was much lower than now appears to be the case. Another source of the inaccuracy, however, may be another pair of assumptions—that deaths do not occur until well after infection and that the time series of infections followed an SIR path with R_0 equal to 2.25 or 2.75 from the very beginning.

In addition to being imprecise, homogeneous SIR-based inferences about epidemic origins may be biased. Growth rates were probably lower in the very early days than they were by the time the epidemic grew to the size where estimates of R_0 were first made. This difference may help reconcile why the fraction that antibody tests indicated had ever been infected was not larger, despite revelations that there was a case in France in late December 2019 and a death in California on February 6, 2020.³² It may also help account for why some models, e.g. that shown in Figure 3 of Baqae et al. (2020), find it difficult to match data on deaths from very early in the epidemic.³³

5 Implications and Conclusions

The most basic message of this paper is that thinking about an epidemic in terms of homogeneous SIR models can lead to mistaken conclusions if the interactions are better described by a model with heterogeneous contact rates. Incorporating at least some heterogeneity need not be so difficult—in many cases what is being done with a single population model could be done quite similarly in a multipopulation model. But the remarkable pace at which the economics literature on COVID-19 progressed makes keeping up with the state of the art sufficiently difficult that my primary hope is that others will take the “heterogeneity matters” message to heart and incorporate it in their work.

Early in any epidemic, there is a great deal of scientific uncertainty about the disease transmission process, and with every new COVID variant there is uncertainty both about transmission and about the current “state” of the system given the immunity conferred by vaccinations and infections with prior strains. This paper’s most important message about the COVID-19 epidemic and its continuing variant-driven waves is that, because there is presumably a great deal of heterogeneity in potentially transmitting behaviors, we will not understand well the dynamics of any new-variant wave until it has become quite prevalent. It is particularly difficult to estimate the parameters describing how a variant is spreading in less active communities, and these parameters are critical to understanding how a wave may progress and how policies will affect it. Recognizing our limitations and doing our best to estimate the hard-to-estimate parameters is important.

The greater speed with which the apparent R_0 can decline in heterogeneous models, particularly when matching is homophilic, also suggests that there may be more uncertainty than has been recognized in papers presenting estimates of the impact of distancing policies and of reopenings using data from earlier in the pandemic. The natural directions of bias are that studies may overstate the impact that temporary mitigation

³²Worobey et al. (2020) provide genome-based evidence that later early cases were not part of the main epidemics in Washington and Italy.

³³Data inaccuracies may, of course, also be relevant here, so it is possible that the model predictions are closer to the truth than are the data.

policies had in slowing the spread of COVID-19 and may underestimate the extent to which subsequent relaxations accelerated the spread. For a similar reason, we could overestimate the extent to which a pandemic was slowed by endogenous reductions in activity. It is particularly important to keep these biases in mind when estimates obtained in some location at some point in time are used to provide advice for dealing with later waves.

A more optimistic implication of heterogeneous SIR models is that COVID-19 waves may often be less severe than standard SIR-based models suggest. Models using growth rates estimated in the early days of any variant's emergence may overstate how rapidly the variant will spread even absent social distancing. And it is possible that epidemic growth can be slowed by herd immunity effects at prevalence levels substantially lower than naïve models suggest. Indeed, many Omicron variants reported to have extremely high values of R_0 do not seem to have spread as widely as standard SIR models suggested they would.

The set of highly active individuals who one would want to target for vaccination to reach the herd immunity region with limited capacity does not seem to align at all well with who has been most interested in getting the vaccine. But it is still somewhat encouraging to think that a wave that quickly confers immunity on many of the most active by infecting them may enter the herd immunity region at lower levels of prevalence than naive calculations suggest because of the combination of these infections and vaccinations. This is especially true if those outside the most active group continue to limit their interactions and wear masks when prevalence levels are high. And the damage caused by each wave can be limited if (as seems to be the case) those in vulnerable populations are especially vigilant. If so, the option of reaching herd immunity through natural infection (conditional on this behavior), becomes less unattractive.

Another important conclusion, however, is that the optimistic message that reaching herd immunity may not be as damaging as feared should not be taken to imply that trying to reach herd immunity is more advisable than earlier analyses suggest. Models with heterogeneity also suggest that controlling variant-driven waves may be easier than thought. Benefits similar to those which herd immunity provides can be obtained by implementing targeted measures to prevent high-contact people, e.g. health care and nursing home workers, from ever being infected. This makes measures such as ensuring nursing home workers have adequate personal protective equipment even more powerful and cost-effective. Permanent and temporary targeted mitigation measures like the use of high quality masks may also be more effective as a means to limit the spread of the epidemic than homogeneous SIR models suggest, even when apparent R_0 values are in the very high ranges that have been provided for some variants. Correctly accounting for heterogeneity in the epidemic process therefore provides “good news” that could bolster the case for policies that mandate mitigation when new variants are spreading. Obviously, it would be valuable to know more about the nature of contact heterogeneity (and about

the long-run health consequences of COVID-19 for survivors) to make this assessment.

I also noted that estimating SIR models on data early in an epidemic may lead one to underestimate the extent to which critical parameters of the epidemic process differ across regions. The changes in the course of epidemics in the aftermath of severe lockdown policies indicate that, in the aggregate, policies and behavioral changes had a large impact on R_0 . The limits to what is safe, however, could be very different in different locations and at different times. It would be valuable to be able to provide tailored guidance.

In addition to the computational challenges, a factor that will limit our ability to calibrate more complex models is the limited data that is available on heterogeneity in contact rates. Although several firms have already made location tracking data available to researchers, privacy concerns have limited public releases to means within various cells. One simple step that could potentially greatly enhance the value of this data is to also release within-cell variances and within-individual time series correlations. Epidemiologists are also able to exploit variation in virus genomes to provide more micro-based estimates of disease-transmission.³⁴ While economists are unlikely to have the expertise to take advantage of genomic data, keeping current on insights coming out of these analyses will be important.

This paper contains a number of “negative” results: noting mistaken conclusions one can reach if models do not include heterogeneity in matching rates; and noting that estimating some important parameters will be difficult. But I do not think of its message as a negative one. I think of it as noting that our models can be fairly easily improved by incorporating heterogeneity in matching rates.

What should an economic model of epidemic policy look like given all of the advances that have been made in the last few years? Any model intended to be relevant to COVID policy should model include age-based (or vulnerability-based) population subgroups to reflect that death rates have been dramatically different across age groups. Models should include some type of behavioral responses to disease prevalence. Although this response has varied across time and space and not always looked like a fully rational response to short-term or lifetime risk, it’s clearly present and makes the extreme prevalence that occurs in some models completely implausible. Models should incorporate what can (and cannot) be done with current vaccine technologies/approval processes relative to the timeline of a pandemic. And I think any model should include substantial heterogeneity in contact rates, whether exogenous or reflecting heterogeneous economic costs or behavioral attitudes toward distancing. Given that model outcomes are sensitive both to the activity levels of the highly active and to the sizes of the subpopulations with R_0 somewhat above and below one, I think that models should try to include at least four activity-differentiated subgroups within each age (or otherwise defined) group. Obviously, it would be ideal to be able to estimate the sizes of the activity-differentiated subgroups from very recent data from whatever

³⁴See, for example, Miller et al. (2020) and Worobey et al. (2020).

pandemic wave is then relevant. Absent this, it would be helpful to have some standard benchmark for relative contact rates produced by applying a clustering algorithm to data on COVID-era interactions. But even absent data-derived estimates, simply putting in some plausible ballpark values like those in my benchmark examples would be progress.

Heterogeneity in contact rates also seems like it should have a more prominent role in the literature on vaccines. The number of people who were highly vulnerable to COVID was quite large, making it quite time-consuming to reach them with vaccines. This is likely to remain true in future pandemics. While the skeptical preferences of some of the most active would make it difficult to achieve similar vaccination rates in the most active groups, the combined effect of vaccinating the willing active and having the unwilling active get infected could potentially be an alternate way to protect the vulnerable.

Accepted manuscript

References

- [1] Daron Acemoglu, Victor Chernozhukov, Iván Werning, and Michael D Whinston. “Optimal Targeted Lockdowns in a Multi-Group SIR Model”. In: *American Economic Review: Insights* 3.4 (2021), pp. 487–502.
- [2] Fernando Alvarez, David Argente, and Francesco Lippi. “A Simple Planning Problem for COVID-19 Lockdown, Testing, and Tracing”. In: *American Economic Review: Insights* 3.3 (2021), pp. 367–82.
- [3] Viggo Andreasen and Freddy B Christiansen. “Persistence of an infectious disease in a subdivided population”. In: *Mathematical Biosciences* 96.2 (1989), pp. 239–253.
- [4] Andrew Atkeson. “How deadly is COVID-19? Understanding the difficulties with estimation of its fatality rate”. NBER Working Paper No. 26965. 2020.
- [5] Andrew Atkeson, Karen Kopecky, and Tao Zha. “Behavior and Transmission of COVID-19”. In: *AEA Papers and Proceedings* 111 (2021), pp. 356–360.
- [6] Christopher Avery, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison. “An Economist’s Guide to Epidemiology Models of Infectious Disease”. In: *Journal of Economic Perspectives* 34 (2020), pp. 79–104.
- [7] Christopher Avery, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison. “Policy implications of models of the spread of coronavirus: Perspectives and opportunities for economists”. In: *Covid Economics* 12 (2020), pp. 21–68.
- [8] Christopher Avery, Frederick Chen, and David McAdams. “Steady-State Social Distancing and Vaccination”. In: *American Economic Review: Insights* 6 (2024), pp. 1–19.
- [9] David Baqaee, Emmanuel Farhi, Michael J Mina, and James H Stock. “Reopening Scenarios”. NBER Working Paper No. 27244. 2020.
- [10] Tom Britton. “Estimation in multitype epidemics”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.4 (1998), pp. 663–679.
- [11] Tom Britton, Frank Ball, and Pieter Trapman. “A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2”. In: *Science* (2020).
- [12] Eric Budish. “ R_1 as an Economic Constraint”. In: *Review of Economic Design* (2024).
- [13] David Champredon, Michael Li, Benjamin M Bolker, and Jonathan Dushoff. “Two approaches to forecast Ebola synthetic epidemics”. In: *Epidemics* 22 (2018), pp. 36–42.
- [14] Nikolaos Demiris and Philip D O’Neill. “Bayesian inference for stochastic multitype epidemics in structured populations via random graphs”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.5 (2005), pp. 731–745.
- [15] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz. “On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations”. In: *Journal of Mathematical Biology* 28.4 (1990), pp. 365–382.
- [16] Michael Droste, Andrew Atkeson, James H. Stock, and Michael J. Mina. “Economic Benefits of COVID-19 Screening Tests”. In: *Review of Economic Design* (2024).
- [17] Michael Droste and James H. Stock. “Adapting to the COVID-19 Pandemic”. In: *AEA Papers and Proceedings* 111 (2021), pp. 351–355.
- [18] Jonathan Dushoff and Simon Levin. “The effects of population heterogeneity on disease invasion”. In: *Mathematical Biosciences* 128.1-2 (1995), pp. 25–40.

- [19] Martin S Eichenbaum, Sergio Rebelo, and Mathias Trabandt. “The macroeconomics of epidemics”. In: *The Review of Financial Studies* 34.11 (2021), pp. 5149–5187.
- [20] Glenn Ellison. “Implications of Heterogeneous SIR Models for Analyses of COVID-19”. NBER Working Paper No. 27373. 2020.
- [21] Maryam Farboodi, Gregor Jarosch, and Robert Shimer. “Internal and external effects of social distancing in a pandemic”. In: *Journal of Economic Theory* 196 (2021), p. 105293.
- [22] Carlo A Favero, Andrea Ichino, and Aldo Rustichini. “Restarting the economy while saving lives under Covid-19”. CEPR Discussion Paper No. DP14664. 2020.
- [23] Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, et al. “Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. 2020”. In: *DOI* 10 (2020), p. 77482.
- [24] Jesús Fernández-Villaverde and Charles I Jones. “Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities”. In: *Journal of Economic Dynamics and Control* 140 (2022), p. 104318.
- [25] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Unwin, Helen Coupland, T Mellan, Harisson Zhu, T Berah, J Eaton, P Perez Guzman, et al. “Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries”. In: (2020).
- [26] Rachel Glenester, Thomas Kelly, Claire McMahon, and Christopher M. Snyder. “Quantifying the Social Value of a Universal COVID-19 Vaccine and Incentivizing its Development”. In: *Review of Economic Design* (2024).
- [27] M Gabriela M Gomes, Ricardo Aguas, Rodrigo M Corder, Jessica G King, Kate E Langwig, Caetano Souto-Maior, Jorge Carneiro, Marcelo U Ferreira, and Carlos Penha-Goncalves. “Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold”. In: *medRxiv* (2020).
- [28] Matthew Goodkin-Gold, Michael Kremer, Christopher M. Snyder, and Heidi Williams. “Optimal Vaccine Subsidies for Epidemic Diseases”. In: *Review of Economics and Statistics* forthcoming (2022).
- [29] Michael Greenstone and Vishan Nigam. “Does social distancing matter?” In: *Covid Economics* 7 (2020), pp. 1–22.
- [30] Herbert W Hethcote. “The mathematics of infectious diseases”. In: *SIAM review* 42.4 (2000), pp. 599–653.
- [31] John A Jacquez, Carl P Simon, and James S Koopman. “Core Groups and the R0s for Subgroups in Heterogeneous SIS and SI Models”. In: *Epidemic Models: Their Structure and Relation to Data* 5 (1995), p. 279.
- [32] Siamak Javadi, Elena Quercioli, and Lones Smith. “Strategically Rational Risk Taking by Age in COVID-19, and the Heterogeneous Agent Behavioral SIR Model”. In: *Review of Economic Design* (2024).
- [33] Siamak Javadi, Elena Quercioli, and Lones Smith. “Strategically Rational Risk Taking by Age in COVID-19, and the Heterogeneous Agent Behavioral SIR Model”. In: *Review of Economic Design* (2024).
- [34] Callum J Jones, Thomas Philippon, and Venky Venkateswaran. “Optimal Mitigation Policies in a Pandemic: Social Distancing and Working from Home”. In: *The Review of Financial Studies* 34.11 (2021), pp. 5188–5223.
- [35] William Ogilvy Kermack and Anderson G McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the Royal Society of London. Series A* 115.772 (1927), pp. 700–721.

- [36] Petra Klepac, Adam J Kucharski, Andrew JK Conlan, Stephen Kissler, Maria L Tang, Hannah Fry, and Julia R Gog. “Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project”. In: (2020).
- [37] Ivan Korolev. “Identification and Estimation of the SEIRD Epidemic Model for COVID-19”. In: *Journal of Econometrics* 220.1 (2021), pp. 63–85.
- [38] Michael Kremer. “Integrating Behavioral Choice into Epidemiological Models of the AIDS Epidemic”. In: *Quarterly Journal of Economics* 111.2 (1996), pp. 549–573.
- [39] Mariana Kudlyak, Lones Smith, and Andrea Wilson. “For whom the bell tolls: avoidance behavior at breakout in an SI3R model of covid”. In: *Virtual Macro Seminar*. 2020.
- [40] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. “Superspreading and the effect of individual variation on disease emergence”. In: *Nature* 438 (2005), pp. 355–359.
- [41] José Lourenço, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. “Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic”. In: *medRxiv* (2020).
- [42] Robert M May and Roy M Anderson. “The transmission dynamics of human immunodeficiency virus (HIV)”. In: *Applied Mathematical Ecology*. Springer, 1989, pp. 263–311.
- [43] David McAdams. “Nash SIR: An Economic-Epidemiological Model of Strategic Behavior During a Viral Epidemic”. In: *Covid Economics* 16 (2020), pp. 115–134.
- [44] David McAdams. “The Blossoming of Economic Epidemiology”. In: *Annual Review of Economics* 13 (2021), pp. 539–570.
- [45] Danielle Miller, Michael A Martin, Noam Harel, Talia Kustin, Omer Tirosh, Moran Meir, Nadav Sorek, Shiraz Gefen-Halevi, Sharon Amit, Olesya Vorontsov, et al. “Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel”. In: *medRxiv* (2020).
- [46] Jo el Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases”. In: *PLOS Medicine* 5.3 (2020), pp. 381–391.
- [47] Lukasz Rachel. “The Second Wave”. In: *Review of Economic Design* (2024).
- [48] Adriano A Rampini. “Sequential lifting of covid-19 interventions with population heterogeneity”. NBER Working Paper No. 27063. 2020.
- [49] Bob RE Rowthorn and Flavio Toxvaerd. “The optimal control of infectious diseases via prevention and treatment”. In: (2020). Cambridge-INET Working Paper.
- [50] James H Stock. “Coronavirus Data Gaps and the Policy Response to the Novel Coronavirus”. In: *Covid Economics* 3 (2020), pp. 1–11.
- [51] Flavio Toxvaerd. “Equilibrium Social Distancing”. In: *Covid Economics* 15 (2020), pp. 110–133.
- [52] Thomas Troger. “Optimal Testing and Social Distancing”. In: *Review of Economic Design* (2024).
- [53] H Juliette T Unwin, Swapnil Mishra, Valerie C Bradley, Axel Gandy, Michaela Vollmer, Thomas Mellan, Helen Coupland, Kylie Ainslie, Charles Whittaker, Jonathan Ish-Horowicz, et al. “State-level tracking of COVID-19 in the United States”. In: (2020).
- [54] Pauline Van den Driessche and James Watmough. “Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission”. In: *Mathematical Biosciences* 180.1-2 (2002), pp. 29–48.

- [55] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, et al. “The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt”. In: *Epidemics* 22 (2018), pp. 13–21.
- [56] Michael Worobey, Jonathan Pekar, Brendan B Larsen, Martha I Nelson, Verity Hill, Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and Philippe Lemey. “The emergence of SARS-CoV-2 in Europe and the US”. In: *bioRxiv* (2020).

Accepted manuscript