

RESEARCH ARTICLE

Novel estimates reveal subnational heterogeneities in disease-relevant contact patterns in the United States

Casey F. Breen^{1*}, Ayesha S. Mahmud¹, Dennis M. Feehan

Department of Demography, University of California, Berkeley, Berkeley, California, United States of America

* caseybreen@berkeley.edu



Abstract

Population contact patterns fundamentally determine the spread of directly transmitted airborne pathogens such as SARS-CoV-2 and influenza. Reliable quantitative estimates of contact patterns are therefore critical to modeling and reducing the spread of directly transmitted infectious diseases and to assessing the effectiveness of interventions intended to limit risky contacts. While many countries have used surveys and contact diaries to collect national-level contact data, local-level estimates of age-specific contact patterns remain rare. Yet, these local-level data are critical since disease dynamics and public health policy typically vary by geography. To overcome this challenge, we introduce a flexible model that can estimate age-specific contact patterns at the subnational level by combining national-level interpersonal contact data with other locality-specific data sources using multilevel regression with poststratification (MRP). We estimate daily contact matrices for all 50 US states and Washington DC from April 2020 to May 2021 using national contact data from the US. Our results reveal important state-level heterogeneities in levels and trends of contacts across the US over the course of the COVID-19 pandemic, with implications for the spread of respiratory diseases.

OPEN ACCESS

Citation: Breen CF, Mahmud AS, Feehan DM (2022) Novel estimates reveal subnational heterogeneities in disease-relevant contact patterns in the United States. *PLoS Comput Biol* 18(12): e1010742. <https://doi.org/10.1371/journal.pcbi.1010742>

Editor: Daniel B. Larremore, University of Colorado Boulder, UNITED STATES

Received: April 28, 2022

Accepted: November 16, 2022

Published: December 2, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1010742>

Copyright: © 2022 Breen et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All code and data used to produce the analyses presented in this manuscript are available from an Open Science

Author summary

Reliable quantitative estimates of contact patterns are in high demand because they are a critical input for modeling and reducing the spread of directly transmitted infectious diseases such as COVID-19 and influenza. But, while national-level contact data are available from surveys collected in many countries—including the United States—local-level estimates of age-specific contact patterns are almost never available. Yet these local-level contact patterns are crucial, since disease dynamics and public health policy vary by geography. Here, we introduce a new method for estimating levels and trends in local-level rates of disease-relevant contacts by combining information from national-level contact data and other locality-specific data sources. We illustrate our method by estimating novel state-level estimates of contact patterns in the US over the course of the COVID-19 pandemic. These estimates are an important resource for researchers and policymakers who wish to understand the state-level dynamics of the COVID-19 pandemic in the United

Framework Repository: <https://doi.org/10.17605/OSF.IO/AECWN>.

Funding: A Berkeley Population Center pilot grant (National Institute of Child Health and Human Development P2CHD073964) was awarded to ASM and DMF. DMF received additional funding from the Hellman Fellows Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

States. The new methods we introduce are applicable to any directly transmitted infectious disease and can be applied in many different settings all over the world.

This is a *PLOS Computational Biology* Methods paper.

Introduction

As of April 2022, the COVID-19 pandemic is estimated to have infected over 505 million people and caused at least 6.2 million deaths around the world [1]. Like many other directly transmitted diseases—including influenza, pertussis, and measles—the pathogen that causes COVID-19 is spread through respiratory droplets that can be transmitted during close, person-to-person interactions [2–6]. For these directly transmitted infectious diseases, reliable quantitative estimates of levels and patterns of interpersonal contact are central to measuring, modeling, and reducing transmission. Contact patterns serve as critical inputs to mathematical disease models, allowing for the direct incorporation of contact heterogeneity by age, geography, and time to more realistically simulate disease transmission dynamics [7, 8].

While national-level estimates of age-specific contact patterns are available in many countries [6, 9–13], local-level estimates are almost never available. Yet these local-level contact patterns are crucial, since disease dynamics and public health policy vary by geography within a country. This need for timely local-level estimates of contact patterns is especially apparent during the COVID-19 pandemic. In the US, for example, state-to-state differences in physical distancing policies, epidemic waves, seasonality, culture, and other factors have led to many heterogeneous COVID-19 outbreaks, rather than one coherent national outbreak. To accurately model these local COVID-19 outbreaks, each state would ideally collect its own periodic contact surveys to quantify changing local patterns of interpersonal interaction over the course of the pandemic. However, repeatedly collecting contact data for each state would require a very large sample, making it prohibitively expensive.

To overcome this challenge, we introduce a novel approach to estimating subnational interpersonal contact patterns by combining information from national-level contact data and other locality-specific data sources—such as census data—using multilevel regression with poststratification (MRP) [14, 15]. Conceptually, our approach proceeds in two steps: first, we fit Bayesian hierarchical models to age-specific contact patterns observed in the national-level contact survey data. The purpose of these models is to predict age-specific contact rates based on geographic and demographic covariates. Second, we poststratify our model-based estimates by aggregating them up to state-level age-stratified contact rates using weights that represent the population composition of each state. Age-stratified contact rates are key because of age-related variation in transmission, mortality, and immunity [16, 17].

We apply our method to estimate daily age-stratified contact matrices for all 50 US states and Washington DC between April 2020 and May 2021. All estimates are publicly available at https://osf.io/aecwn/?view_only=0de9a7de36b241d1a5f3f7174629e368. Our results reveal substantial state-level heterogeneity in levels and trends of contacts across the US over the course of the COVID-19 pandemic, with overall contact intensity being highest for the Southern and Rocky Mountain states and lowest for states on the West Coast and in New England. Contact patterns generally became increasingly assortative with respect to age over our observation window, with significant state-level differences in these trends. Taken together, our results demonstrate that national-level contact patterns are not a suitable replacement for state-level contact patterns. Our method can be applied in many other settings where national-level

contact data is available, but local-level estimates are needed. The resulting age-stratified contact matrices are relevant for a range of infectious diseases, such as COVID-19, influenza, pertussis, and measles.

Data and methods

Data

We use data from Waves 1–6 of the Berkeley Interpersonal Contact Study (BICS), a nationally-representative study of interpersonal interaction in the US [11]. Respondents were asked to report on the number of household and non-household persons they had *conversational contact* with the day before the survey. Conversational contact was defined as a two-way conversation with three or more words in the physical presence of the other person. This is the definition used by the POLYMOD project [9]. Respondents were asked to provide age and sex information on all household members and detailed information on up to three of their non-household contacts, including demographic profile, length of interaction, and where the conversation took place. These detailed contacts were up-weighted to reflect the total number of contacts, for which detailed information may or may not have been collected (see Section 1.1 in S1 Appendix for details).

Six waves of data were collected from April 2020 to May 2021. Waves were generally conducted over a few weeks. Table 1 shows the number of interviews conducted per month.

New approach to estimating subnational contact patterns

Our paper introduces a new method for estimating levels and trends in state-level rates of disease-relevant contacts in the United States over the course of the COVID-19 pandemic by combining information from national-level contact data and other locality-specific data sources. The methods we introduce are applicable to any directly transmitted infectious disease—including, for example, influenza, pertussis, and measles—and can be applied in many different settings all over the world.

Researchers construct age-structured contact matrices to summarize patterns of interpersonal contact in a population. These age-structured contact matrices describe average rates of contact between age groups and serve as a key input to most mathematical models of directly transmitted infectious diseases. Formally, we assume the population has been divided into A discrete age groups. Let c_{ij} be the average number of contacts that a person in age group i has with people in age group j over a given time period. Then the contact matrix $C = (c_{ij})$ is the $A \times A$ matrix that is composed of the expected number of contacts between each pair of age groups.

A growing body of methodological work is concerned with how best to estimate the entries c_{ij} of the contact matrix C [6, 9, 10, 12, 13, 18, 19]. A conventional strategy, called indirect

Table 1. Number of BICS interviews administered monthly.

Wave	Month	Year	Interviews
1	April	2020	856
1	May	2020	525
2	June	2020	1113
3	September	2020	1486
4	December	2020	1094
5	February	2021	1304
6	May	2021	2412

<https://doi.org/10.1371/journal.pcbi.1010742.t001>

estimation, is based on fitting models that back-estimate the entries in the contact matrix C from epidemiological data, such as case-counts or estimates of seroprevalence. Indirect estimation is appealing because these methods do not require special data collection. However, this convenience comes at a cost—indirect estimation requires making several assumptions that are hard to validate or check. More recently, attention has turned to direct estimation methods which involve collecting data to estimate each entry c_{ij} of the contact matrix. Several different types of data can be collected to achieve this goal: previous studies have explored estimating c_{ij} from time-use data [10, 18], wearable sensors [20, 21], and surveys, including diary studies [9, 11, 22].

This study focuses on using surveys that ask respondents to report about their face-to-face interactions with other people [6]. This direct estimation approach is appealing because it is extremely flexible. Surveys can be used by researchers to: (1) make inferences about contact patterns in a population that is not constrained to a small geographic area, (2) gather information about the interactions most relevant to the spread of directly transmitted pathogens, and (3) collect important information beyond the contacts themselves, including contextual information about interpersonal interactions (e.g., their location and duration) and about the people who interacted (e.g., their age, sex, occupation, and relationship to one another). This contextual information can be very useful when developing policies to limit contacts in the hopes of slowing the spread of a directly transmitted infectious disease.

While surveys or diary studies have now been used to produce national-level estimates of age-specific contact patterns in many countries around the world [6, 9–11, 23–25], the majority of countries still lack nationally representative contact surveys. One stream of research has developed an approach for extrapolating synthetic national-level contact matrices in countries where no contact data are available at all [12, 13]. Principled national-level extrapolations of contact patterns are vastly preferable to no information at all. These principled extrapolation methods can be seen as a complement to the methods we introduce here, which can instead be used when national-level contact data are available but researchers wish to make fine-grained, local-level inferences. Our approach is thus more similar to interpolation, as opposed to extrapolation.

In order to use nationally representative BICS data to estimate contact matrices at the state-level [26–31], we extend a promising approach called multilevel regression with poststratification (MRP). First developed by Gelman and Little (1997), MRP has recently been extended to estimate subnational public opinion [14, 15, 32, 33], forecast elections [34], and estimate small-area population health outcomes [35]. This study is the first application of MRP to the estimation of social contact patterns.

MRP consists of two main steps. In the first step (multilevel regression), we fit Bayesian hierarchical models to estimate contact patterns for all poststratification “cells”—mutually exclusive groups created by cross-classifying key covariates. For example, one cell used in our study is Non-Hispanic Black women aged 25–35 living in New York in a household of size 4. In the second step (poststratification), we aggregate the cell-level estimates up to the state-level by weighting each cell by its relative proportion in the state. For our study, we add a third step, which takes state-level estimates and aggregates them up to contact matrices.

MRP has traditionally been used to study static outcomes. However, we needed to estimate contact matrices for each US state over many time periods. One approach is to use a separate MRP model for each time period (e.g., year) and combine these separate MRP models into a single time-series model. However, combining separate MRP models doesn’t make full use of all available information and is less flexible for incorporating time trends into a model. Therefore, we fit a single “dynamic MRP” model explicitly incorporating a time trend [15]. This model allows us to produce estimates for all 51 states and 411 different time periods (days). Dynamic MRP allows us to take advantage of all data available when making predictions.

To summarize, we follow three steps to estimate state-level contact matrices from nationally-representative BICS data: (1) we fit Bayesian hierarchical models to the national-level survey data; (2) we poststratify model estimates; (3) we calculate state-level contact matrices from the poststratified estimates. We now describe each of these steps in turn.

Models

Our hierarchical models overcome the instability of estimates from very finely partitioned cells by estimating cell-level averages from a hierarchical model that allows for “random” or “modeled” effects for some predictors. Random effects allow for partial pooling towards the group mean, with greater pooling for less-populated cells [27]. When little to no data are available for a cell, estimates are based on data from similar cells; when lots of data are available for a cell, estimates are closer to the original survey responses.

There are fundamental differences between household contacts and non-household contacts; for example, we would expect non-household contacts to change substantially over the course of the pandemic while household contacts remain relatively stable. Therefore, we develop two flexible hierarchical models for the contact data in the BICS survey: one for non-household contacts and one for household contacts. We estimate the total number of contacts by summing the predicted number of household and non-household contacts.

For the non-household contacts models, we include a national-level cubic time trend and state-level hierarchical cubic time trend, which allows for state-level deviations from the national time trend. We chose our final model by fitting a series of models with different specifications and performing model selection using Pareto-smoothed importance sampling leave-one-out cross validation (PSIS LOO-CV) to select the model with the best performance [36]. For further details, see Section 2.3 in [S1 Appendix](#).

We incorporate Google Community Mobility reports on visits to seven different place categories—grocery and pharmacy, parks, transit stations, retail and recreation, residential, and workplaces [37, 38]—into our model. We follow [39] by aggregating these seven signals into one predictor, the three-week rolling average of the first principal component. We find this mobility predictor only marginally increases our model’s performance (see Table A in [S1 Appendix](#)). However, we include the mobility predictor in our model to highlight the model’s ability to incorporate state-level contextual predictors. Highly comparable estimates from a model without the mobility predictor are presented in Section 3.4 in [S1 Appendix](#), demonstrating this method can be used without mobility data.

For our model, we let $y_{i,c}$ be the number of contacts person i has with people in age category c for $1 \leq c \leq C$. Then $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,C})^T$ is a vector with the number of contacts person i has with people in each age group. We fit separate models to predict the total number of contacts (calculated by up-weighting detailed contacts) in each of the 7 defined age categories: 0–17, 18–24, 25–34, 35–44, 45–54, 55–64, and 65+. We assume each $y_{i,c}$ follows a negative binomial distribution,

$$y_{i,c} \sim \mathcal{NB}(\mu_{i,c}, \phi), \quad (1)$$

where $\mu_{i,c}$ is the expected number of contacts. We model the log of the expected number of contacts using

$$\begin{aligned} \log(\mu_{i,c})_{\text{non-household}} = & \alpha_{0,c} + a_{j[i],c}^{\text{gender}} + a_{j[i],c}^{\text{age category}} + a_{j[i],c}^{\text{household size}} + a_{j[i],c}^{\text{race}} + a_{j[i],c}^{\text{mobility}} \\ & + a_{j[i],c}^{\text{gender:agecat}} + a_{j[i],c}^{(\text{day} + \text{day}^2 + \text{day}^3)} + \\ & + b_{j[i],c}^{(1 + \text{day} + \text{day}^2 + \text{day}^3)|\text{state}} \end{aligned} \quad (2)$$

and

$$\log(\mu_{i,c})_{\text{household}} = \alpha_{0,c} + a_{j[i],c}^{\text{gender}} + a_{j[i],c}^{\text{age category}} + a_{j[i],c}^{\text{household size}} + a_{j[i],c}^{\text{race}} + a_{j[i],c}^{\text{gender:agecat}} + b_{j[i],c}^{1|\text{state}}, \quad (3)$$

where $\alpha_{0,c}$ is the fixed intercept term and the terms $a_{j[i],c}^{\text{var}}$ denote the population-level effects (fixed effects) associated with each categorical variable. We use the subscript $j[i]$ to denote the group membership for the i th respondent. For example, the $a_{j[i],c}^{\text{race}}$ variable can take values $\{a_{\text{Black},c}^{\text{race}}, a_{\text{Hispanic},c}^{\text{race}}, a_{\text{White},c}^{\text{race}}, a_{\text{OtherRace},c}^{\text{race}}\}$ depending on the racial self-identification of respondent i . The terms $b_{j[i],c}^{\text{var}|\text{state}}$ denote the group effects (random effects or hierarchically modeled effects).

We include group effects because they allow the model to take advantage of partial pooling: when many observations are available, the estimates are closer to the original survey responses, and when few observations are available for a group, the estimates are pulled away from survey responses towards the overall mean. By assuming that the group effects for a given variable (e.g. the $b_{j[i],c}^{1|\text{state}}$ for the 51 states) are exchangeable, we can treat the variance $\sigma^{2,\text{state}}$ as a free parameter that will affect the degree of partial pooling. Thus, the amount of partial pooling—how much cell estimates are pulled towards the overall mean—is estimated from the data. In the household contact models, we model the state-specific intercepts as $b_{j[i],c}^{1|\text{state}} \sim N(0, \sigma^{2,\text{state}})$. We use a weakly informative hyper-prior for the variance parameters to avoid boundary estimates (i.e., to avoid a group-level variance estimate of 0 where such an estimate is a priori implausible [40, 41]). Specifically, we assume a half-t-distribution (truncated from below at 0) with 3 degrees of freedom for the prior of the standard deviation, $\sigma^{\text{state}}: \sigma^{\text{state}} \sim \text{half-}t_3(0, 2.5)$.

In the non-household contact models, we have state-specific cubic time trends, which require estimating a vector of four parameters per state. To benefit from partial pooling, we model these four state-specific parameters hierarchically via

$$b_{j[i],c}^{(1+\text{day}+\text{day}^2+\text{day}^3)|\text{state}} \sim N(0, V), \quad (4)$$

where the variance-covariance V is

$$\begin{aligned} V &= \text{diag}(\sigma)\Sigma\text{diag}(\sigma) \\ \Sigma &\sim \text{LKJ}(1) \\ \sigma &\sim \text{half-}t_3(0, 2.5). \end{aligned} \quad (5)$$

Here, we employ the strategy of Bürkner (2017), which separately models the marginal variances σ (a vector with four entries) and the correlation structure between the four parameters. The LKJ(1) prior implies a uniform prior distribution over correlation matrices. Similar to the household contacts model, we adopt a weakly informative half-t-distribution hyperprior for the variances. Specifically, we use a half-t-distribution with 3 degrees of freedom (truncated from below at 0) for the standard deviation.

Finally, we specify relatively uninformative priors for some of the model parameters:

$$\begin{aligned} \phi &\sim \text{Gamma}(\alpha = 0.01, \beta = 0.01) \\ \alpha_{0,c} &\sim t_3(-2.3, 2.5). \end{aligned} \quad (6)$$

All other parameters have a flat prior.

Poststratification

For our post-stratification step, we aggregate the cell-level estimates up to the state-level by weighting each cell by its relative share of the state's population. We define the target population using the Public Use Micro Data Samples from the American Community Survey 2014–2018 5-year sample [42]. We split the target population into 14,688 cells based on 5 categorical variables: gender (2 categories), state (51 categories), age group (6 categories), household size (6 categories), and Race / Hispanic origin (4 categories). Note that, under traditional poststratification, our survey would not be large enough to produce estimates for these 14,688 cells: we would have little or no data for most cells, and the resulting estimates would be extremely noisy. However, this number of cells is not a problem for MRP, which borrows information from similar cells to smooth out estimates for otherwise noisy cells.

Fig 1 illustrates our method by comparing our MRP-adjusted estimates for April 2020 with the original survey responses for two states: California, where many interviews were conducted

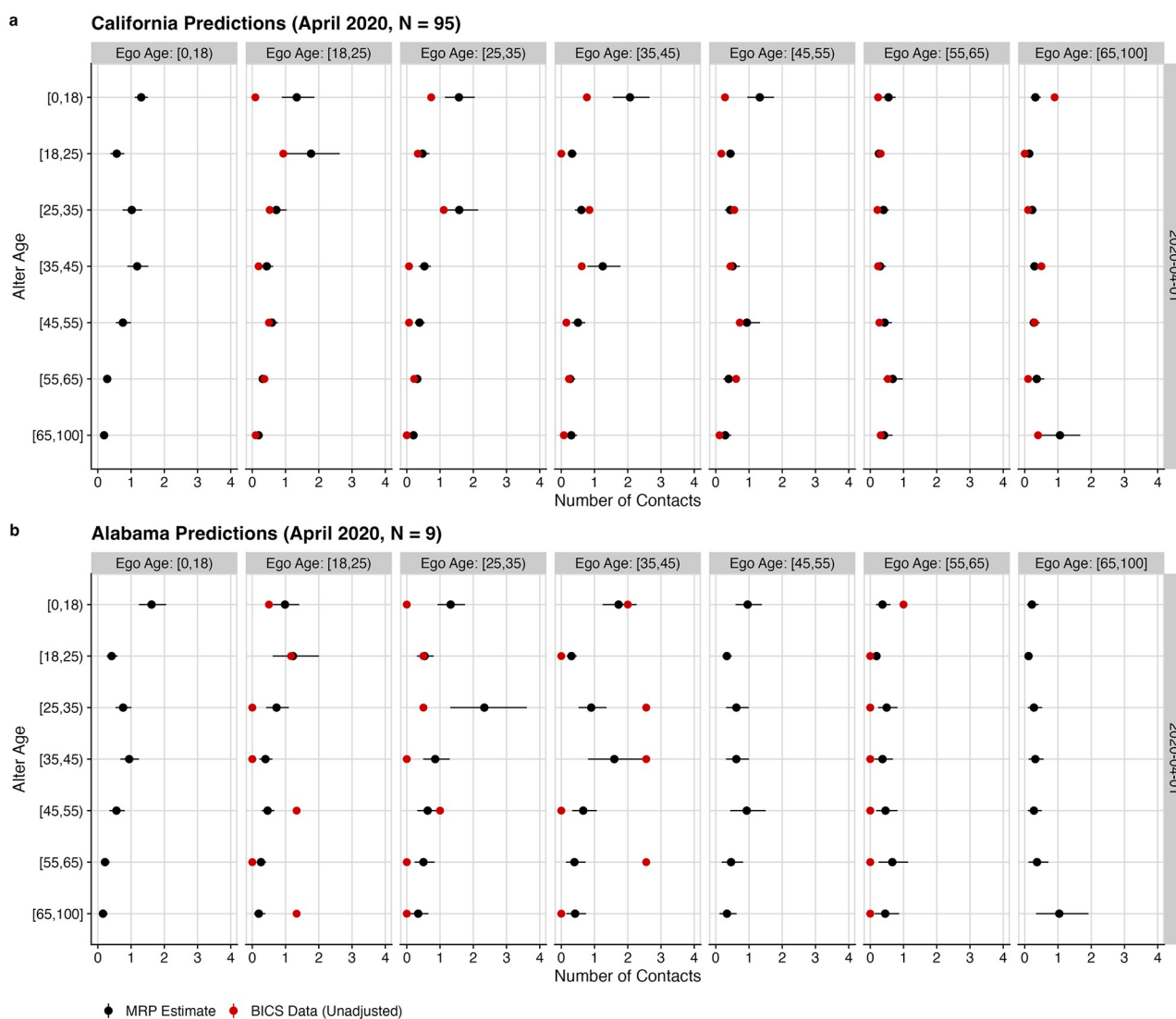


Fig 1. Comparison of MRP-adjusted estimates (black) with unweighted BICS estimates for Alabama (panel A) and California (panel B) from April of 2020. MRP estimates are presented with 95% credible intervals. Where more data are available, predictions more closely track original survey estimates.

<https://doi.org/10.1371/journal.pcbi.1010742.g001>

($N = 95$), and Alabama, where fewer interviews were conducted ($N = 9$). While the MRP estimates align closely with the original survey responses for California, they differ significantly for Alabama. This reflects the partial pooling of the hierarchical model: since Alabama has little data available, MRP estimates use information from similar cells to smooth out the noisy raw estimates produced from just 9 interviews. This partial pooling is particularly relevant for months where we collected no data. For example, no interviews were conducted in Alabama in November, but we are still able to use the dynamic MRP model to make predictions for November.

Next, we used the models to predict average age-specific contacts in each cell, and then poststratified the model predictions using the American Community Survey 2014–2018 5-year sample. The resulting estimates of state-level contact patterns for April 2020 to May 2021 were adjusted for reciprocity and arranged into contact matrices. As an illustration, Fig 2 shows the estimated contact matrices for South Dakota and Washington in July 2020. Section 3.6 in S1 Appendix shows the estimated contact matrices for all states and all months from April 2020 to May 2021.

For each estimated contact matrix $\hat{C}$, we define the relative contact intensity index to be equal to $\frac{\rho(\hat{C})}{\rho(C^{UK})}$, where $\rho(\hat{C})$ is the dominant eigenvalue of the estimated contact matrix and $\rho(C^{UK})$ is the dominant eigenvalue of the contact matrix for the UK POLYMOD study. We follow other studies in using contact patterns from UK participants in the POLYMOD study as a pre-pandemic baseline reference [9, 11]. Assuming no age-specific differences in susceptibility or infectiousness, the contact intensity index is directly proportional to the ratio of the R_0 values implied by the two contact matrices [43]. Additionally, we calculate each age group's relative contribution to overall transmission. This is proportional to the leading left eigenvector of

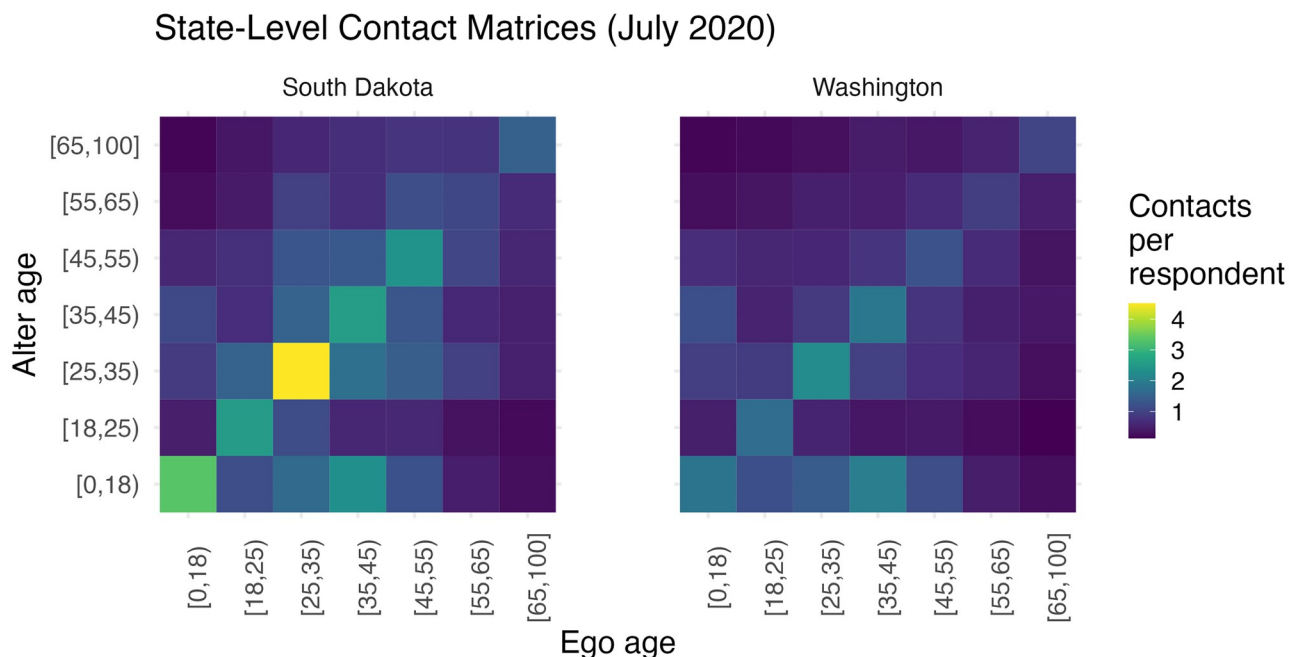


Fig 2. MRP estimates of state-level contact matrices for Washington and South Dakota in July, 2020. Each entry i, j in the contact matrix represents the average number of contacts that a person (ego) in age group i has with other people (alter) in age group j in the past 24 hours. Each cell is shaded according to the estimated mean number of total (household and non-household) contacts. The contact matrices have been adjusted for reciprocity; however, this is not necessarily reflected visually in the contact matrices due to the differential size of the population subgroups.

<https://doi.org/10.1371/journal.pcbi.1010742.g002>

the contact matrix, assuming no age-specific variation in transmissibility or susceptibility [44]. The leading left eigenvector is defined as a row vector $\mathbf{v}_L$ such that $\mathbf{v}_L \hat{C} = \rho(\hat{C}) \mathbf{v}_L$, where $\hat{C}$ is the estimated contact matrix and $\rho(\hat{C})$ is the leading eigenvalue.

Finally, to assess the degree of assortative mixing with respect to age (i.e., the tendency for people to have higher contact with others similar to them in age), we estimate the q-index. The q-index gives a measure of departure from perfectly proportionate mixing: a value of 0 indicates perfectly proportionate mixing and a value of 1 indicates a full assortativity. Specifically, the q-index is defined as:

$$\text{q-index} = \frac{|\hat{\lambda}_2|}{|\hat{\lambda}_1|} \quad (7)$$

where $\hat{\lambda}_1$ is the largest eigenvalue and $\hat{\lambda}_2$ is the second largest eigenvalue.

We use our model to estimate daily age-stratified contact matrices for all 50 US states and for Washington DC. The model combines information from (1) the Berkeley Interpersonal Contact Study (BICS) waves 1–6 ($N = 8,790$), which collected responses at the US national level from April 2020 to May 2021; and (2) locality-specific data sources, including the American Community Survey and digital estimates of population mobility from Google (“Methods”). Our model produces estimated age-stratified contact rates for each state and each day, but we focus here on monthly average estimates for each state in order to investigate changes over moderate time scales and to smooth predicted values, as recommended in the small-area estimation literature [45, 46].

Fig 2 compares predicted contact matrices for Washington and South Dakota—two states chosen to illustrate an example of a state with low and high contact, respectively—over all days in July 2020. Section 3.6 in S1 Appendix shows additional estimated contact matrices for all states and all months from April 2020 to May 2021. Two substantive findings emerge from Fig 2 and Section 3.6 in S1 Appendix. First, the levels of interpersonal contact differ significantly across states. For example, in July 2020, members of the [18–25] age group in South Dakota had on average 11.6 total contacts in the past 24 hours, while members of the [18–25] age group in Washington had only 6.4 total contacts in the past 24 hours. Second, estimated state-level contact patterns have high levels of assortative mixing by age for all states and across the course of the pandemic: at each month, the leading diagonal stripes visible in both state’s estimated contact matrix in Fig 2 show highest relative rates of contact between members of the same age group at all or nearly all ages. Members of the [0–18] age group are also estimated to have high average daily contacts with members of age groups [25–35] and [35–45] across all states and time periods, suggesting the persistent influence of generational and household structure on interpersonal interaction. These findings are consistent with previous national-level contact studies conducted in countries outside the US both before and during the COVID-19 pandemic [9, 47].

To compare the intensity of interpersonal contact across states and time, we define a quantity called the *relative contact intensity index*. The relative contact intensity index summarizes how high or low age-structured contact rates are in a specific state at a specific time, relative to a baseline. A relative contact intensity value of 1 means that contact intensity is unchanged since baseline; a value less than 1 means that contact intensity has reduced since baseline, and a value greater than 1 means that contact intensity has increased since baseline (for full mathematical details, see “Methods”).

Fig 3 shows how the relative contact intensity index for each state changes from March of 2020 to May of 2021, using UK participants in the POLYMOD study as a pre-pandemic

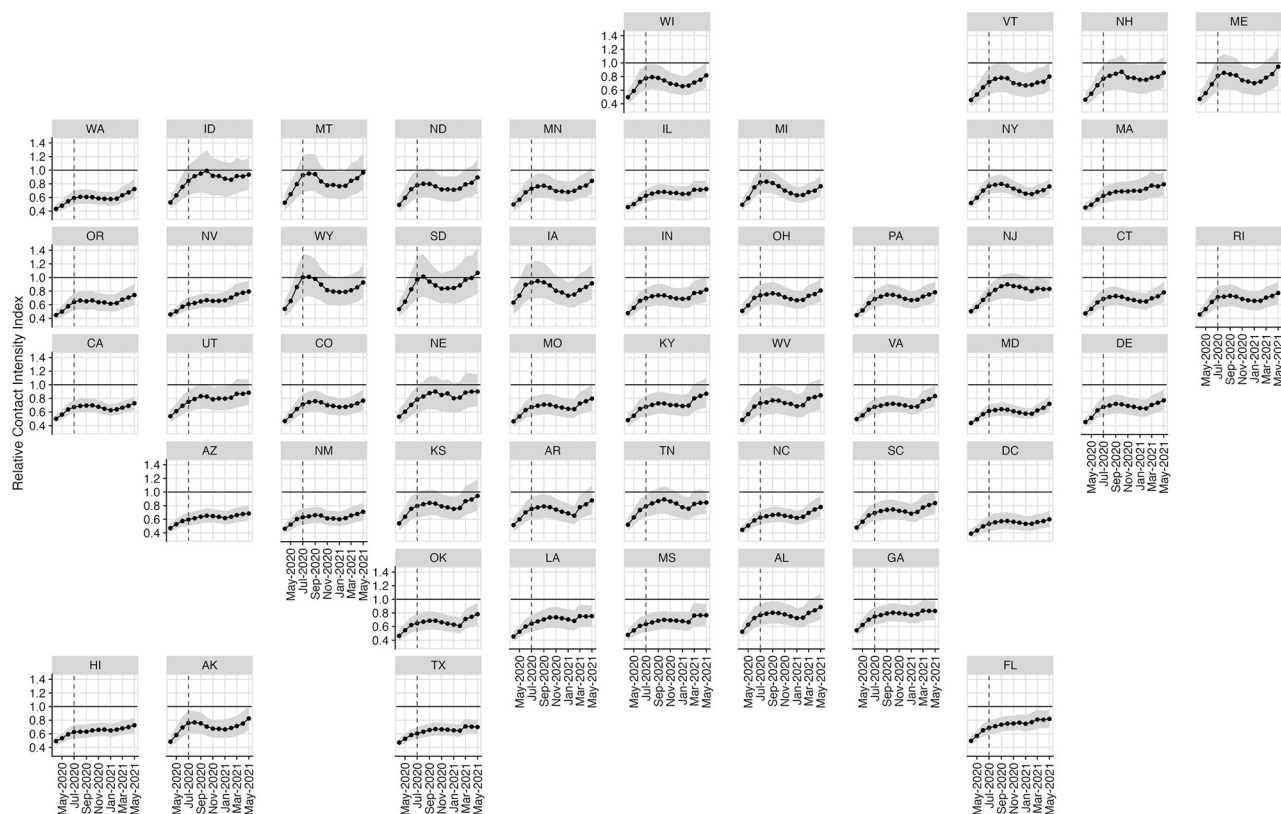


Fig 3. Monthly state-specific estimates of relative contact intensity index over time. Horizontal black line shows pre-pandemic baseline contact intensity. Vertical dashed line denotes July 2020, the month we present contact matrices for in Fig 2.

<https://doi.org/10.1371/journal.pcbi.1010742.g003>

baseline reference [9]. Consistent with the cross-state variation Fig 2 revealed for July 2020, Fig 3 shows significant heterogeneity in total levels of contact across states at all time periods. This heterogeneity suggests that national-level contact rates would be a misleading proxy for the local level: in July 2020, national contact rates would over-estimate contact intensity for Washington by 18%, while under-estimating contact intensity in South Dakota by 34%. All states experienced an upward trend in contacts over time. Relative contact intensity is lowest in March 2020 and increases until August 2020. Beginning in September 2020, some states experience a substantial decline in contact intensity (Iowa, Kansas, and New York), while other states (Florida, Hawaii, and Massachusetts) experienced a modest increase. In January 2021, contact intensity began increasing in almost every state until the end of our observation window in May 2021. The variation in contact intensity was lowest in April 2020, and peaked in August 2020 (see Fig D in S1 Appendix for full analysis).

The relative contact intensity index is a useful measure of the overall intensity of interpersonal contact. But, by summarizing the level of contact across all age groups with a single quantity, the relative contact intensity index can mask important age-specific variation over time. For example, because many diseases such as COVID-19 and influenza have strong age-specific patterns of mortality [16], age-specific contact rates are key to understanding underlying transmission dynamics. To demonstrate the implications of age-specific heterogeneity in contact rates, we create an index capturing each age group's relative contribution to overall transmission, assuming no age-specific differences in susceptibility or infectiousness [48]. Fig 4 shows the results for each age group in three large US states with different population and

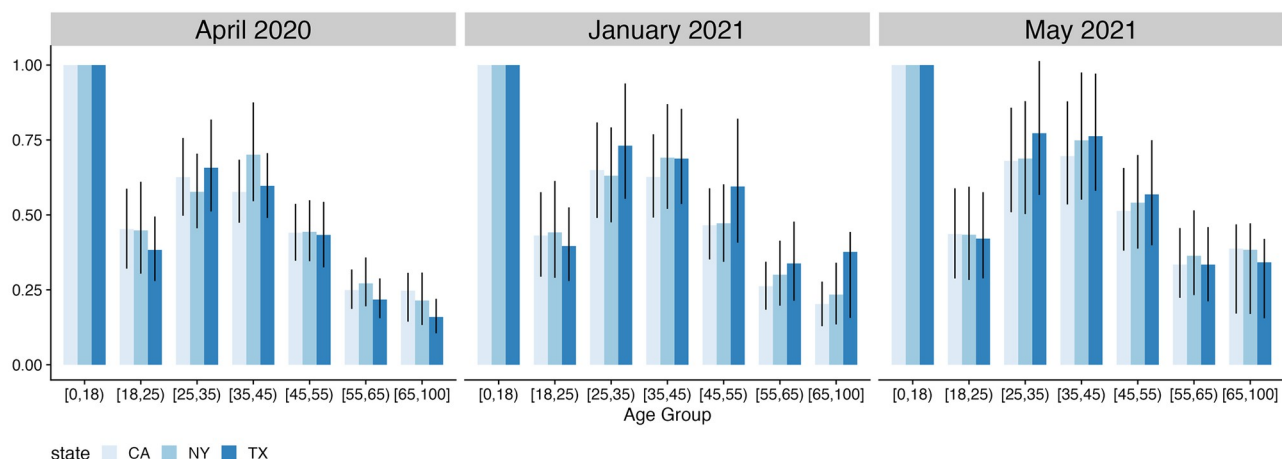


Fig 4. Relative contribution of each age-group to overall transmission in California, New York, and Texas assuming no age-specific differences in transmissibility or susceptibility. We pick the youngest age group [0–18] as a reference group which all other estimates are rescaled to. Error bars represent 80% credible intervals.

<https://doi.org/10.1371/journal.pcbi.1010742.g004>

policy dynamics: California, New York, and Texas. For interpretability, we normalize each value by dividing by the value for the [0–18] age group within each state and month. The contribution to overall transmission is highest for the [0–18] age group and lowest for the [55–65] and [65–100] age groups. These differences vary across states. For example, in January 2021, the relative contribution for the [65–100] age group was 87% larger for Texas than for California.

Finally, to understand how assortativity with respect to age changed over time, we estimate monthly, state-level estimates of the q -index. The q -index takes on values between 0 (proportionate mixing) and 1 (fully assortative mixing) [49]; see “Methods” for full mathematical definition. Fig 5 shows contacts generally became more age-assortative over time, consistent with previous findings in Wuhan, China [49]. However, trends vary across states: New Jersey’s age assortativity peaks in January 2021, with a gradual decline until May of 2021. In California, age assortativity remains low until January 2021, when there is a sharp uptick.

Previous work has found that non-pharmaceutical interventions (NPIs), such as state-level social distancing regulations and their enforcement, are associated lower aggregate-level contact intensity [25, 50–53]. As a robustness check, we refit our models adding the Oxford Stringency Score as a covariate [54]. The Oxford Stringency Score is an NPI that measures the stringency of “lock-down” policies that primarily restrict people’s social behavior. We find no significant improvement the predictive accuracy of the model (see Table B in S1 Appendix). Incorporating data on NPIs in other settings may be very useful in predicting contact rates, and we emphasize that our modeling framework makes this possible.

Discussion

In this paper, we introduce a dynamic MRP model for estimating state-level contact patterns from national-level survey data. We apply this model to national-level contact survey data collected from the Berkeley Interpersonal Contact study to produce age-stratified contact matrices for all 50 US states and Washington DC. This method for estimating subnational contact patterns can be easily adopted to other settings in countries that have recently collected national-level contact data and have available poststratification targets from a census or similar data source.

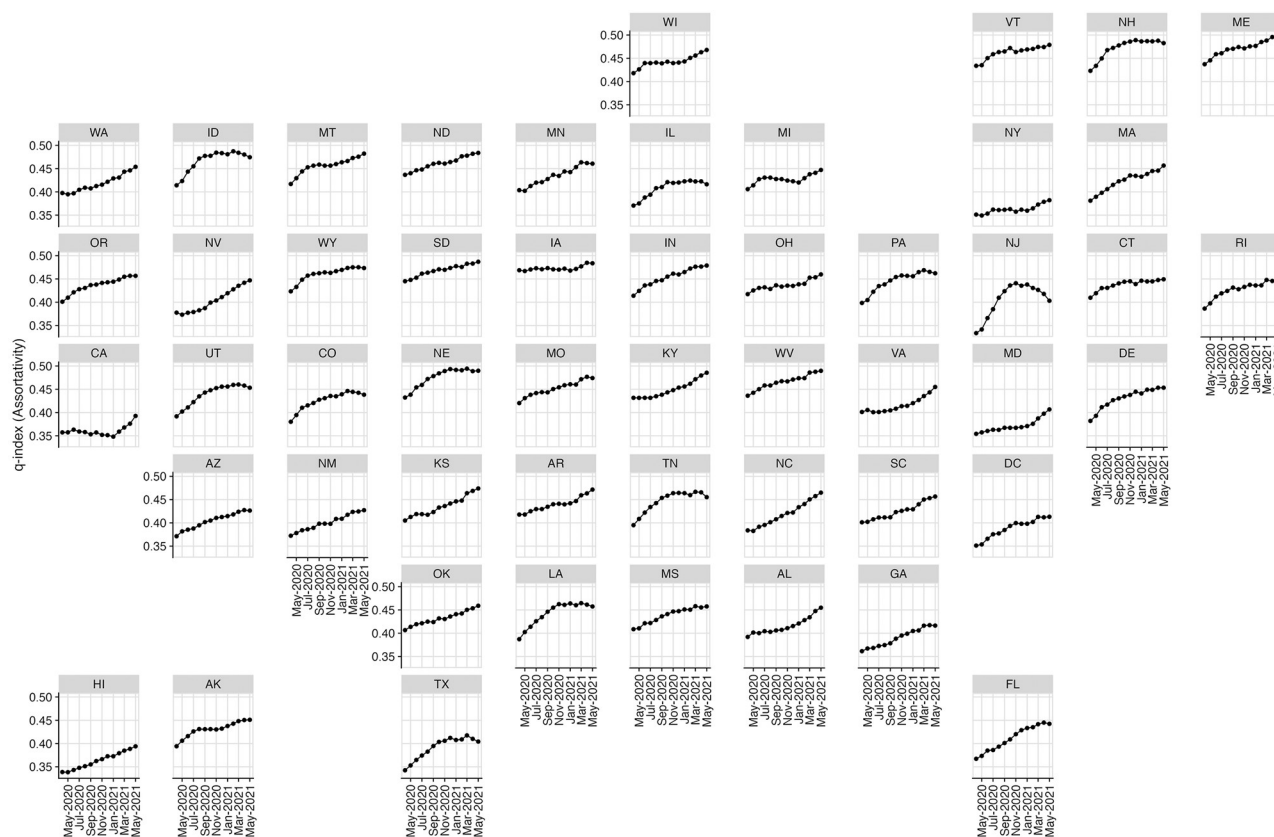


Fig 5. Change in age assortativity, as measured by the q-index, over the course of the pandemic. A higher q-index value corresponds to higher age assortativity.

<https://doi.org/10.1371/journal.pcbi.1010742.g005>

Our results reveal several key substantive insights into social contact patterns in the US during the COVID-19 pandemic. First, there is significant heterogeneity across states in overall levels of contact intensity. Second, while trends varied by state, contact intensity changed dramatically over time, with a low in April 2020, a spike in the summer of 2020 followed by a decline until January of 2021, and a gradual increase between February 2021 and May 2021. Finally, we observe a general increase in age-related contact assortativity over time, again with distinct state-level patterns of change. In sum, our findings demonstrate the crucial importance of state-level contact patterns; we find that during the COVID-19 pandemic in the US, national-level contact patterns mask important state-level variation and are not a suitable proxy for state-level estimates.

We have several methodological recommendations for other studies that may wish to produce subnational contact rate estimates from national contact data. First, we recommend researchers collect data on both household and non-household contacts and model them separately, as predictors for household and non-household contacts may differ in important ways. Second, if the contact data was collected over time, a time trend should be included in the model [15]. The cubic time-trend specified in our model may not generalize to other situations. Third, we recommend researchers fit several models with different time trends and use principled model selection techniques to pick the final model. Finally, we echo past advice in the small-area literature: model at the finest level possible and aggregate to the desired geography and time scale [45, 46].

Despite the promise of this approach, our study has important limitations and there are several directions for future work. We employ an ad-hoc approach to enforcing reciprocity in aggregate contacts between age groups (see “Methods”). Future work could explicitly incorporate these reciprocity constraints into the modeling approach. This would allow for a fully Bayesian strategy with credible intervals that may better capture the full uncertainty inherent in our model. The model could be adapted to more directly capture the age structure of reported contacts. This will involve more explicitly modeling the relationship between reported contacts in different age groups (as opposed to fitting parallel negative binomial models, as we did here). Because we have no survey respondents under the age of 18, we use the principle of reciprocity to impute the average daily number of contacts for this age group [11]; future surveys could try to interview children under 18 or obtain proxy reports from children’s parents. The reported contacts by age could be smoothed within the modeling framework, thereby reducing large jumps between neighboring cells in the total number of contacts. The BICS survey did not directly interview respondents under age 18; thus, we follow previous studies in inferring the within-group contact from all (excluding school) contacts from the UK POLYMOD (see Section 1.2 in S1 Appendix for details). As some schools in the US were open during this period, our estimate for the relative contribution the [0–18] group makes to overall transmission may be an underestimate for some states and time periods. Additionally, we assumed that all household members were contacts, but future research could relax this assumption and model trends in contact among household members over time. Perhaps most importantly, these results could be validated against other estimates from a high-quality state-level contact study, which ideally captures several states during several different time periods. These are important components of future work, which will advance our ability to produce accurate and timely small-area estimates of contact patterns from national-level contact data.

Supporting information

S1 Appendix. File containing additional detail about the data, model, and supplemental tables and figures. Fig A. The R_0 value implied by contact patterns in each state. Fig B. State-specific MRP estimate of total contact (combined household and non-household contacts). Fig C. Illustration of partial pooling. Fig D. Monthly standard deviation of contact intensity across states. Fig E. Model input and output for state of California. Fig F. State-level change in relative contact intensity index over time from a model without mobility data predictors. Fig G—Fig T. Estimated State-level Contact Matrices by month. Table A. Model selection metrics. Table B. Model performance with and without the Oxford Stringency Score. (PDF)

Acknowledgments

For helpful feedback on earlier versions of the manuscript, the authors would like to thank Sophia Rabe-Hesketh, Monica Alexander, Carl Schmertmann, and members of the Population Association of America 2021 session “Bayesian Demography.”

Author Contributions

Conceptualization: Dennis M. Feehan.

Data curation: Ayesha S. Mahmud, Dennis M. Feehan.

Formal analysis: Casey F. Breen.

Funding acquisition: Ayesha S. Mahmud, Dennis M. Feehan.

Investigation: Casey F. Breen.

Methodology: Casey F. Breen.

Project administration: Dennis M. Feehan.

Software: Casey F. Breen.

Supervision: Ayesha S. Mahmud, Dennis M. Feehan.

Visualization: Casey F. Breen.

Writing – original draft: Casey F. Breen.

Writing – review & editing: Casey F. Breen, Ayesha S. Mahmud, Dennis M. Feehan.

References

1. WHO. World Health Organization COVID-19 Dashboard; 2022.
2. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*. 2020; 323(13):1239. <https://doi.org/10.1001/jama.2020.2648> PMID: 32091533
3. Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 Pathophysiology: A Review. *Clinical Immunology*. 2020; 215:108427. <https://doi.org/10.1016/j.clim.2020.108427> PMID: 32325252
4. Rohani P, Zhong X, King AA. Contact Network Structure Explains the Changing Epidemiology of Pertussis. *Science*. 2010; 330(6006):982–985. <https://doi.org/10.1126/science.1194134> PMID: 21071671
5. Goeyvaerts N, Hens N, Ogunjimi B, Aerts M, Shkedy Z, Damme PV, et al. Estimating Infectious Disease Parameters from Data on Social Contacts and Serological Status. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2010; 59(2):255–277.
6. Wallinga J, Teunis P, Kretzschmar M. Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents. *American Journal of Epidemiology*. 2006; 164(10):936–944. <https://doi.org/10.1093/aje/kwj317> PMID: 16968863
7. Mistry D, Litvinova M, Pastore y Piontti A, Chinazzi M, Fumanelli L, Gomes MFC, et al. Inferring High-Resolution Human Mixing Patterns for Disease Modeling. *Nature Communications*. 2021; 12(1):323. <https://doi.org/10.1038/s41467-020-20544-y> PMID: 33436609
8. Liu QH, Zhang J, Peng C, Litvinova M, Huang S, Poletti P, et al. Model-Based Evaluation of Alternative Reactive Class Closure Strategies against COVID-19. *Nature Communications*. 2022; 13(1):322. <https://doi.org/10.1038/s41467-021-27939-5> PMID: 35031600
9. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*. 2008; 5(3):e74. <https://doi.org/10.1371/journal.pmed.0050074> PMID: 18366252
10. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, Edmunds WJ. Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases. *American Journal of Epidemiology*. 2008; 168(9):1082–1090. <https://doi.org/10.1093/aje/kwn220> PMID: 18801889
11. Feehan DM, Mahmud AS. Quantifying Population Contact Patterns in the United States during the COVID-19 Pandemic. *Nature Communications*. 2021; 12(1):1–9. <https://doi.org/10.1038/s41467-021-20990-2> PMID: 33563992
12. Prem K, Cook AR, Jit M. Projecting Social Contact Matrices in 152 Countries Using Contact Surveys and Demographic Data. *PLOS Computational Biology*. 2017; 13(9):e1005697. <https://doi.org/10.1371/journal.pcbi.1005697> PMID: 28898249
13. Prem K, van Zandvoort K, Klepac P, Eggo RM, Davies NG, Group CftMMoIDCW, et al. Projecting Contact Matrices in 177 Geographical Regions: An Update and Comparison with Empirical Data for the COVID-19 Era. *PLOS Computational Biology*. 2021; 17(7):e1009098. <https://doi.org/10.1371/journal.pcbi.1009098> PMID: 34310590
14. Park DK, Gelman A, Bafumi J. Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*. 2004; 12(4):375–385. <https://doi.org/10.1093/pan/mph024>
15. Gelman A, Lax J, Phillips J, Gabry J, Trangucci R. Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion. 2018; p. 48.

16. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-Specific Mortality and Immunity Patterns of SARS-CoV-2. *Nature*. 2021; 590(7844):140–145. <https://doi.org/10.1038/s41586-020-2918-0> PMID: 33137809
17. Davies NG, Klepac P, Liu Y, Prem K, Jit M, CMMID COVID-19 working group, et al. Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics. *Nature Medicine*. 2020; 26(8):1205–1211. <https://doi.org/10.1038/s41591-020-0962-9> PMID: 32546824
18. Dorélien AM, Ramen A, Swanson I, Hill R. Analyzing the Demographic, Spatial, and Temporal Factors Influencing Social Contact Patterns in U.S. and Implications for Infectious Disease Spread. *BMC Infectious Diseases*. 2021; 21(1):1009. <https://doi.org/10.1186/s12879-021-06610-w> PMID: 34579645
19. Iozzi F, Trusiano F, Chinazzi M, Billari FC, Zagheni E, Merler S, et al. Little Italy: An Agent-Based Approach to the Estimation of Contact Patterns- Fitting Predicted Matrices to Serological Data. *PLoS Computational Biology*. 2010; 6(12):e1001021. <https://doi.org/10.1371/journal.pcbi.1001021> PMID: 21152004
20. Barrat A, Cattuto C, Tozzi AE, Vanhems P, Voirin N. Measuring Contact Patterns with Wearable Sensors: Methods, Data Characteristics and Applications to Data-Driven Simulations of Infectious Diseases. *Clinical Microbiology and Infection*. 2014; 20(1):10–16. <https://doi.org/10.1111/1469-0691.12472> PMID: 24267942
21. Mastrandrea R, Fournet J, Barrat A. Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. *PLOS ONE*. 2015; 10(9):e0136497. <https://doi.org/10.1371/journal.pone.0136497> PMID: 26325289
22. Smieszek T, Burri EU, Scherzinger R, Scholz RW. Collecting Close-Contact Social Mixing Data with Contact Diaries: Reporting Errors and Biases. *Epidemiology and Infection*. 2012; 140(4):744–752. <https://doi.org/10.1017/S0950268811001130> PMID: 21733249
23. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A Systematic Review of Social Contact Surveys to Inform Transmission Models of Close-contact Infections. *Epidemiology (Cambridge, Mass)*. 2019; 30(5):723–736. <https://doi.org/10.1097/EDE.0000000000001047> PMID: 31274572
24. Mousa A, Winskill P, Watson OJ, Ratmann O, Monod M, Ajelli M, et al. Social Contact Patterns and Implications for Infectious Disease Transmission—a Systematic Review and Meta-Analysis of Contact Surveys. *eLife*. 2021; 10:e70294. <https://doi.org/10.7554/eLife.70294> PMID: 34821551
25. Liu CY, Berlin J, Kiti MC, Del Fava E, Grow A, Zagheni E, et al. Rapid Review of Social Contact Patterns During the COVID-19 Pandemic. *Epidemiology*. 2021; 32(6):781–791. <https://doi.org/10.1097/EDE.0000000000001412> PMID: 34392254
26. Efron B, Morris C. Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*. 1975; 70(350):311–319. <https://doi.org/10.1080/01621459.1975.10479864>
27. Fay RE, Herriot RA. Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*. 1979; 74(366a):269–277. <https://doi.org/10.1080/01621459.1979.10482505>
28. Schmertmann CP, Cavenaghi SM, Assunção RM, Potter JE. Bayes plus Brass: Estimating Total Fertility for Many Small Areas from Sparse Census Data. *Population Studies*. 2013; 67(3):255–273. <https://doi.org/10.1080/00324728.2013.795602> PMID: 24143946
29. Assunção RM, Schmertmann CP, Potter JE, Cavenaghi SM. Empirical Bayes Estimation of Demographic Schedules for Small Areas. *Demography*. 2005; 42(3):537–558. <https://doi.org/10.1353/dem.2005.0022> PMID: 16235612
30. Alexander M, Zagheni E, Barbieri M. A Flexible Bayesian Model for Estimating Subnational Mortality. *Demography*. 2017; 54(6):2025–2041. <https://doi.org/10.1007/s13524-017-0618-7> PMID: 29019084
31. Rao JNK, Molina I. Small Area Estimation: Rao/Small Area Estimation. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2015.
32. Gelman A, Little TC. Poststratification Into Many Categories Using Hierarchical Logistic Regression. 1997;.
33. Lax JR, Phillips JH. How Should We Estimate Sub-National Opinion Using MRP? Preliminary Findings and Recommendations. 2013; p. 39.
34. Wang W, Rothschild D, Goel S, Gelman A. Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting*. 2015; 31(3):980–991. <https://doi.org/10.1016/j.ijforecast.2014.06.001>
35. Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*. 2014; 179(8):1025–1033. <https://doi.org/10.1093/aje/kwu018> PMID: 24598867

36. Vehtari A, Gelman A, Gabry J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing*. 2017; 27(5):1413–1432. <https://doi.org/10.1007/s11222-016-9709-3>
37. Google L. COVID-19 Community Mobility Reports; 2021. <https://www.google.com/covid19/mobility/>.
38. Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al. Reduction in Mobility and COVID-19 Transmission. *Nature Communications*. 2021; 12(1):1090. <https://doi.org/10.1038/s41467-021-21358-2> PMID: 33597546
39. Jewell S, Futoma J, Hannah L, Miller AC, Foti NJ, Fox EB. It's Complicated: Characterizing the Time-Varying Relationship between Cell Phone Mobility and COVID-19 Spread in the US; 2021.
40. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika*. 2013; 78(4):685–709. <https://doi.org/10.1007/s11336-013-9328-2> PMID: 24092484
41. Gelman A. Prior Distributions for Variance Parameters in Hierarchical Models(Comment on Article by Browne and Draper). *Bayesian Analysis*. 2006; p. 20.
42. Ruggles S, Flood S, Goeken R, Grover J, Meyer E, Pacas J, et al. IPUMS USA: Version 10.0 [Dataset]. Minneapolis, MN: IPUMS <https://doi.org/10.18128/D010.V10.0>. 2020;.
43. Diekmann O, Heesterbeek JAP, Metz JAJ. On the Definition and the Computation of the Basic Reproduction Ratio R_0 in Models for Infectious Diseases in Heterogeneous Populations. *Journal of Mathematical Biology*. 1990; 28(4). <https://doi.org/10.1007/BF00178324> PMID: 2117040
44. Hurford A, Cownden D, Day T. Next-Generation Tools for Evolutionary Invasion Analyses. *Journal of The Royal Society Interface*. 2010; 7(45):561–571. <https://doi.org/10.1098/rsif.2009.0448> PMID: 19955121
45. Mills CW, Johnson G, Huang TTK, Balk D, Wyka K. Use of Small-Area Estimates to Describe County-Level Geographic Variation in Prevalence of Extreme Obesity Among US Adults. *JAMA Network Open*. 2020; 3(5):e204289. <https://doi.org/10.1001/jamanetworkopen.2020.4289> PMID: 32383746
46. Downes M, Gurrin LC, English DR, Pirkis J, Currier D, Spittal MJ, et al. Multilevel Regression and Post-stratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples. *American Journal of Epidemiology*. 2018; 187(8):1780–1790. <https://doi.org/10.1093/aje/kwy070> PMID: 29635276
47. Verelst F, Hermans L, Vercruysse S, Gimma A, Coletti P, Backer JA, et al. SOCRATES-CoMix: A Platform for Timely and Open-Source Contact Mixing Data during and in between COVID-19 Surges and Interventions in over 20 European Countries. *BMC Medicine*. 2021; 19(1):254. <https://doi.org/10.1186/s12916-021-02133-y> PMID: 34583683
48. Klepac P, Kucharski AJ, Conlan AJ, Kissler S, Tang ML, Fry H, et al. Contacts in Context: Large-Scale Setting-Specific Social Mixing Matrices from the BBC Pandemic Project. *Epidemiology*; 2020.
49. Zhang J, Litvinova M, Liang Y, Wang Y, Wang W, Zhao S, et al. Changes in Contact Patterns Shape the Dynamics of the COVID-19 Outbreak in China. *Science*. 2020; 368(6498):1481–1486. <https://doi.org/10.1126/science.abb8001> PMID: 32350060
50. Coletti P, Wambua J, Gimma A, Willem L, Vercruysse S, Vanhoutte B, et al. CoMix: Comparing Mixing Patterns in the Belgian Population during and after Lockdown. *Scientific Reports*. 2020; 10(1):21885. <https://doi.org/10.1038/s41598-020-78540-7> PMID: 33318521
51. Gimma A, Munday JD, Wong KLM, Coletti P, van Zandvoort K, Prem K, et al. Changes in Social Contacts in England during the COVID-19 Pandemic between March 2020 and March 2021 as Measured by the CoMix Survey: A Repeated Cross-Sectional Study. *PLOS Medicine*. 2022; 19(3):e1003907. <https://doi.org/10.1371/journal.pmed.1003907> PMID: 35231023
52. Latsuzbaia A, Herold M, Bertemes JP, Mossong J. Evolving Social Contact Patterns during the COVID-19 Crisis in Luxembourg. *PLOS ONE*. 2020; 15(8):e0237128. <https://doi.org/10.1371/journal.pone.0237128> PMID: 32760114
53. Tizzani M, De Gaetano A, Jarvis CI, Gimma A, Wong K, Edmunds WJ, et al. Impact of Tiered Measures on Social Contact and Mixing Patterns in Italy during the Second Wave of COVID-19. In Review; 2022.
54. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, et al. A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*. 2021; 5(4):529–538. <https://doi.org/10.1038/s41562-021-01079-8> PMID: 33686204

Supporting Information

Novel Estimates Reveal Subnational Heterogeneities in Disease-Relevant
Contact Patterns in the United States

Casey F. Breen ^{*} Ayesha S. Mahmud [†] Dennis M. Feehan [‡]

August 23, 2025

^{*}Corresponding author. Department of Demography, University of California, Berkeley. casey-breen@berkeley.edu.

[†]Department of Demography, University of California, Berkeley.

[‡]Department of Demography, University of California, Berkeley.

Contents

1	BICS Contact Survey	S3
1.1	Contact-level Weights	S3
1.2	Youngest Age Group	S4
2	Methodological Details	S4
2.1	Reciprocity Constraints	S4
2.2	Poststratification	S5
2.3	Model Selection	S6
2.4	Summarizing contact intensity	S7
2.5	Alternative Contact Intensity Measures	S8
3	Additional Analyses	S12
3.1	Internal validation check	S12
3.2	Variation in Contact Intensity	S12
3.3	Mobility data illustration for CA	S14
3.4	Model without mobility predictors	S14
3.5	Model Code	S17
3.6	Contact Matrices by month	S18

1 BICS Contact Survey

1.1 Contact-level Weights

Respondents were asked to report the age and sex of all household contact members and to provide detailed information on up to three non-household contacts, including contact age, sex, contact location, and their relation to the contact. In all waves, some respondents reported more than three non-household contacts, but these respondents were prompted to report only on the first three contacts that came to mind. To account for this, within-respondent weights were constructed to provide detailed information on all contacts. Each of the three initial contacts receives the following weight a_i :

$$a_i = \frac{\text{Total number of contacts}}{\text{Total number of contacts for which detailed data was collected}}. \quad (1)$$

To illustrate, if respondent i reports 12 contacts and provides detailed information on 3 contacts, each of the 3 original contacts are assigned a weight $a_i = 4$. Intuitively, the within-respondent weights represent the total number of contacts represented by each of the three contacts who were reported about in detail. This approach assumes that the first three contacts who come to mind are a randomly selected sample of all contacts. For a more detailed explanation, see Feehan and Cobb [8]. We round each respondent's weighted number of contacts in each age category to the closest integer using stochastic rounding:

$$\text{Round}(\text{contacts}) = \begin{cases} \lfloor \text{contacts} \rfloor, & \text{with probability } 1 - (\text{contacts} - \lfloor \text{contacts} \rfloor) \\ \lfloor \text{contacts} \rfloor + 1, & \text{with probability } (\text{contacts} - \lfloor \text{contacts} \rfloor) \end{cases}$$

For example, if the weighted number of alters in a certain age category was 5.3 alters, it would be rounded to 5 with probability 0.7 and to 6 with probability 0.3. Stochastic rounding preserves the expected value of the number of weighted contacts, which might not be the case for conventional rounding.

1.2 Youngest Age Group

For the youngest age group $[0 - 18)$, for which there are no survey respondents, we use the principle of reciprocity to impute the average daily number of connections [7]. Specifically, we assume that the number of contacts for the youngest age group is:

$$\hat{c}_{1j} = \frac{c_{j1}N_j}{N_1} \quad (2)$$

where c_{ij} is the unadjusted, average daily contacts respondents in age group i reported to their contacts in age group j , and N_j is the population size of group j . Intuitively, we are using reciprocity principles to infer the number of connections from the $[0 - 18)$ group to other groups using reports from other groups to the $[0 - 18)$ group. However, we cannot use this method to infer the average daily number of contacts the $[0 - 18)$ group has to other members of the $[0 - 18)$ group. To do this, we calculate the leading eigenvalue of each of our contact matrices and compare it to the leading eigenvalue of the contact matrix from the UK POLYMOD. We then scale the number of contacts in the UK POLYMOD $[0 - 18)$ by the ratio of the leading eigenvalues; we use this imputed value for the average daily contacts from members of the $[0 - 18)$ group to contacts in the $[0 - 18)$ group.

While we must make several assumptions to estimate contact matrices for the (0-18] age class, these assumptions do not affect our estimates of contact patterns among adults.

2 Methodological Details

2.1 Reciprocity Constraints

We aggregated our poststratified estimates of contact rates by state and time into crude contact matrices and then adjusted these crude contact matrices to enforce reciprocity, an important feature of contact matrices. To simply motivate why contact matrices should be reciprocal: if person A had conversational contact with person B, person B must have also had conversational contact with person A. Therefore, a contact matrix should have the same total number of contacts from age-group i to age-group j as from age group j to age group

i. In practice, estimated contact patterns from survey data (and poststratified estimates) are not necessarily reciprocal. To ensure reciprocity [13], we use the following formula to estimate adjusted mean contacts $\hat{c}_{ij}$, the average number of contacts that a member of group *i* has with people in group *j*:

$$\hat{c}_{ij} = \frac{1}{2N_i} \times (c_{ij}N_i + c_{ji}N_j) \quad (3)$$

where

- N_i = size of age-group *i* from ACS
- N_j = size of age-group *j* from ACS
- c_{ij} = estimated average number of contacts that a member of group *i* has with people in group *j*
- c_{ji} = estimated average number of contacts that a member of group *j* has with people in group *i*

2.2 Poststratification

Survey-based estimates are commonly adjusted using poststratification, a technique that incorporates auxiliary information about the population of interest to improve estimates from a sample. The traditional poststratification approach first assigns respondents to mutually exclusive “cells” – groups cross-classified by key covariates. For example, one cell might be women aged 25-35 living in New York. Next, the cell-level estimates are generated by taking the mean sample response within each cell. Finally, the cell-level estimates are aggregated by weighting each cell relative to its proportion in the population.

This classic poststratification approach works well when the assumptions it relies upon are satisfied. In particular, classic poststratification assumes that each person within a cell has the same chance of being sampled. This assumption is generally violated in practice unless cells have been very finely partitioned. However, once cells have been finely partitioned, the number of interviews per cell typically becomes very small, resulting in unstable cell-level averages. Thus, under classic poststratification there is a tension between how finely cells are partitioned and how stable cell-level average estimates are.

	Mobility	Time Trend	ELPD (Diff)	SE Diff
Model 3	Yes	national cubic + state-specific cubic random effect	0.0	0.0
Model 2	No	national cubic + state-specific cubic random effect	-1.2	6.89
Model 1	No	state-specific cubic random effect	-60.9	17.9

Table A: Model selection metrics. A larger ELPD denotes a more predictive model. This table shows that Model 3 is the best model, as Model 3 has the lowest ELPD.

Hierarchical models like the one proposed here have the potential to help resolve this tension. The idea is to overcome the instability of estimates from very finely partitioned cells by estimating cell-level averages from a hierarchical model that allows for “random” or “modeled” effects for some predictors. Random effects allow for partial pooling towards the group mean, with greater pooling for less-populated cells [6]: when little to no data are available for a cell, estimates are based on data from similar cells; when lots of data are available for a cell, estimates are closer to the original survey responses.

2.3 Model Selection

We fit a series of models with different levels of complexity to identify a model that is as flexible as possible without overfitting. To perform model selection, we use Pareto-smoothed importance sampling leave-one-out cross validation (PSIS LOO-CV) [14]. Conventional LOO-CV requires omitting a single observation and refitting the model. However, for complex models or larger datasets, this can be computationally expensive (e.g., for our model, LOO-CV would require repeating estimation 8,920 times). PSIS LOO-CV is a computationally efficient method for estimating the pointwise out-of-sample predictive accuracy of a model using the Pareto-smoothed importance sampling. PSIS LOO-CV calculates importance weights which come from evaluating the ratio of a set of draws from the full posterior distribution to a set of draws from the leave-one-out posterior distribution. PSIS LOO-CV then fits a Pareto distribution to the upper tail of importance weights to smooth out extreme values. The results are then combined into a single measure, the expected log pointwise predictive density (ELPD). The ELPD is a summary of out-of-sample predictive accuracy using posterior simulations, and can be used both to assess the (out-of-sample) predictive accuracy

Model	ELPD (Diff)	SE Diff
Model 3	0.0	0.0
Model 3 + Oxford Stringency Index	-8.4	6.7

Table B: Incorporating the Oxford Stringency Score into our best-performing model (model 3) decreased our models performance, as measured by the expected log predictive density (ELPD). A larger ELPD denotes a more predictive model.

of the model and for model comparison. [Table A](#) summarizes the ELPD predictive density for four different model fits. Based on these results, we chose Model 3 for our analysis: it had the best performance as measured by its LOO-CV ELPD.

Fits shown here were obtained using the `brms` package [1]. Models were run with 1,000 warm-up iterations and 1,000 sampling iterations. Each model was run on four different sampling chains; the R-Hat statistics were between 1 and 1.01, suggesting that the chains had mixed successfully.

[Table B](#) shows that incorporating the Oxford Stringency Score [9]—an NPI that measures the stringency of “lock-down” policies that primarily restrict people’s social behavior—as a covariate into our model led to a slight decrease in model performance. Understanding why incorporating NPIs into our model did not improve its predictive accuracy is a potential avenue for future research.

2.4 Summarizing contact intensity

While a contact matrix contains valuable detail about age-specific transmission rates, a single summary measure of total level of social contact can facilitate comparisons across time and place. We use the leading eigenvalue of the contact matrix C as a summary measure of the overall levels of contact implied by a contact matrix.

To motivate this choice, we briefly review the link between the contact matrix C and the next generation matrix N , which is useful in modeling the early stages of a disease outbreak; see Diekmann, Heesterbeek and Metz [5] and Diekmann, Heesterbeek and Britton [4] for a full account. The next generation matrix $N = (n_{ij})$ is an $A \times A$ matrix that maps a vector with counts of infected individuals in each age group at generation g , ϕ_g , to a vector of counts of infected individuals in each age group in the subsequent generation, $\phi_{g+1} = N\phi_g$.

The dominant eigenvalue of the next generation matrix, $\rho(N)$, is the basic reproductive number, R_0 [5]. R_0 is a critical quantity in infectious disease epidemiology: it summarizes the number of secondary infections expected to result from one infected individual in an entirely susceptible population.

The social contact hypothesis says that the next generation matrix is proportional to the contact matrix, i.e., $N = \alpha C$ where $\alpha > 0$ is a proportionality constant that depends on the pathogen and other factors, assuming that there is no age-specific heterogeneity in susceptibility and transmissibility [16, 10, 4]. Under this hypothesis, the dominant eigenvalue of the social contact matrix is $\rho(C) = \rho(\alpha N) = \alpha R_0$. Thus, the dominant eigenvalue of a social contact matrix can be a useful summary of the aggregate level of contact.

The social contact hypothesis can also help compare contact matrices across place or time. Given two contact matrices C_1 and C_2 , there will be R_0 values associated with each one, say R_0^1 and R_0^2 . Assume that the pathogen of interest is the same and that factors other than contact are held constant, so that the proportionality constant α is the same for both matrices. Then under the social contact hypothesis, $\frac{\rho(C_1)}{\rho(C_2)} = \frac{\alpha R_0^1}{\alpha R_0^2} = \frac{R_0^1}{R_0^2}$. Thus, if the dominant eigenvalue of C_2 is twice the dominant eigenvalue of C_1 , then the R_0 value associated with C_2 is twice as big as the R_0 value associated with C_1 . Under the social contact hypothesis, this ratio can be identified even if the proportionality constant α is not known.

2.5 Alternative Contact Intensity Measures

In the main body of our paper, we quantify contact rates using the relative contact intensity. Here, we present two alternative measures of contact intensity: (1) the average number of total daily contacts and (2) a measure of R_0 for a respiratory pathogen such as COVID-19 (with age-specific differences in the probability of transmission given contact with an infectious individual). We calculate R_0 using a deterministic, age-stratified compartmental model that has been used previously to study COVID-19 dynamics [3]. The model assumes that all people are initially susceptible, and then can become exposed through a contact with an infected person. After exposure, there is a latent period before individuals become infected either clinically or sub-clinically. Sub-clinical infections are assumed to be less infectious compared with clinical infections. After the infectious state, individuals then

move to the recovered state. To calculate R_0 , we take the absolute value of the leading eigenvalue of the next-generation matrix (NGM), which is defined as [3]:

$$NGM = D_u C D_y D_x \quad (4)$$

where

- D_u is a diagonal matrix with elements u_i , the probability of transmission given contact with an infected person for an individual in age-class i (i.e. the relative susceptibility of individuals in age-class i)
- C is the state-specific age-structured contact matrix
- D_y is a diagonal matrix with elements y corresponding to the mean infectious period
- D_x is a diagonal matrix with elements $\rho_j + (1 - \alpha)\rho_j$; ρ_j is the probability of having clinical symptoms in age-class j and α is the relative infectiousness of sub-clinical cases

The probability of transmission given contact with an infected person in a given age-class (u_i) and the probability of having clinical symptoms per age group (ρ_j) were drawn from the literature [3]. We assume a relative infectiousness of clinical vs. subclinical individuals (α) of 0.5 [12] and a mean infectious period of 6 days [11]. We re-scale these u_i values such that a model fit to a baseline contact matrix from UK participants in the POLYMOD study [13, 7, 2] produces the R_0 value of 2.5 [15]. We then fit this model with the rescaled u_i values on all estimated contact matrix; estimates are shown in Figure A.

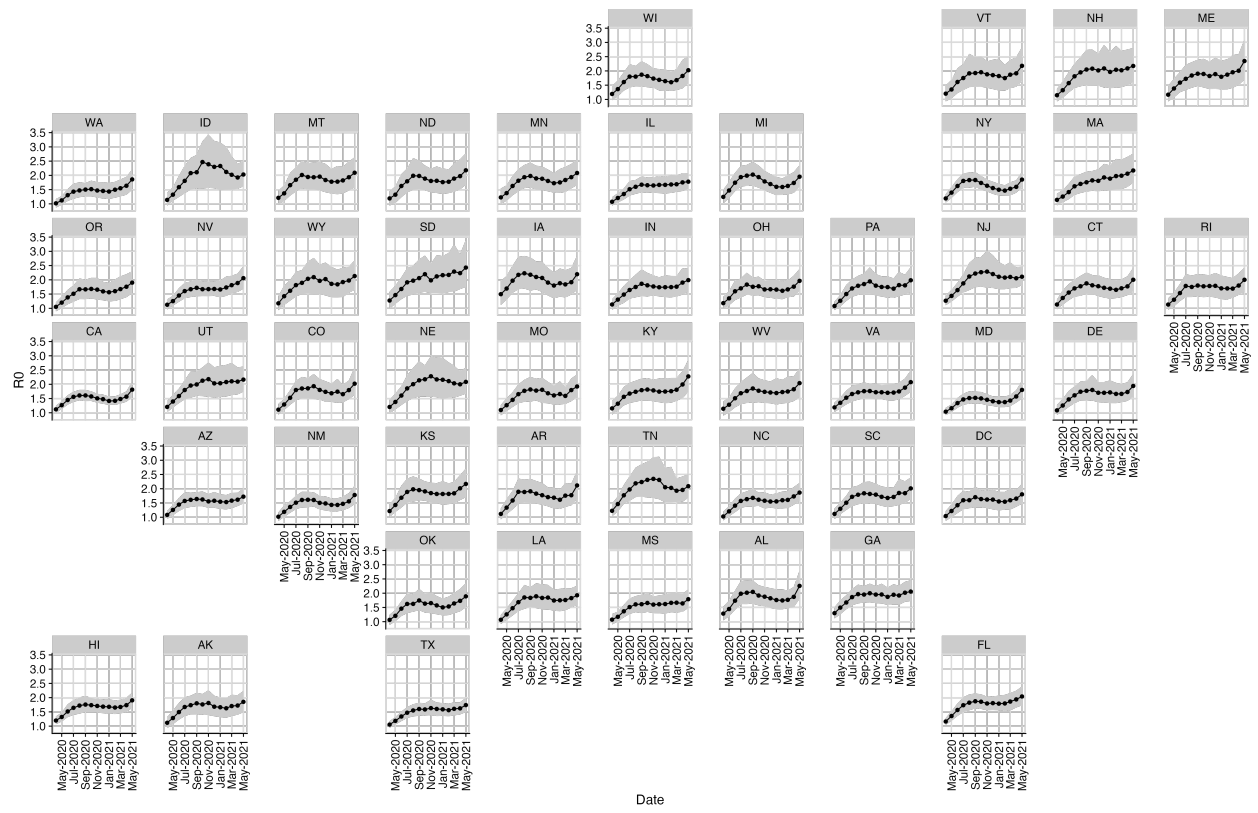


Figure A: The R_0 value implied by contact patterns in each state. R_0 was estimated using an age-stratified deterministic compartmental model.

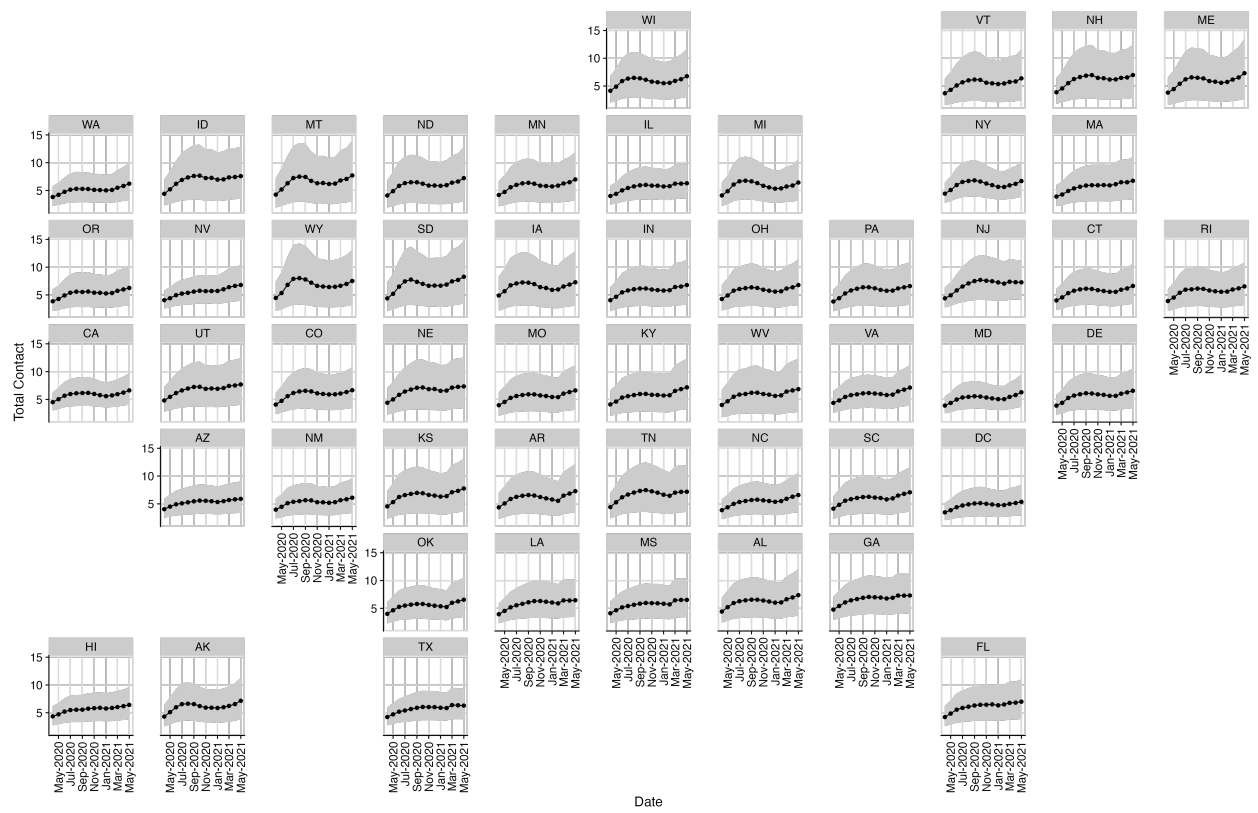


Figure B: State-specific MRP estimate of total contact (combined household and non-household contacts).

3 Additional Analyses

3.1 Internal validation check

To assess the internal validity of our method, we re-estimate our models on two subsamples. These subsamples are intended to illustrate the effects of having less data in two states: Florida and Washington. We chose these two states to illustrate partial pooling because they deviated from the overall mean in distinct ways. Florida deviates from the overall mean because there is no noticeable decline in contacts over the period between November 2020 and March 2021. Washington broadly shares similar trends over time to the overall mean, but has a lower contact rate. For the first subsample, we randomly drop all but 20% of the responses for each wave from Florida and Washington; we do not drop any data from other states. For the second subsample, we randomly drop all but 60% of the responses for each wave from Florida and Washington; again, we do not drop any data from other states.

[Figure C](#) plots the the estimated contact intensity from models fit to these two subsamples and to the full sample. The figure demonstrates how partial pooling works: when there is less data available (60% subsample), the estimates are pulled towards the overall mean. When more data is available (20% subsample), the estimates more closely resemble the raw survey data. This is clearly illustrated in Florida and Washington, as estimates from 60% subsample and full sample are pulled away from the overall mean. Additionally, this provides some internal validation check for this method, as the estimates are fairly reasonable even when there is little data.

3.2 Variation in Contact Intensity

[Figure D](#) shows the standard deviation of the relative contact intensity index across states by month. This measure of how variable contact intensity was across states reveals several interesting patterns. The state-level similarity in contact intensity changed dramatically over the course of the COVID-19 pandemic. Contact intensity was most similar across states during April 2020, when social contact was lowest. When contact intensity spiked during July and August of 2020, state-to-state variation in contact intensity was highest.

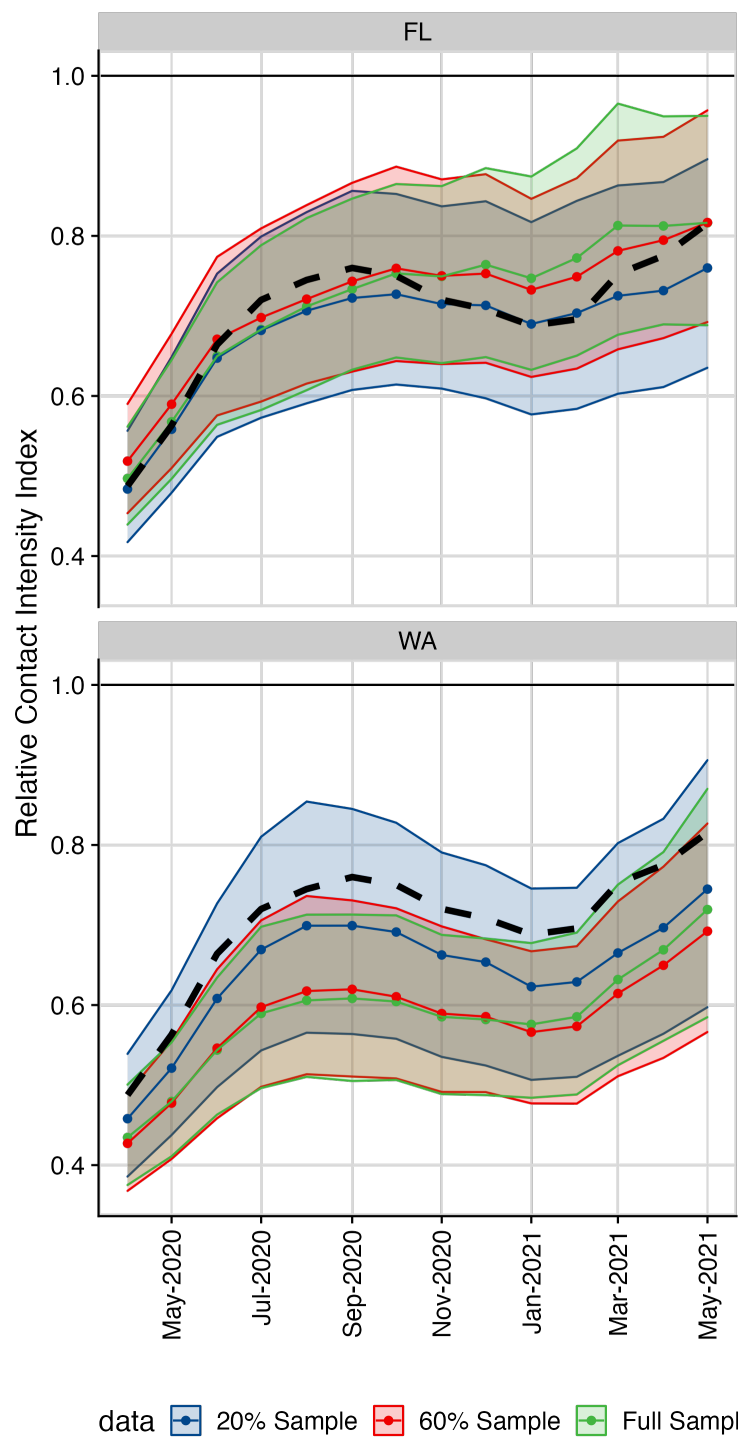


Figure C: Illustration of partial pooling. The black dashed line shows the average estimated relative contact intensity across all states.

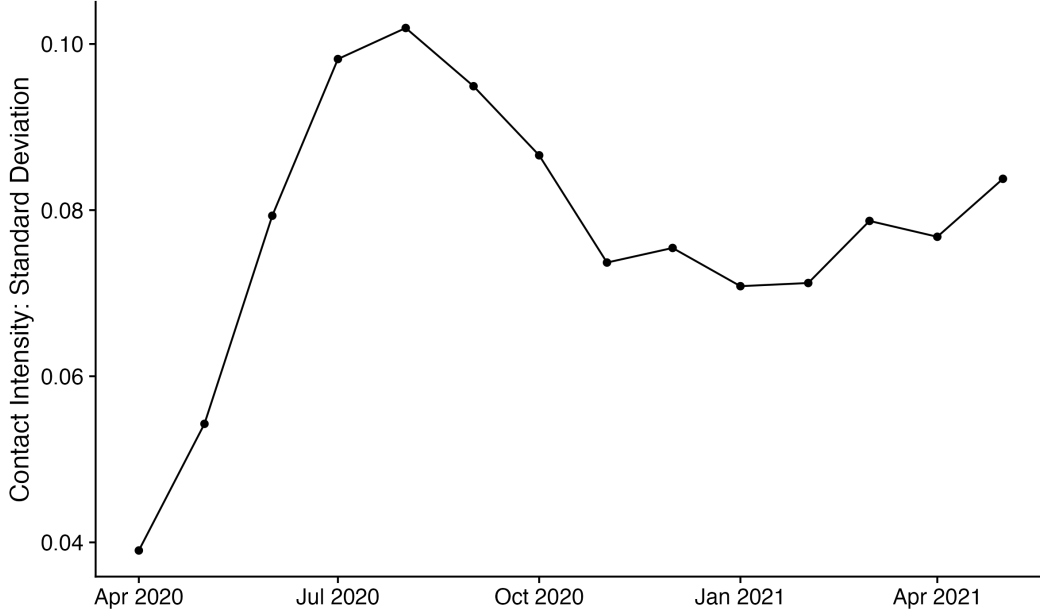


Figure D: Monthly standard deviation of contact intensity across states.

3.3 Mobility data illustration for CA

Figure E shows model input and predictions for California to give intuition about the relationship between model input and ultimate predictions of contact intensity. Specifically, panel (a) shows the first principal component of the Google mobility data and panel (b) shows the rolling three-week average of first principal component of mobility data. We used the rolling-three week average of the first principal component to smooth the data. A higher value of the first principal component indicates higher levels of mobility. Panel (c) shows the average monthly reported number of contacts in California. Panel (d) shows the implied R_0 value and panel (e) shows daily predictions of the relative contact intensity index. The cubic trend – observed in both BICS unadjusted survey data and mobility data – is also reflected in both the predicted relative contact intensity index and the monthly estimate of implied R_0 .

3.4 Model without mobility predictors

While we find mobility data is only marginally predictive (see Table A), we include it in our model to highlight the model’s flexible ability to incorporate state-level contextual predictors.

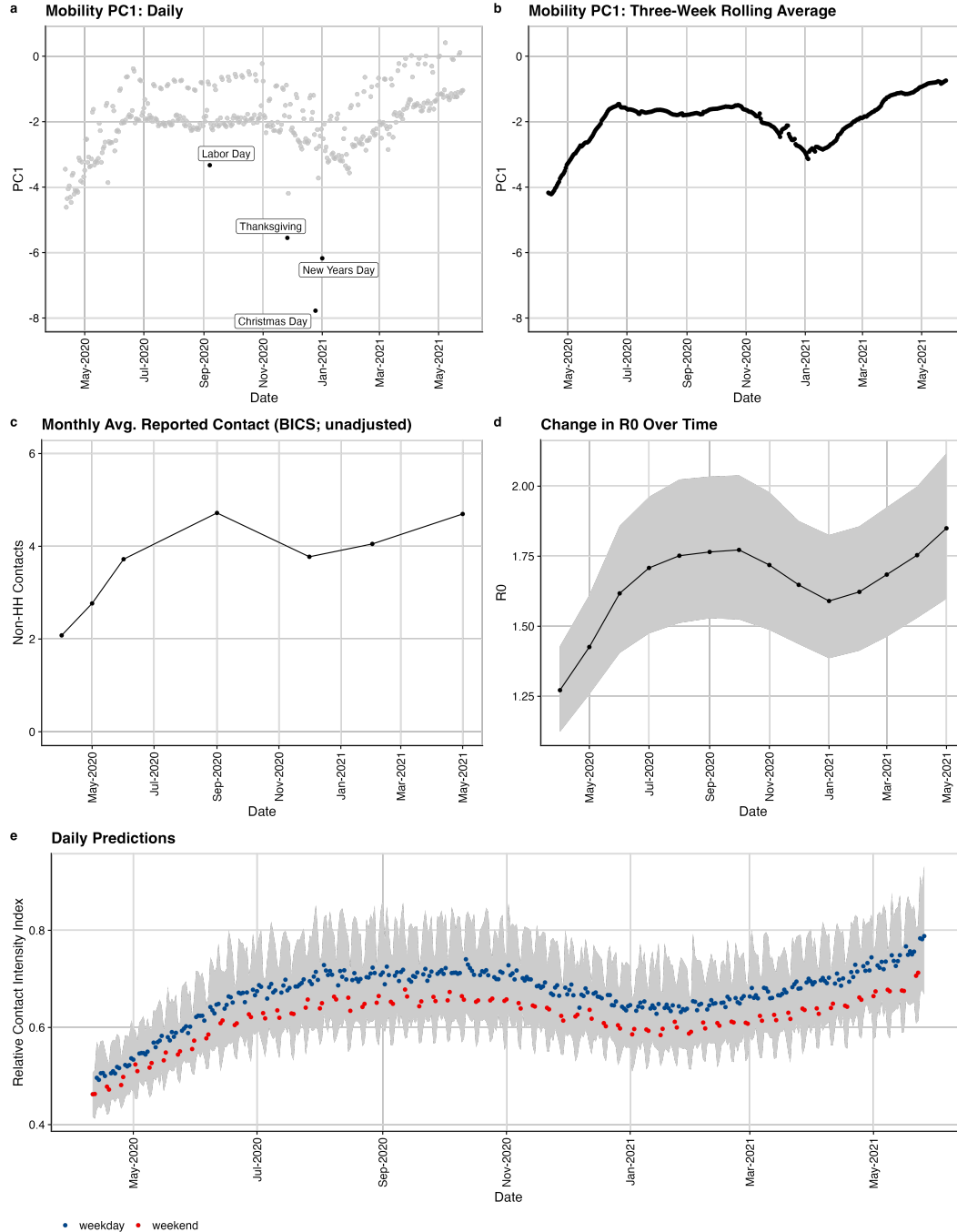


Figure E: Model input and output for state of California. Panel a shows the first principle component of seven different Google mobility signals. Panel b shows the three-week rolling average of the first principle component of seven different Google mobility signals. Panel c shows the unadjusted estimates of contact from BICS survey. Panel d shows the change in R_0 over time. Panel e shows the daily MRP model predictions.

In [Figure F](#), we show the estimated state-level relative contact intensity index over time from a model fit *without* the mobility predictor. The estimates from our model without the

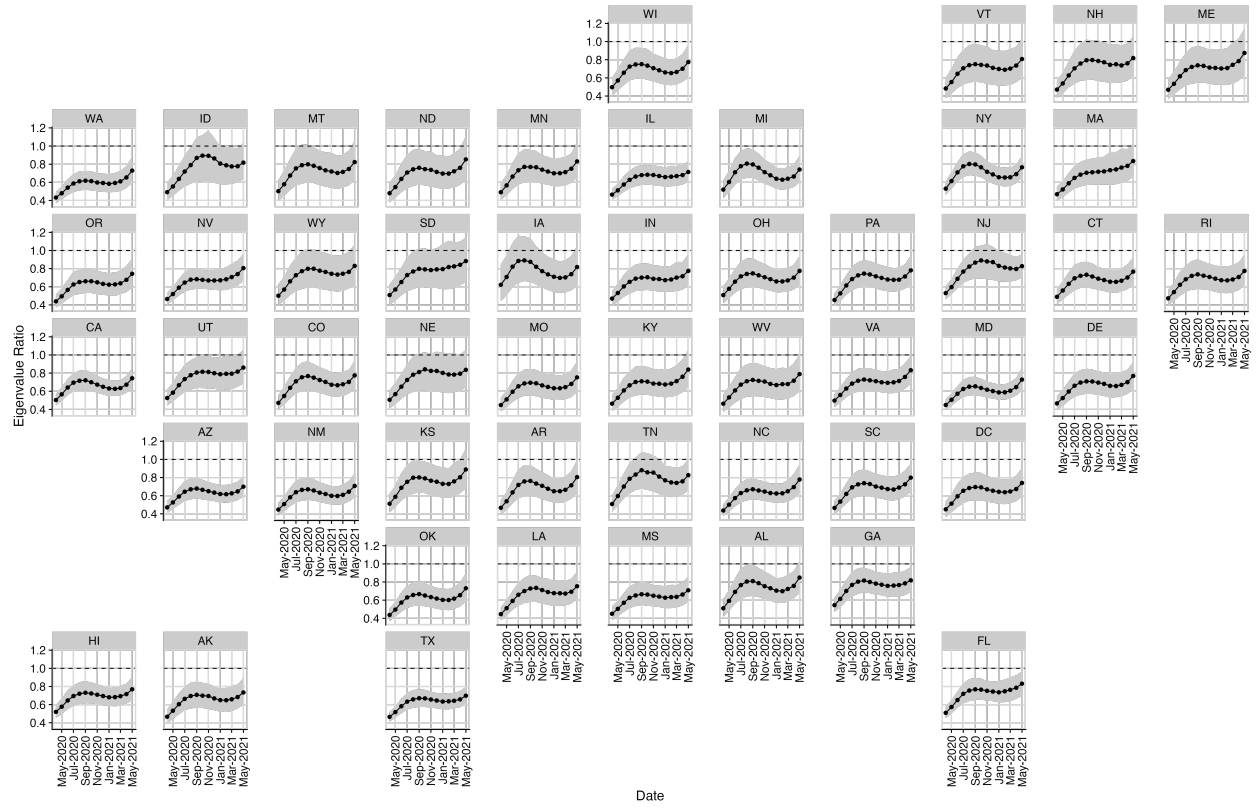


Figure F: State-level change in relative contact intensity index over time from a model without mobility data predictors.

mobility predictor closely track our estimates from the model with the mobility predictor, demonstrating that mobility data is not a prerequisite for this approach.

3.5 Model Code

All models were fit using the `brms` package in R. Full `brms` syntax below:

```
## non-household model
nonhh_model <- brm(
  formula = mvbind(alter_age_0_18, alter_age_18_25, alter_age_25_35,
    alter_age_35_45, alter_age_45_55, alter_age_55_65, alter_age_65_100) ~
    agecat +
    gender +
    hhsize +
    race +
    weekday +
    agecat:gender +
    rolling_pca_3wks + ## mobility data
    poly(day_std, 3) + ## national polynomial trend
    (poly(day_std, 3) | estimated_state), ## state level hierarchical polynomial trend
  family = 'negbinomial',
  inits = 0,
  data = df_formodel_topcode,
  chains = 4,
  cores = 16,
  backend = "cmdstanr",
  threads = threading(4),
  # control = list(adapt_delta = 0.9),
  iter = 4000)

## household model
hh_model <- brm(
  formula = mvbind(alter_age_0_18_hh, alter_age_18_25_hh, alter_age_25_35_hh,
    alter_age_35_45_hh, alter_age_45_55_hh, alter_age_55_65_hh, alter_age_65_100_hh) ~
    agecat +
    gender +
    hhsize +
    race +
    agecat:gender +
    (1 | state),
  family = 'negbinomial',
  inits = 0,
  data = df_formodel_topcode_household,
  chains = 4,
  cores = 16,
  iter = 4000)
```

3.6 Contact Matrices by month

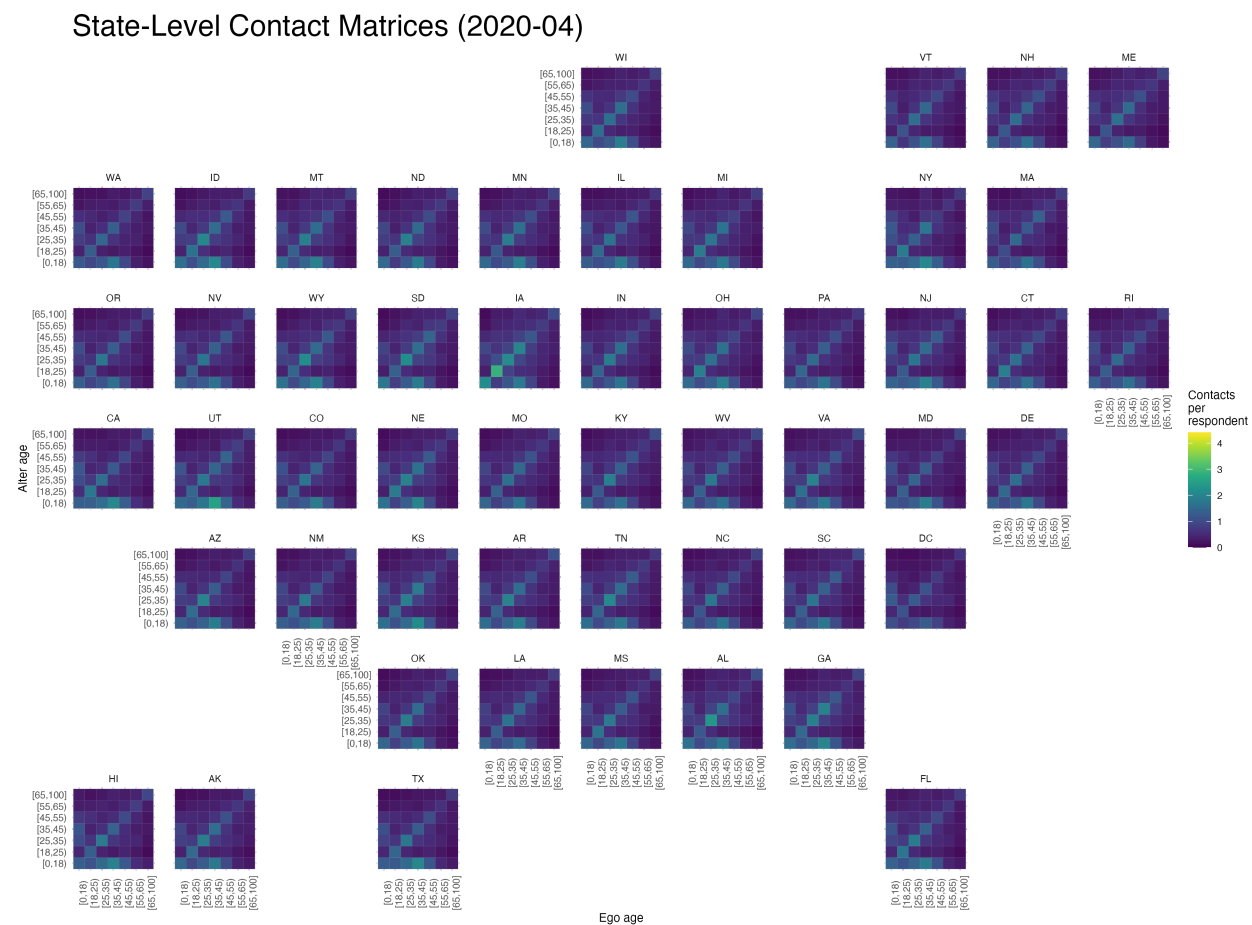


Figure G: Estimated contact matrix for 2020-04.

State-Level Contact Matrices (2020-05)

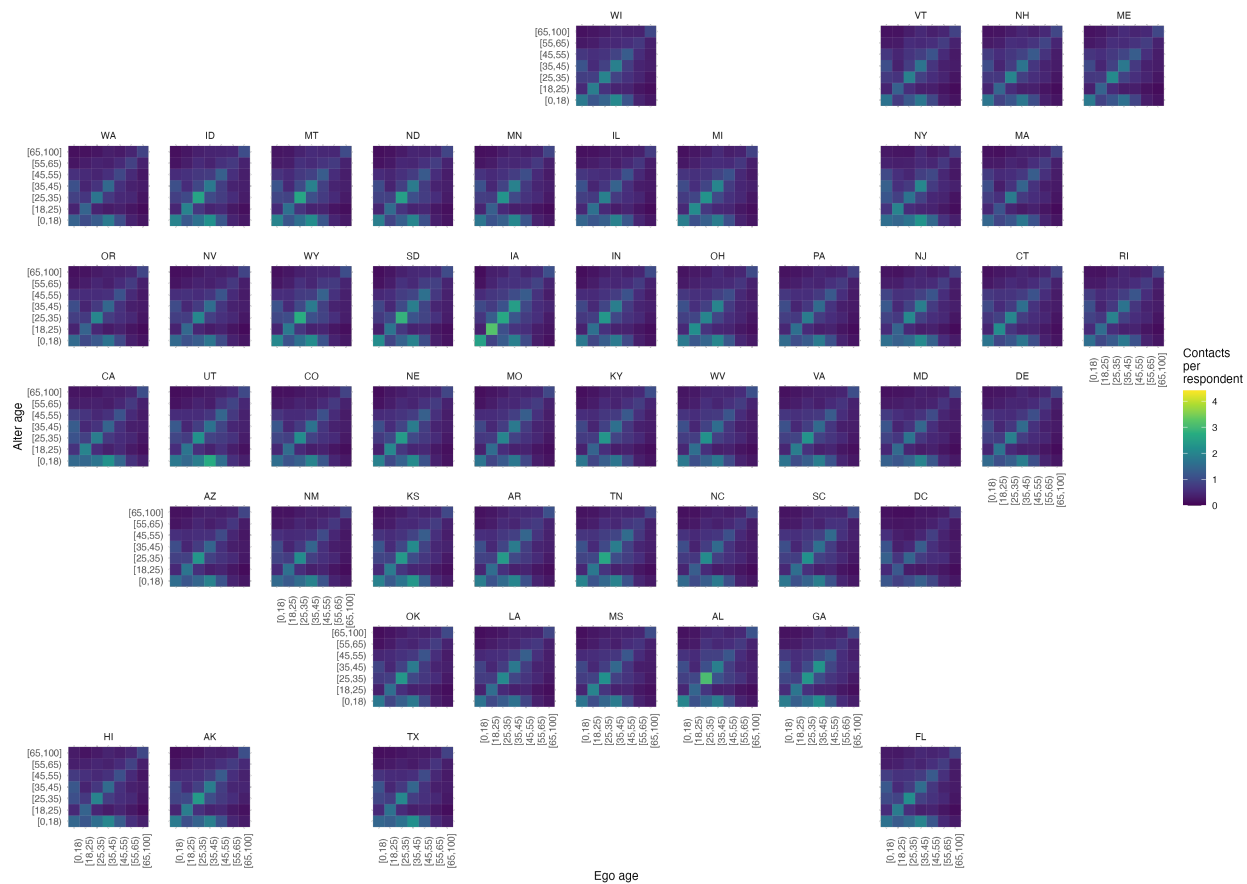


Figure H: Estimated contact matrix for 2020-05.

State-Level Contact Matrices (2020-06)

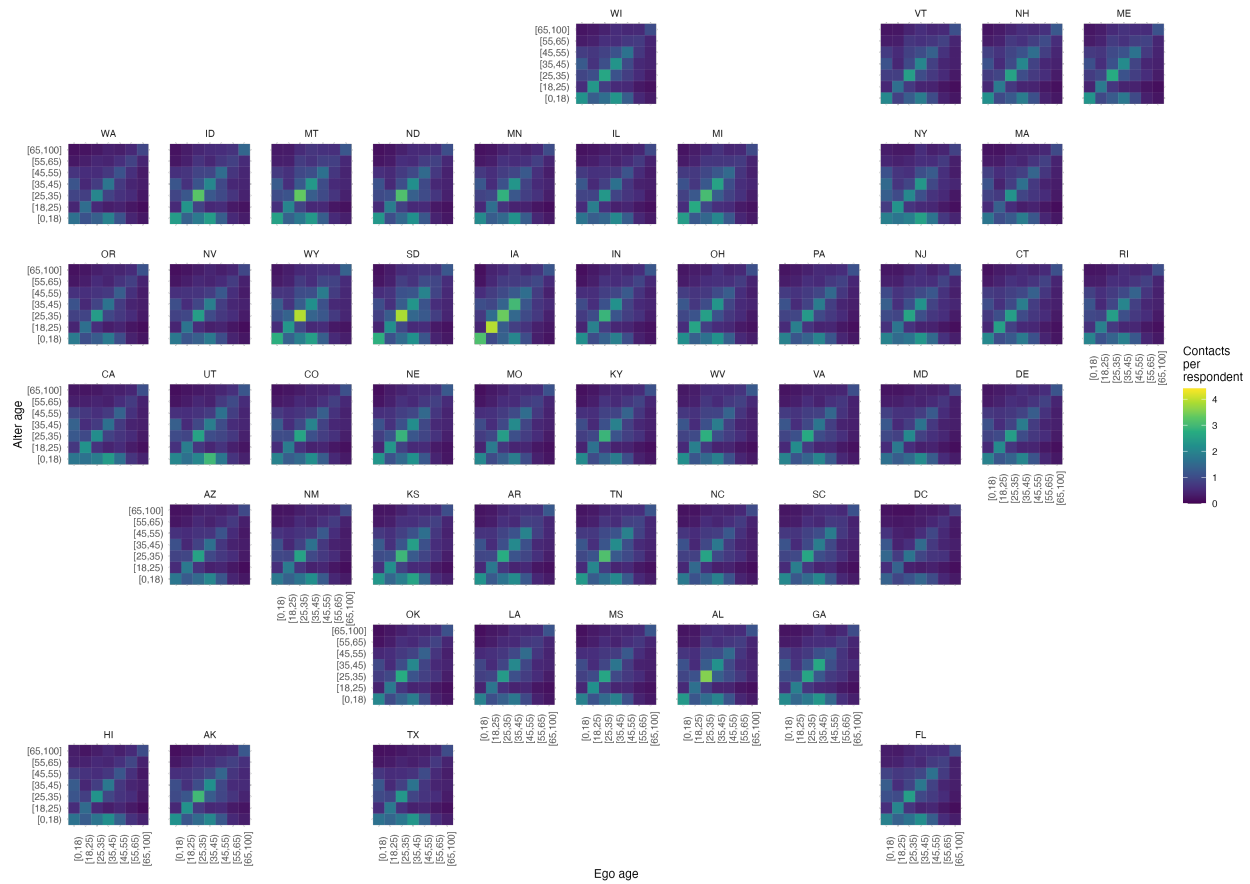


Figure I: Estimated contact matrix for 2020-06.

State-Level Contact Matrices (2020-07)

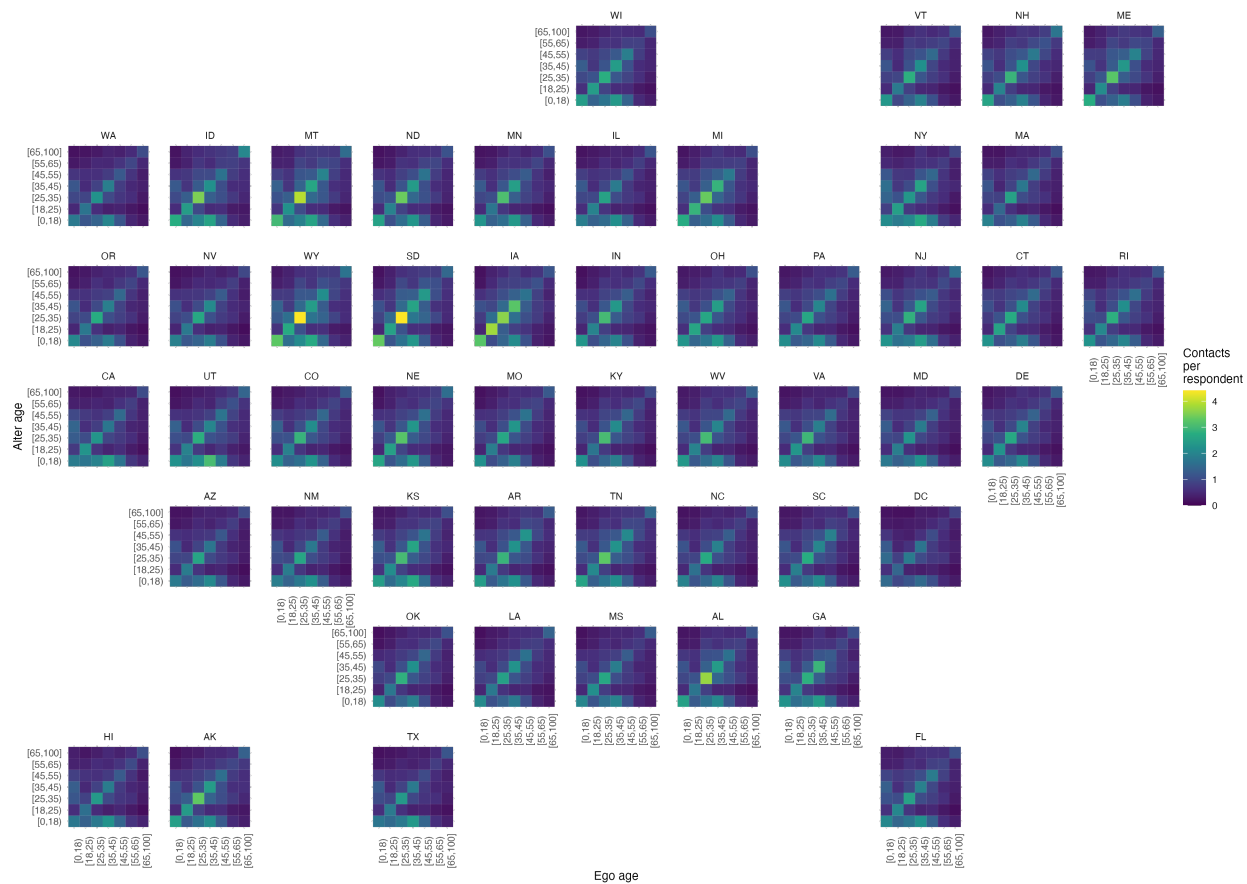


Figure J: Estimated contact matrix for 2020-07.

State-Level Contact Matrices (2020-08)

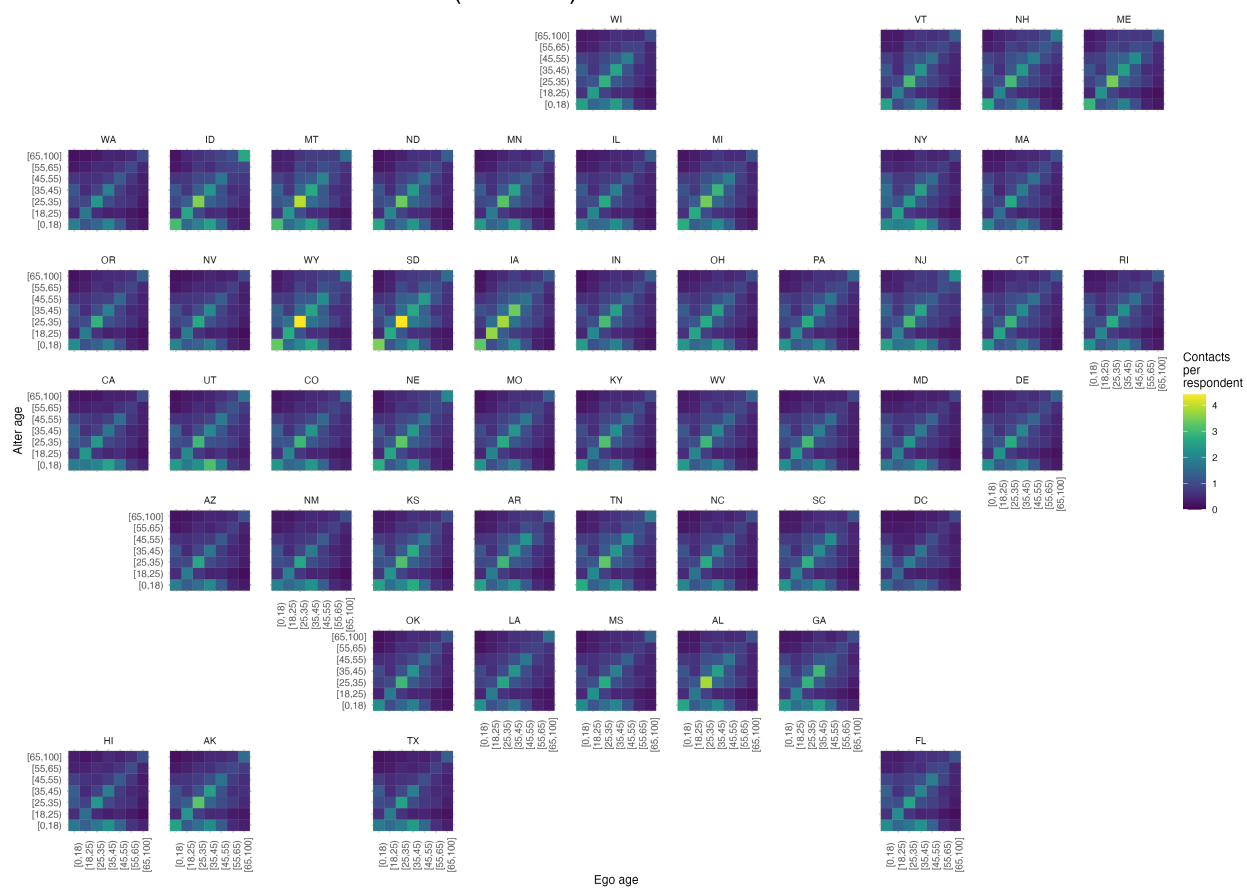


Figure K: Estimated contact matrix for 2020-08.

State-Level Contact Matrices (2020-09)

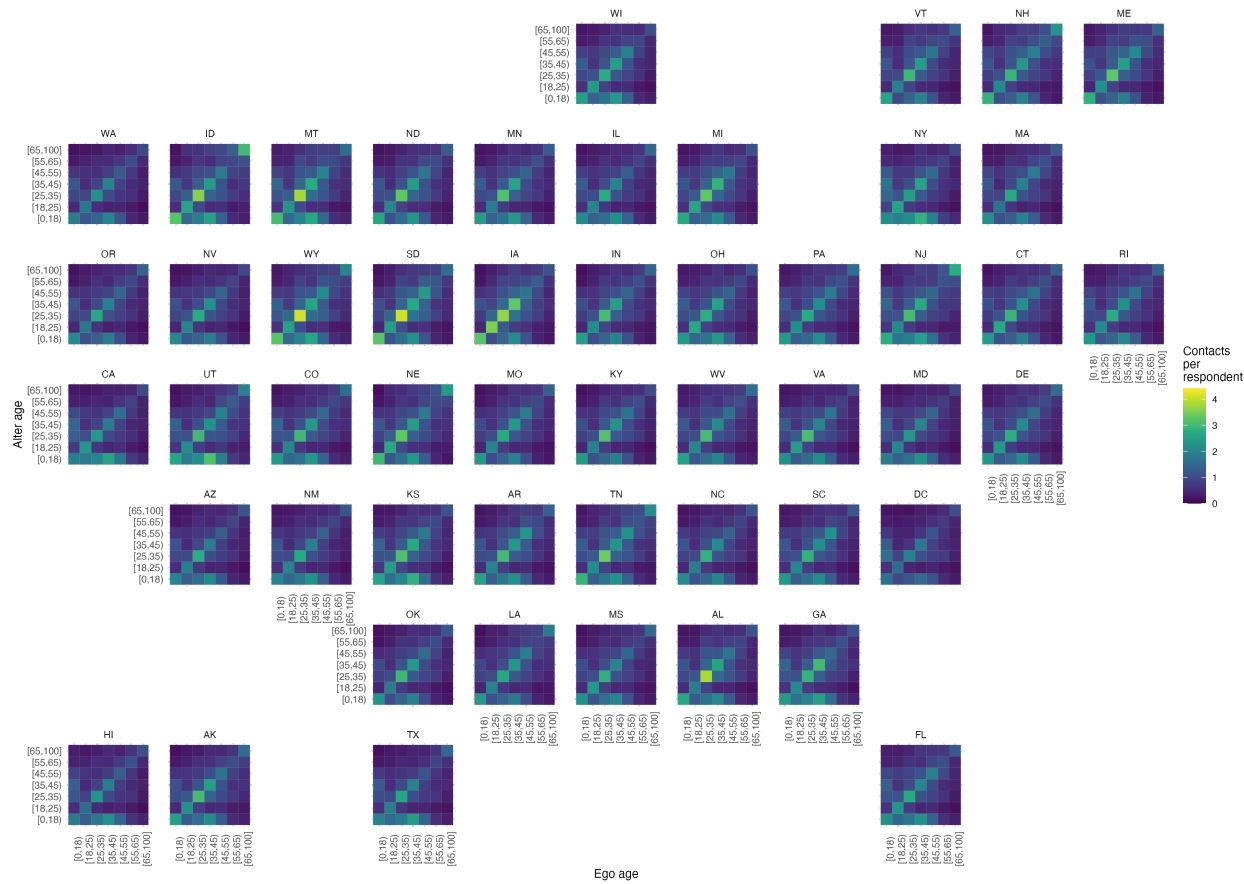


Figure L: Estimated contact matrix for 2020-09.

State-Level Contact Matrices (2020-10)

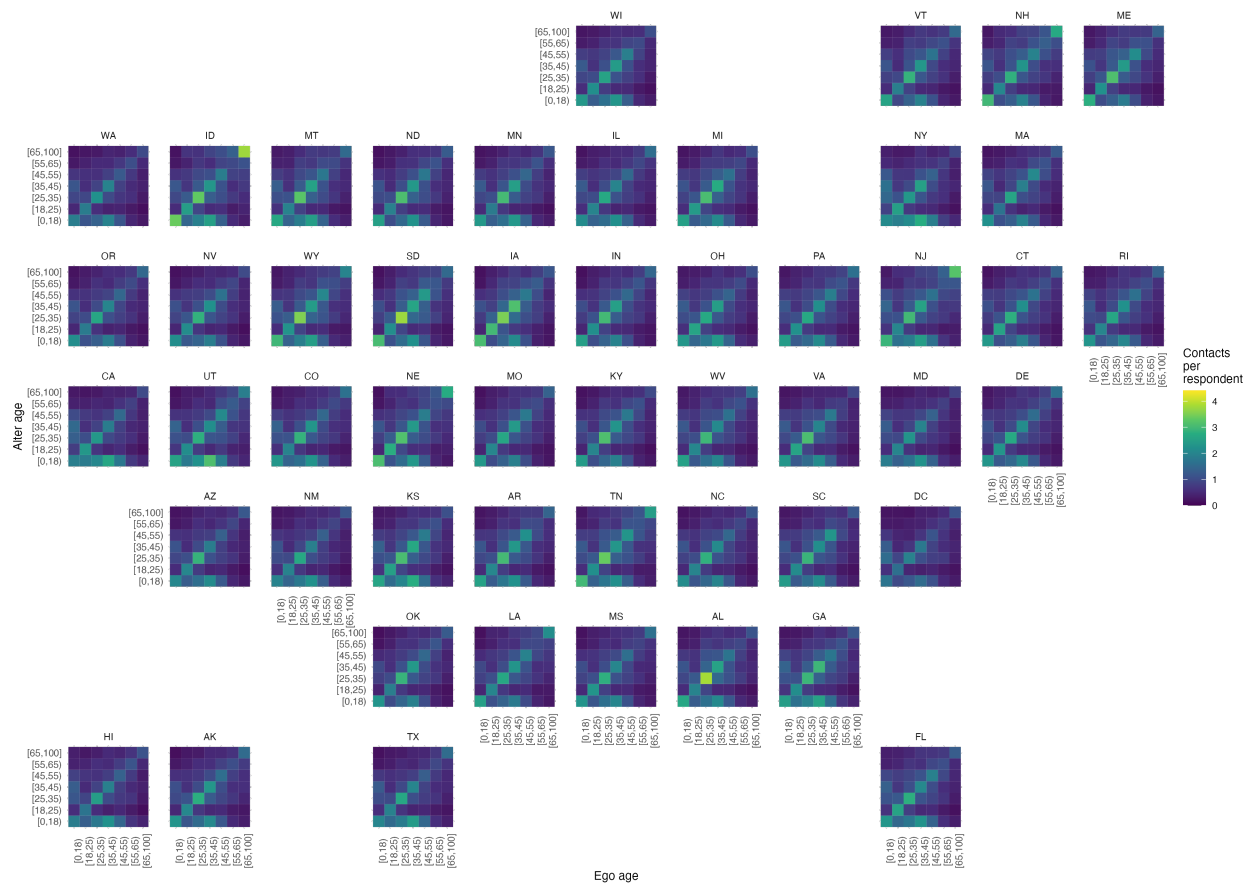


Figure M: Estimated contact matrix for 2020-10.

State-Level Contact Matrices (2020-11)

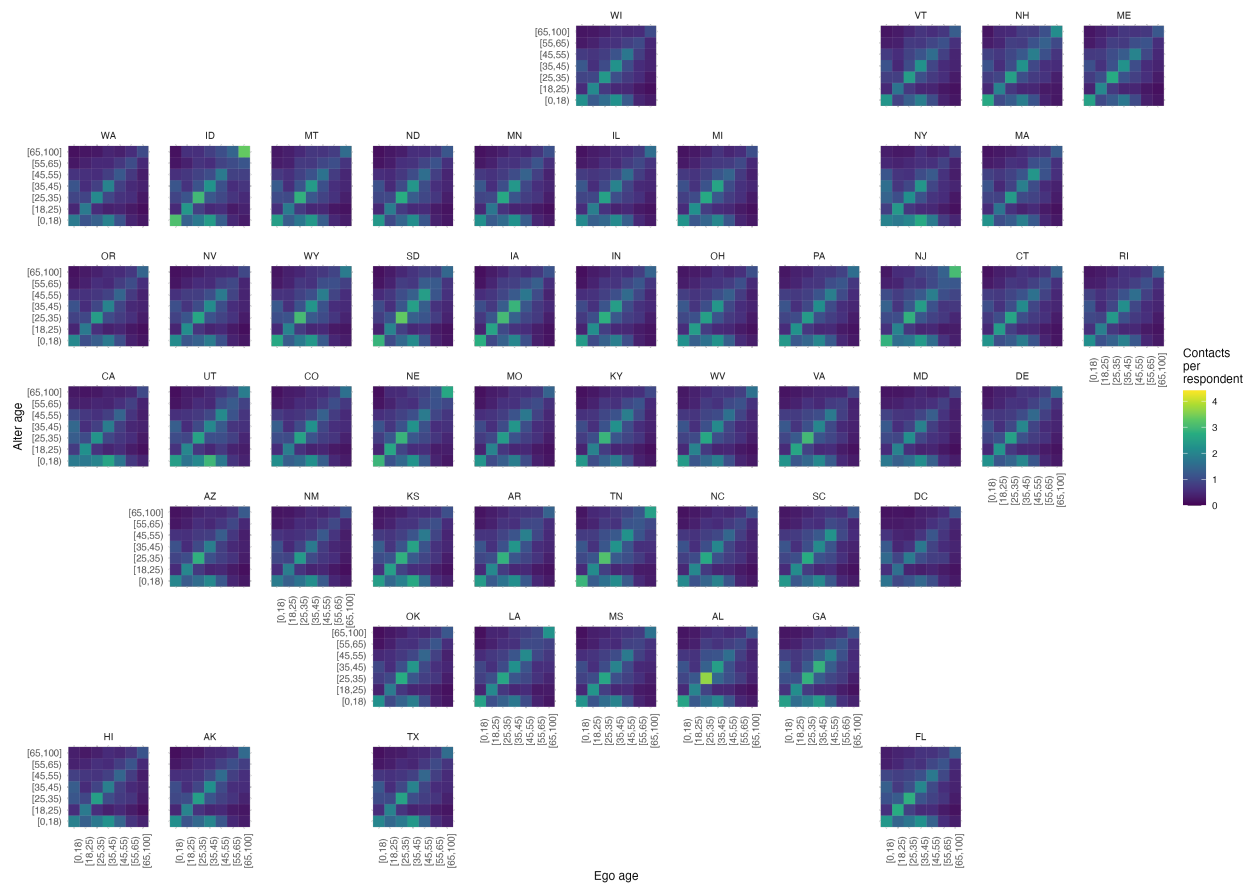


Figure N: Estimated contact matrix for 2020-11.

State-Level Contact Matrices (2020-12)

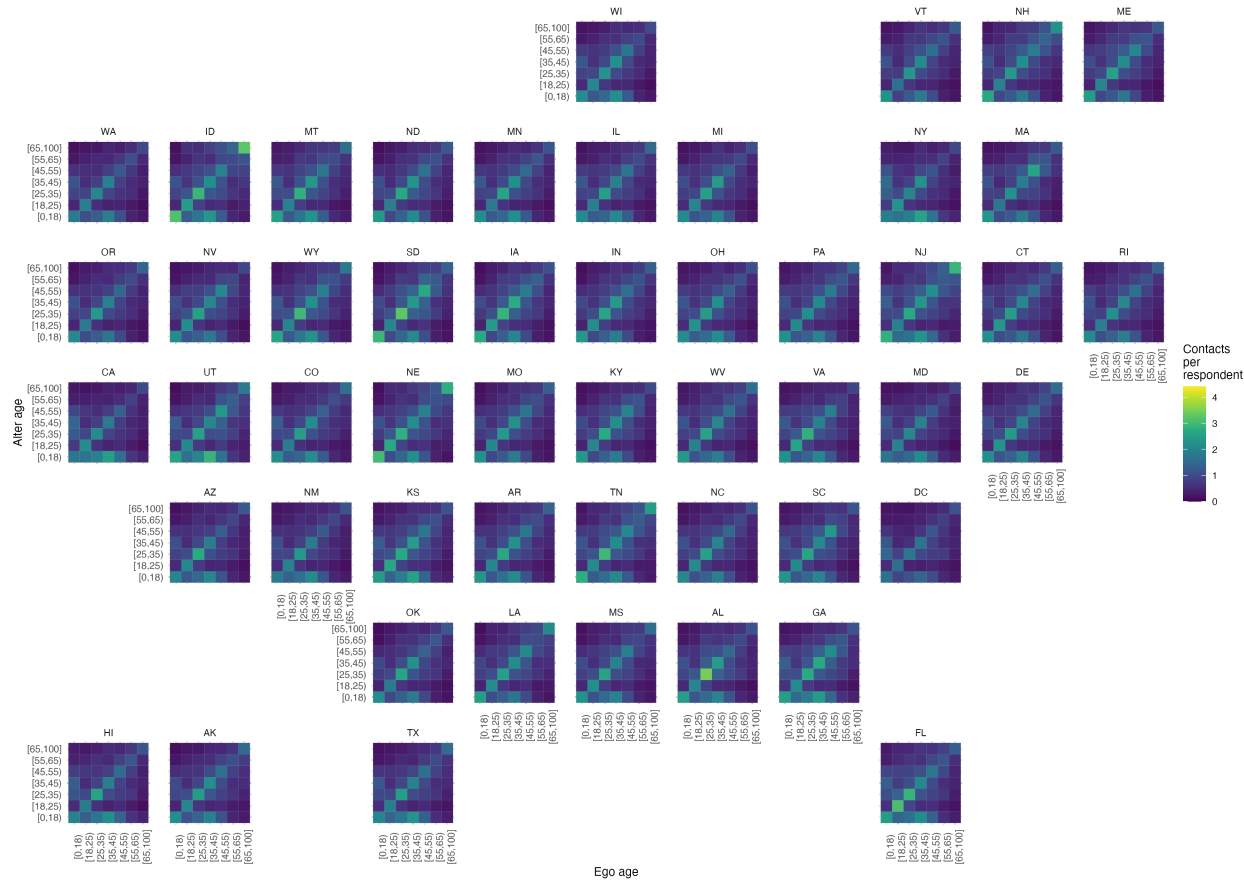


Figure O: Estimated contact matrix for 2020-12.

State-Level Contact Matrices (2021-01)

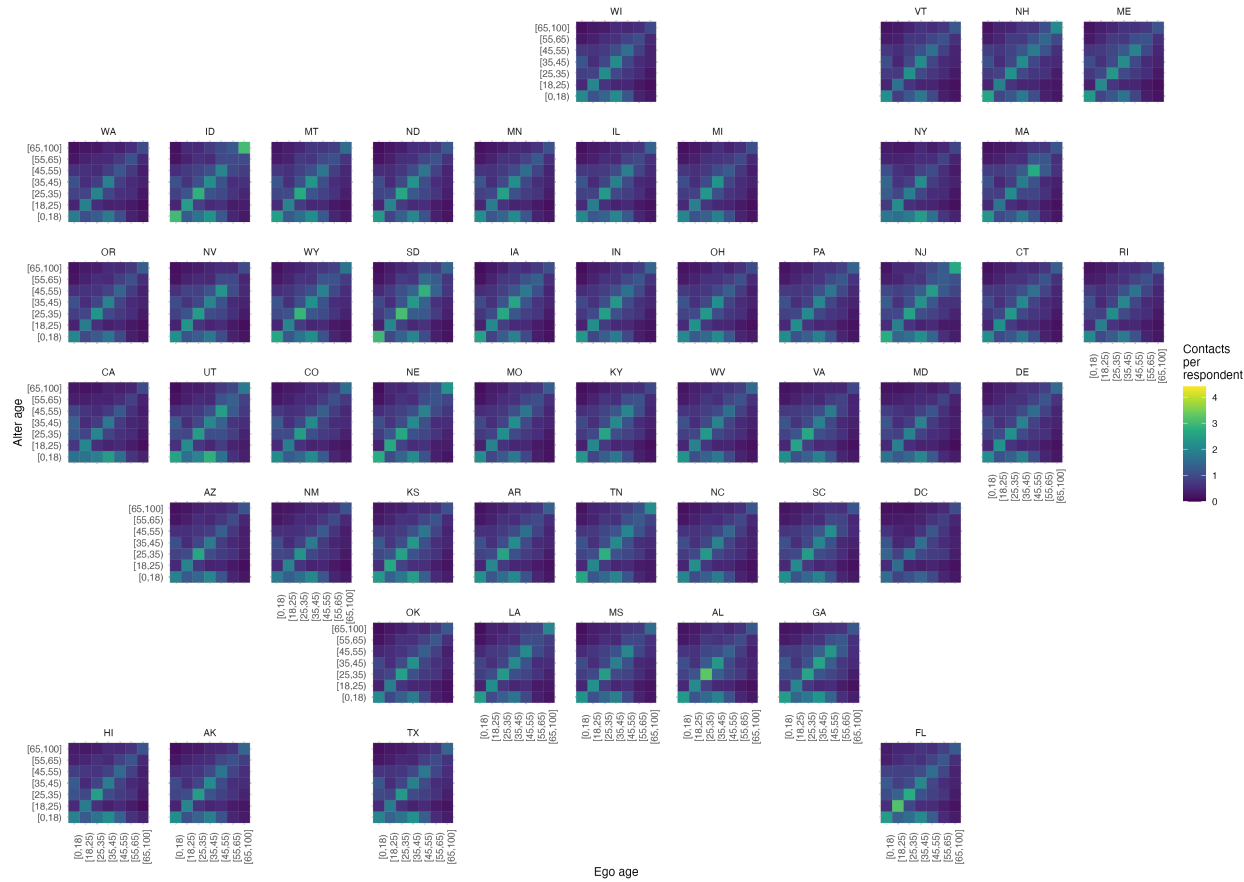


Figure P: Estimated contact matrix for 2021-01.

State-Level Contact Matrices (2021-02)

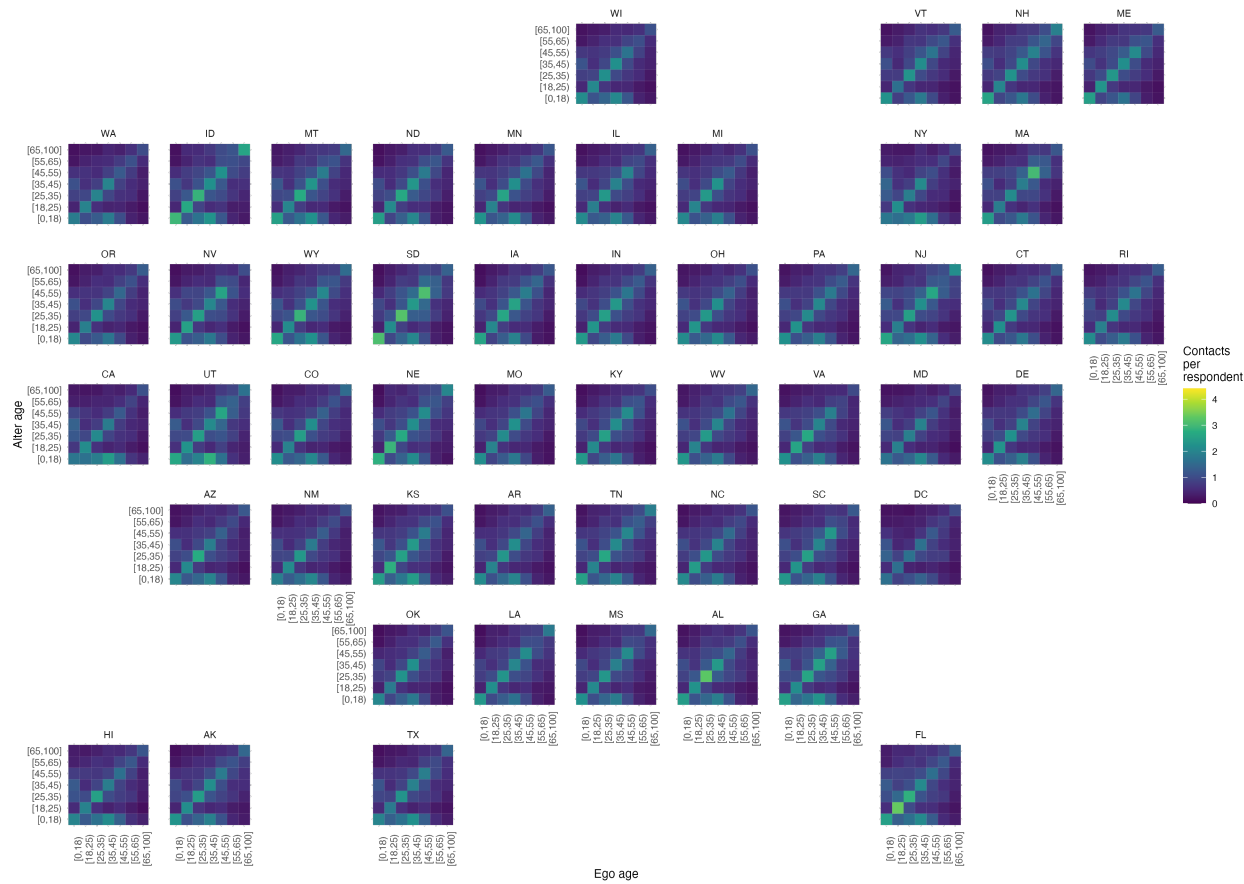


Figure Q: Estimated contact matrix for 2021-02.

State-Level Contact Matrices (2021-03)

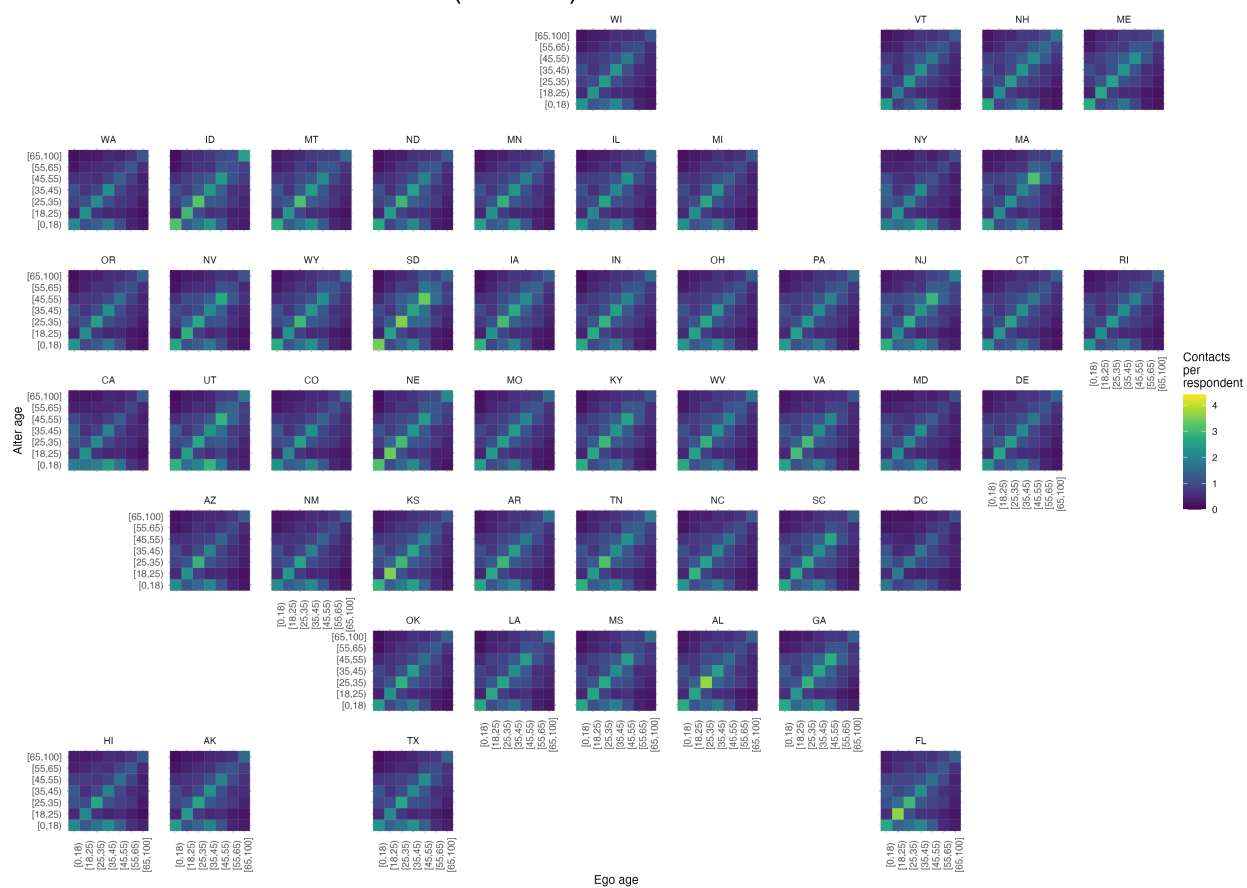


Figure R: Estimated contact matrix for 2021-03.

State-Level Contact Matrices (2021-04)

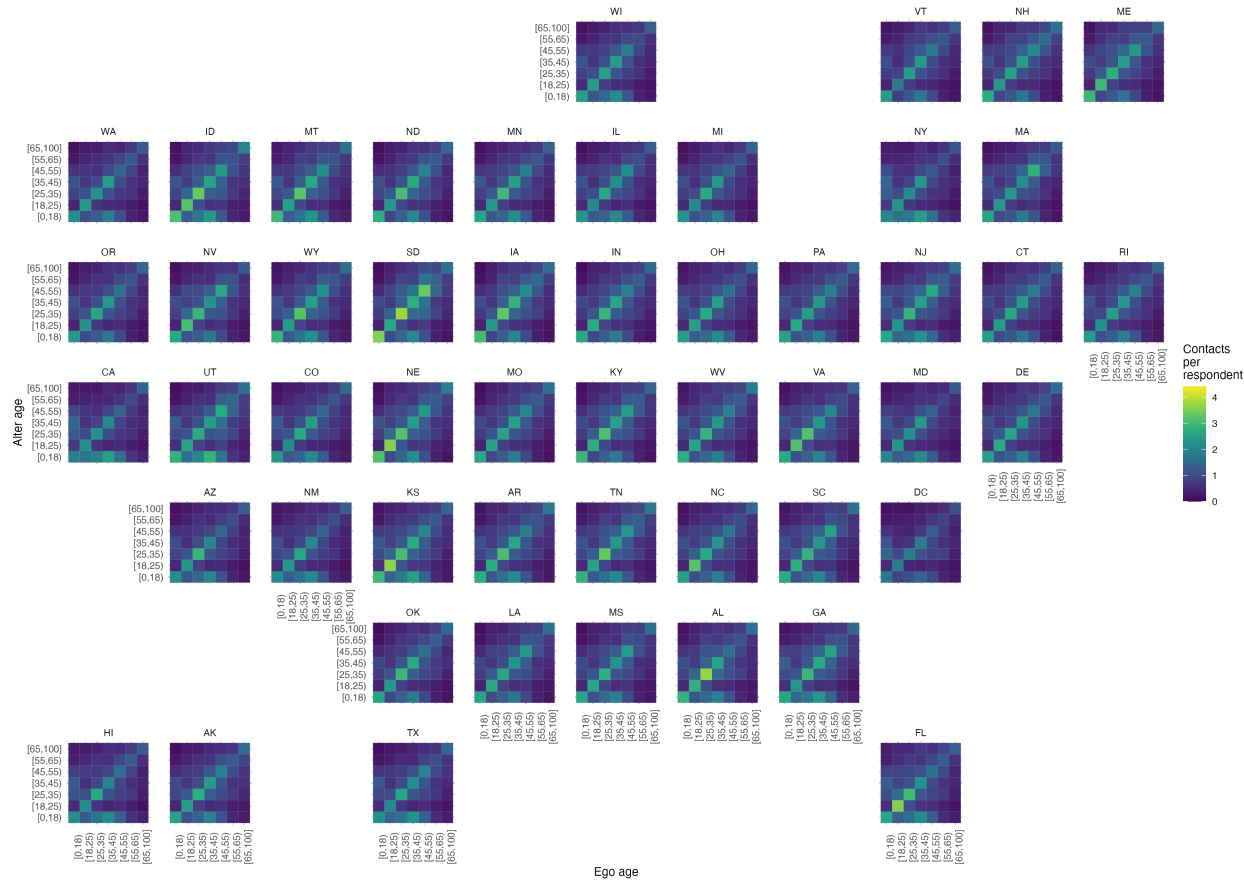


Figure S: Estimated contact matrix for 2021-04.

State-Level Contact Matrices (2021-05)

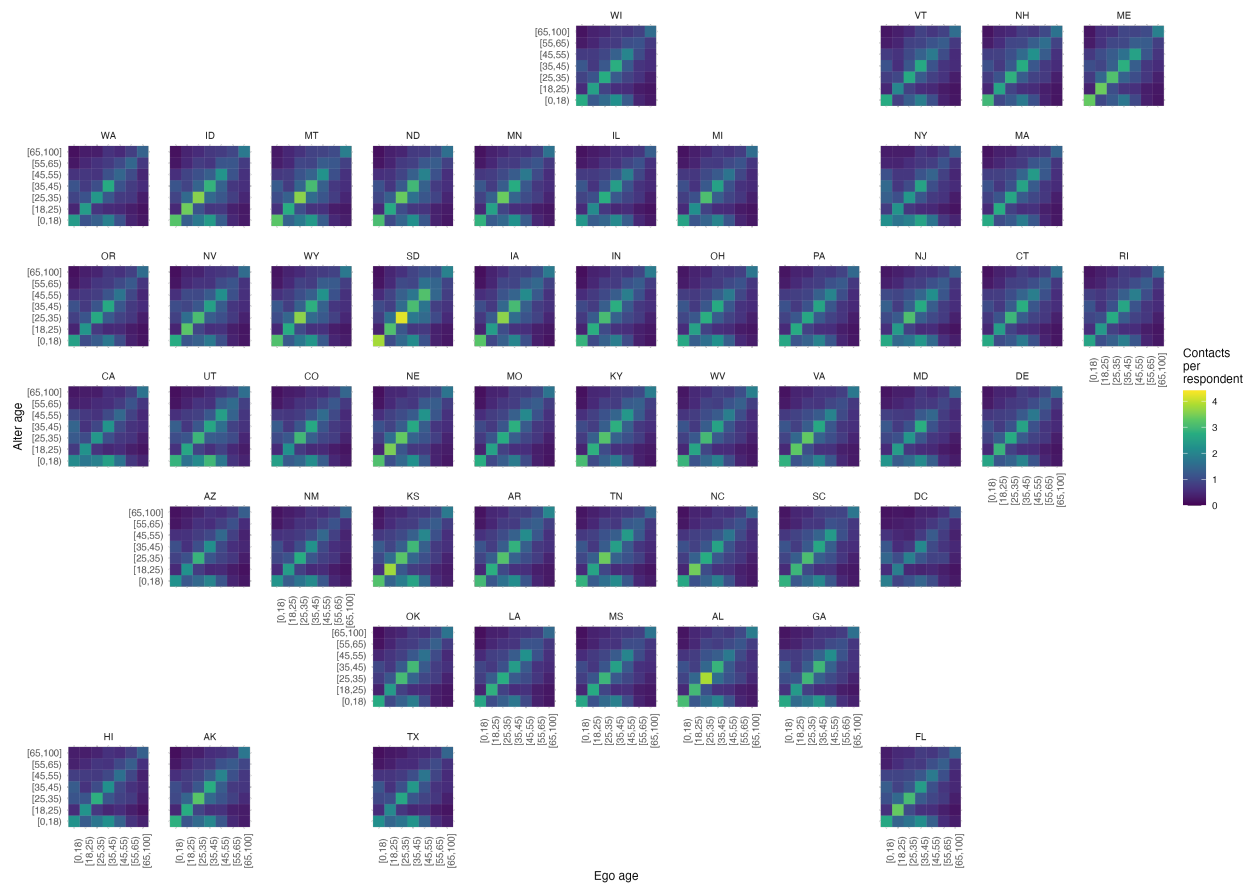


Figure T: Estimated contact matrix for 2021-05.

References

1. Bürkner, Paul-Christian. 2017. “Brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01 .
2. Chin, Taylor, Dennis M. Feehan, Caroline O. Buckee and Ayesha S. Mahmud. 2021. Contact Surveys Reveal Heterogeneities in Age-Group Contributions to SARS-CoV-2 Dynamics in the United States. Preprint Epidemiology.
3. Davies, Nicholas G., Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, CMMID COVID-19 working group, Carl A. B. Pearson, Billy J. Quilty, Adam J. Kucharski, Hamish Gibbs, Samuel Clifford, Amy Gimma, Kevin van Zandvoort, James D. Munday, Charlie Diamond, W. John Edmunds, Rein M. G. J. Houben, Joel Hellewell, Timothy W. Russell, Sam Abbott, Sebastian Funk, Nikos I. Bosse, Yueqian Fiona Sun, Stefan Flasche, Alicia Rosello, Christopher I. Jarvis and Rosalind M. Eggo. 2020. “Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics.” *Nature Medicine* 26(8):1205–1211.
4. Diekmann, O., H. Heesterbeek and T. Britton. 2013. *Mathematical Tools for Understanding Infectious Disease Dynamics*. EBSCO Ebook Academic Collection Princeton University Press.
5. Diekmann, O., J.A.P. Heesterbeek and J.A.J. Metz. 1990. “On the Definition and the Computation of the Basic Reproduction Ratio R_0 in Models for Infectious Diseases in Heterogeneous Populations.” *Journal of Mathematical Biology* 28(4).
6. Fay, Robert E. and Roger A. Herriot. 1979. “Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.” *Journal of the American Statistical Association* 74(366a):269–277.
7. Feehan, Dennis M. and Ayesha S. Mahmud. 2021. “Quantifying Population Contact Patterns in the United States during the COVID-19 Pandemic.” *Nature Communications* 12(1):893.
8. Feehan, Dennis M. and Curtiss Cobb. 2019. “Using an Online Sample to Estimate the Size of an Offline Population.” *Demography* 56(6):2377–2392.
9. Hale, Thomas, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar and Helen Tatlow. 2021. “A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker).” *Nature Human Behaviour* 5(4):529–538.
10. Hens, N., Z. Shkedy, M. Aerts, C. Faes, P. Van Damme and P. Beutels. 2012. *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*. Statistics for Biology and Health Springer New York.
11. Holmdahl, Inga, Rebecca Kahn, James Hay, Caroline O. Buckee and Michael Mina. 2020. “Frequent Testing and Immunity-Based Staffing Will Help Mitigate Outbreaks in Nursing Home Settings.”

12. McEvoy, David, Conor G. McAloon, Áine B. Collins, Kevin Hunt, Francis Butler, Andrew W. Byrne, Miriam Casey, Ann Barber, John Griffin, Elizabeth Ann Lane, Patrick Wall and Simon J. More. 2020. “The Relative Infectiousness of Asymptomatic SARS-CoV-2 Infected Persons Compared with Symptomatic Individuals: A Rapid Scoping Review.”.
13. Mossong, Joël, 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. 2008. “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases.” *PLoS Medicine* 5(3):e74.
14. Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. “Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC.” *Statistics and Computing* 27(5):1413–1432.
15. Viceconte, Giulio and Nicola Petrosillo. 2020. “COVID-19 R0: Magic Number or Conundrum?” *Infectious Disease Reports* 12(1):8516.
16. Wallinga, Jacco, Peter Teunis and Mirjam Kretzschmar. 2006. “Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents.” *American Journal of Epidemiology* 164(10):936–944.