nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20544-y OPEN

Inferring high-resolution human mixing patterns for disease modeling

Dina Mistry[1,2], Maria Litvinova[2,3,4], Ana Pastore y Piontti[2], Matteo Chinazzi[2], Laura Fumanelli[5], Marcelo F. C. Gomes[6], Syed A. Haque[2], Quan-Hui Liu[7], Kunpeng Mu[2], Xinyue Xiong[2], M. Elizabeth Halloran[8,9], Ira M. Longini Jr.[10], Stefano Merler[5], Marco Ajelli[2,4]✉ & Alessandro Vespignani[2,3]✉

Mathematical and computational modeling approaches are increasingly used as quantitative tools in the analysis and forecasting of infectious disease epidemics. The growing need for realism in addressing complex public health questions is, however, calling for accurate models of the human contact patterns that govern the disease transmission processes. Here we present a data-driven approach to generate effective population-level contact matrices by using highly detailed macro (census) and micro (survey) data on key socio-demographic features. We produce age-stratified contact matrices for 35 countries, including 277 sub-national administratvie regions of 8 of those countries, covering approximately 3.5 billion people and reflecting the high degree of cultural and societal diversity of the focus countries. We use the derived contact matrices to model the spread of airborne infectious diseases and show that sub-national heterogeneities in human mixing patterns have a marked impact on epidemic indicators such as the reproduction number and overall attack rate of epidemics of the same etiology. The contact patterns derived here are made publicly available as a modeling tool to study the impact of socio-economic differences and demographic heterogeneities across populations on the epidemiology of infectious diseases.

[1] Institute for Disease Modeling, Global Health Division, Bill and Melinda Gates Foundation, Seattle, WA, USA. [2] Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, USA. [3] ISI Foundation, Turin, Italy. [4] Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA. [5] Bruno Kessler Foundation, Trento, Italy. [6] Fiocruz, Scientific Computing Program, Grupo de Métodos Analíticos em Vigilância Epidemiológica, Rio de Janeiro, Brazil. [7] College of Computer Science, Sichuan University, Chengdu, Sichuan, China. [8] Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [9] Department of Biostatistics, University of Washington, Seattle, WA, USA. [10] Department of Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL, USA. ✉email: marco.ajelli@gmail.com; a.vespignani@northeastern.edu

Mathematical and computational models of infectious disease transmission are increasingly used to provide scenario analysis and forecasts during epidemic outbreaks and quantitative answers to complex public health questions such as devising the effectiveness of control strategies (vaccination, school closure, etc.) during health threat emergencies[1]. Modeling approaches have thus moved away from the classic homogeneous and stylized framework[2,3], progressively incorporating heterogeneities that depend on between- and within-country population variability, disease timescale, transmission settings, as well as specific pathogen characteristics. For instance, geographically structured models allow evaluation of spatially heterogeneous interventions in both animal and human diseases[4,5], while individual-based models lay down the possibility of simulating all micro-details of the transmission process and tracking in time and space each individual of the simulated population[6–8]. When data-driven, these approaches have highlighted the importance of the social, demographic, and economic characteristics of the population in determining the actual mesh of contacts underlying disease spreading among individuals. For this reason, a broad range of methodologies have been used to study human-mixing patterns, including surveys[9–12], contact diaries[13–19], wearable sensors[20,21], analysis of time-use data[22], development of synthetic populations[23–25], and mixed approaches for instance integrating diary-based contact data with time-use data[26,27] or combining contact data with modeling techniques[26,28,29]. However, each methodology has different limitations and assumptions because contact patterns among individuals vary according to the geographical scale (from census blocks to the national level), the disease under consideration, and the detailed socio-economic and demographic characteristics of the population.

Here, we present a data-driven approach to generate effective descriptions of complex contact patterns that can be used to inform infectious disease modeling approaches, including the widely adopted compartmental modeling framework. We make use of highly detailed macro (census) and micro (survey) data from publicly available sources on key socio-demographic features (e.g., age structure, household composition and members' age gaps, employment rates, school structure) to construct synthetic populations of interacting agents, each one representing a hypothetical individual in the real population. The proposed method relies on both macro- and micro-level data for multiple socioeconomic characteristics and can be adapted to different geographical contexts and diseases; something that is not possible in a "one-model-fits-all" approach.

We provide synthetic contact matrices for nations around the world with substantially large and diverse populations. Specifically, we report contact patterns at the subnational level in the following countries: Australia, Canada, China, India, Israel, Japan, Russia, South Africa, and the United States of America. These populations account for 277 subnational administrative regions (such as states, provinces, prefectures, territories, etc. depending on the considered country), cover ~38% of the world's surface area, and account for ~3.5 billion people of the world's 7.6 billion population. The resulting synthetic populations are used to generate age-stratified contact matrices for the most common social settings, in which individuals spend their time interacting with each other (i.e., households, schools, workplaces, and the general community). The resulting contact matrices capture differences at the subnational level that reflect the high degree of cultural and societal diversity of the focus countries. This approach allows us to provide a mesoscopic description of the human contact patterns that can be used in the mathematical and computational analysis of infectious disease spread (see Fig. 1).

To illustrate the importance of considering national and subnational heterogeneities in the analysis of infectious disease epidemiology, we construct the contact matrix relevant for airborne infectious diseases by calibrating the combination of setting-specific (household, workplace, school) contact matrices using as ground truth seven diary-based contact matrices (six European countries[14] and Russia[18]). The resulting matrices are validated against out-of-sample contact data collected in France[30], Japan[31], and China[32]. These contact matrices are then used in the modeling of influenza transmission patterns at the national and subnational levels. The influenza modeling simulations, although considering identical disease etiology, highlights considerable heterogeneities in reproduction number and attack rates across regions of the world included in this study, reflecting differences in key demographic properties such as average age and student population.

As a service to the community, a database containing the inferred setting-specific matrices as well as the overall contact matrices for all locations (and countries) is available on the dedicated online repository: https://github.com/mobs-lab/mixing-patterns. Python codes to work with the contact matrices and examples of how to use them in age-structured compartmental models are available on the same website as well. This presented work can be easily generalized to other countries and settings, and arm the community with a general framework that can be used to make inference on important epidemiological parameters in the modeling of infectious diseases.

Results

We use a data-driven computational approach to infer the contact networks in the social settings where people interact and spend most of their time. In particular, we focus on four social settings (household, school, workplace, and the general community), which are particularly relevant for influenza transmission[7,33]. To reconstruct the synthetic population in each context we use a wide variety of national and subnational micro-level, census, and demographic data that provide the separate characteristics of the population, and the association of multiple characteristics. Micro-level data drawn from socio-demographic surveys are especially useful as no assumptions on the rules of disaggregation are required (the data are already on the required level of disaggregation).

Contacts between individuals in the real-world populations are inferred by analysis of the generated data-driven synthetic networks by measuring the frequency of links between individuals (living, going to school, or working together) in the synthetic contact networks of the different social settings. Then we compare summary statistics derived from the generated synthetic population for each geographical area to those reported in official (macro) statistics (e.g., census data). Examples of the summary statistics used in the approach are the age structure of the population, distributions of household size, type, number of children by household size, and so on, depending on the summary statistics available from official sources. The generated data is compared to the distributions of summary statistics by using the goodness-of-fit tests at the desired level of significance (generally 5%). We use a non-parametric bootstrap procedure to test the uncertainty level of our sampling. This procedure is iterated until a satisfactory fit is reached. In the case of inadequate microdata (e.g., sub-optimal sample size), we use the microdata to extrapolate rules on the age gaps between household members conditioned on the age of the household head, household size, and the relation between the members (e.g., age gap between spouses, age gap between siblings). Note that the same arguments are extended to other settings (e.g., schools, workplaces, hospitals)

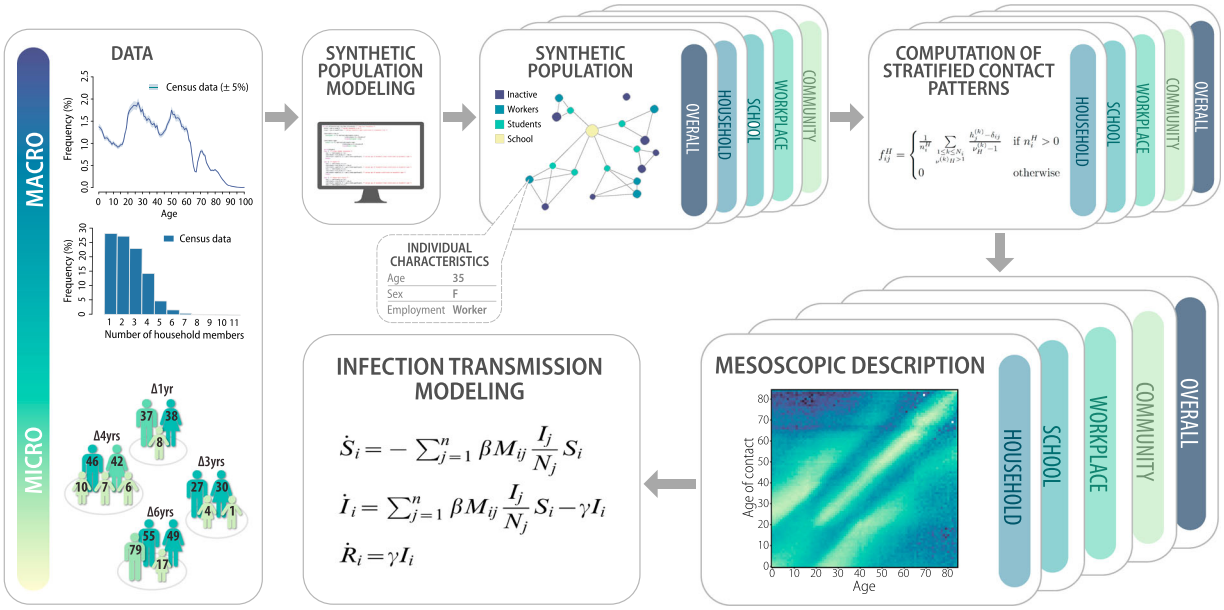

Fig. 1 Modeling framework. Schematic representation of the workflow for modeling human-mixing patterns and infection transmission dynamics.

and can be extended to further stratifications relevant for other diseases (e.g., easy access to health care facilities). An illustration of the matrices construction workflow is reported in Fig. 1, while the full technical description is reported in "Methods" and Supplementary Information.

Setting-specific contact matrices. We report here the results for populations of 277 subnational administrative regions of Australia, Canada, China, India, Israel, Japan, Russia, South Africa, and the United States of America, characterizing contact patterns for about 3.5 billion individuals. We also include data at the national level for 26 European countries[23]. The inferred age-specific contact matrices reveal strong patterns, of which many are common to the diverse locations under study. Figure 2 shows the age-mixing patterns F_{ij}^{k} defined as the per capita frequency of contact of an individual of age i with an individual of age j in setting k.

Starting with the household setting in Fig. 2a, we observe that contacts between individuals can largely be characterized as that between couples living together, and parents and their children in the same household[14,23]. The increased frequency of contact between adults of similar ages along the main diagonal of the household contact matrix represents couples of similar ages living together, while the bands of high frequency above and below the main diagonal indicate contact between parents and children. While most locations share these overall features, the contact matrices show different age-mixing patterns. For instance, in China (Fig. 2a), the lower frequency of contact between children within households is the reflection of the country's so-called "One-child policy". The policy, enacted in 1979 up to 2016, has resulted in over a generation of many Chinese youths growing up without siblings, and hence having less contact on average with other children in this setting. This is a stark contrast with the United States and India (Fig. 2a), where the presence of multiple children born to a family results in an increased frequency of contact between this age group in the household matrix. The presence of multigenerational families in countries like India is also evident from the increased frequency of contact between all age groups, notably between the elderly (60 years and older) and young children. The same feature was observed in ref. [19] for Zimbabwe. Even within the same country, contact patterns may be markedly different. Figure 2a shows the age-mixing patterns within households for two different provinces of China: Beijing and Guizhou. While the household contact patterns in Beijing show a clear signal of the "One-child policy", Guizhou shows the presence of multigenerational families, as well as an increased presence of multiple children living in the same household. This can be traced back to the fact that the Guizhou Province is characterized by a large frequency of minority groups and the "One-child policy" was less strictly applied for minorities.

Figure 2b, c shows the inferred contact matrices in the school and workplace settings for China, the United States, and India. In both settings, the age-mixing patterns vary strongly, reflecting differences in the educational systems, and economic conditions unique to each location. For all locations in our study, the school setting consistently exhibits the highest frequencies of contact between children and young adults attending school together. Interaction with older adults in this setting reflects the contact students have with instructors and other staff members in school. The variability of age-mixing patterns between children in India (Fig. 2b) also reflects the many different kinds of schools that children can attend throughout the country and the different age groups found in those schools. In the workplace environment, most interaction takes place between individuals in the range of 20–65 years of age, with the age range depending on local retirement, employments regulations, and culture. For instance, in many parts of the world it is common for teenagers to be fully or partially employed (see the work contact matrix for the US—Fig. 2c); in India, census records for employment list even children among the population of workers.

Statistical validation of the contact matrices against summary statistics of a large set of socio-demographic indicators has been performed to validate our results (see Supplementary Information).

Human-mixing patterns for influenza transmission. The contact matrices obtained in each setting acquire epidemiological relevance when combined together to generate the descriptions of human-mixing patterns relevant to the spreading of a specific disease. Here, we define the matrix of effective contacts relevant to influenza transmission based on the relative contribution of the household, school, and workplace. Here, by "effective", it is

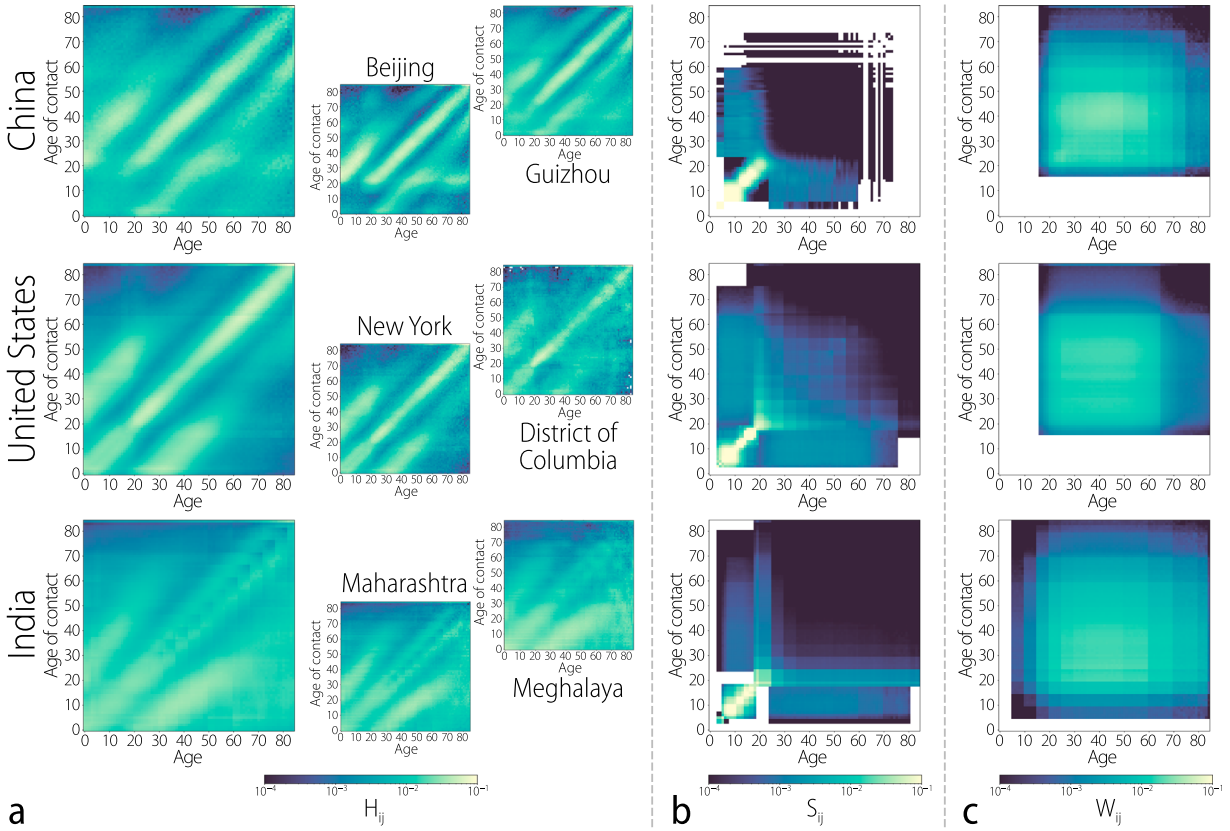

Fig. 2 Age-mixing patterns by setting. Each heatmap represents the average frequency of contact between an individual of a given age (*x* axis) and all of their possible contacts (*y* axis). **a** Matrices of household contacts by age at the national level for China, the United States, and India. The six smaller panels in the center and on the left show household contact matrices at the subnational level in two provinces of China (Beijing, Guizhou), two locations of the United States (the state of New York and the District of Columbia), and two states of India (Maharashtra, Meghalaya). **b** Matrices of school contacts by age at the national level (from top to bottom: China, the United States, India). **c** Matrices of work contacts by age at the national level (from top to bottom: China, the United States, India).

indicated a contact that can lead to the disease transmission. In addition to these three social settings, we consider also the contribution of less structured casual encounters in the population[34], by considering a community contact matrix that assumes individuals as potentially fully mixed[23]. To combine the different matrices, we propose a weighted linear combination of the derived matrices for the four considered social settings, and compute the overall matrix of contacts between individuals of age i and individuals of age j, M (whose elements are denoted as M_{ij}), as a weighted linear combination of setting-specific contact matrices:

$$M_{ij} = \sum_k \omega_k F_{ij}^k \qquad (1)$$

where the element M_{ij} represents the average number of contacts with individuals of age j for an individual of age i per day, and each $\omega_k \geq 0$ is indicating the number of contacts in each setting k.

Generally, the ω_k are unknown disease-specific weights accounting for the relative importance of the different social settings in the transmission of a specific infectious disease. In the case of airborne infectious diseases, we leverage on diary-based survey contact matrices reported in[14,18] for Finland, Germany, Italy, Luxembourg, The Netherlands, the United Kingdom, and the Tomsk Oblast of Russia. For European countries, we relied on data and the setting-specific contact matrices developed in Fumanelli et al.[23] that covers 26 countries. Unfortunately, Poland, and Belgium, which are included in the POLYMOD study[14] used to calibrate the overall contact matrix are not included in ref. [23].

We perform a multiple linear regression analysis to find the values of ω_k such that the resulting M_{ij} best fits the empirical data. Note that the empirical matrices derived in refs. [14,18] describe the average number of contacts of age j for an individual of age i, and in "Methods" we show how ω_k is related to an average number of contacts $\langle c \rangle$ per individual. The regression yields 4.11 contacts (standard error, SE 0.41) in the household setting, 11.41 contacts (SE 0.27) in schools, 8.07 contacts (SE 0.52) in workplaces, and 2.79 contacts (SE 0.48) for the general community setting. It is worth remarking that the estimated weight for household contacts is larger than the average household size. This likely reflects the definition of contacts at home (rather than with household members) used in the POLYMOD study[14] that has been used to calibrate the weights. The rationale for using the POLYMOD and the Russian studies[14,18] in estimating the weights used to assemble the setting-specific synthetic matrices lies in the extensive validation of those contact patterns in epidemiological studies of a set of airborne infectious diseases, including influenza[29,35–39].

Our approach provides overall best matching ω_k and that, in principle, some of the differences in the social behavior of specific countries may not be captured by this approach. For this reason, as a validation of this calibration method, in Fig. 3a we report the correlation between the resulting synthetic matrices for France, Japan, and the Shanghai Province of China and the available empirical matrices for these additional locations[30–32]. We find significant (P value < 0.001) Pearson correlations of 0.92, 0.9, and 0.8 for France, Japan, and Shanghai Province, respectively.

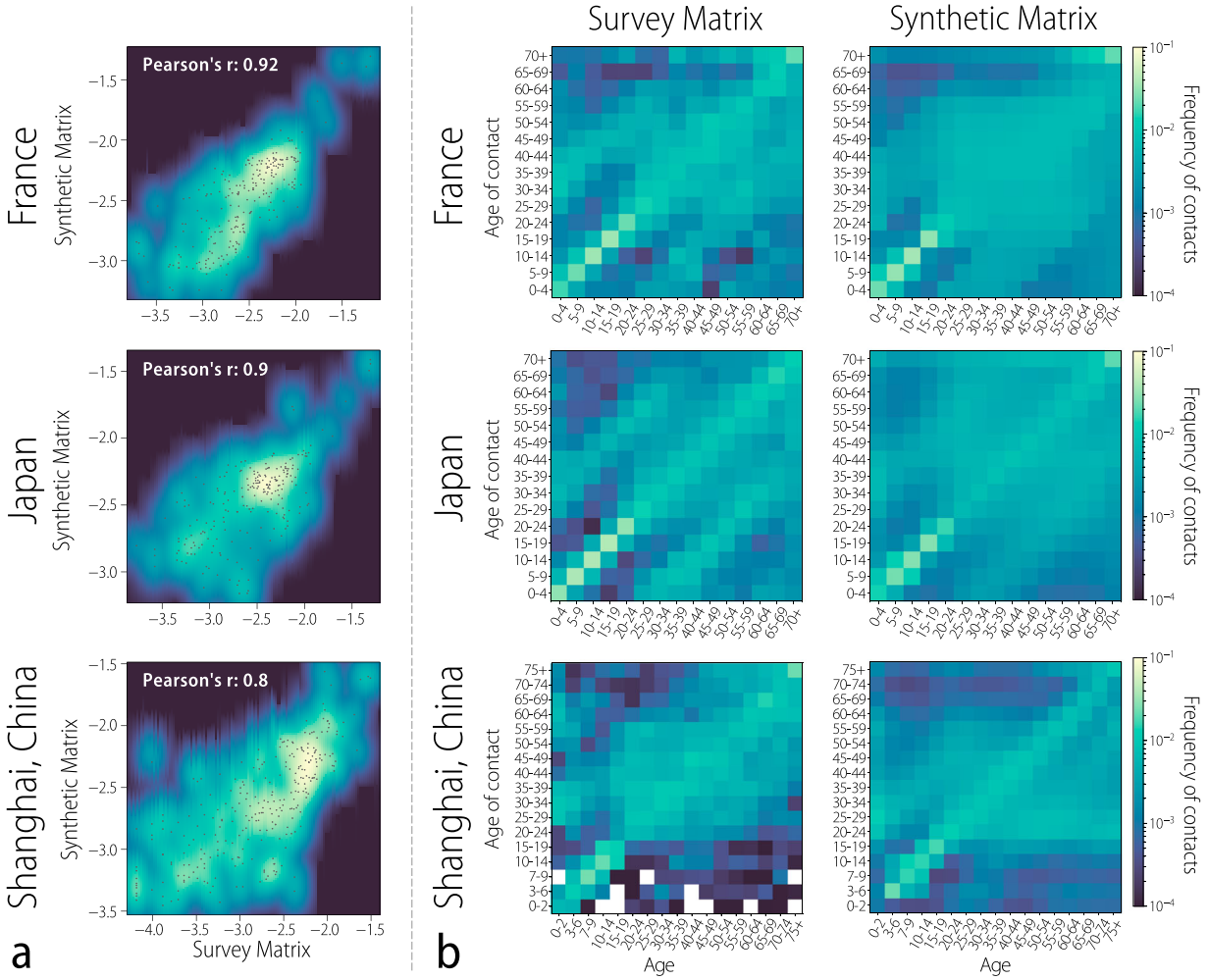

Fig. 3 Comparison to out-of-sample survey matrices. a Density plots showing the correlation of survey-based contact matrices for three out-of-sample locations (France, Japan, and the Shanghai Province of China) and their respective synthetic contact matrices (all normalized to sum to one). The points represent the actual values of the survey and synthetic contact matrices. The linear correlation between the elements of each survey matrix and the corresponding elements of the synthetic matrix is reported in terms of the Pearson correlation coefficient, whose values are reported in each plot. **b** Heatmaps representing the normalized survey matrices and the normalized overall synthetic matrices for France, Japan, and the Shanghai Province of China.

Moreover, we use the Canberra distance as a measure of the similarity between two contact matrices[23] (see "Methods" for the definition of the Canberra distance). We estimate the distance between the seven survey-based matrices used in the calibration phase and their respective synthetic matrices to be 0.21 on average (range: 0.17–0.28). (Note that the resulting Canberra distance is normalized by the square of the number of elements of the contact matrix to account for the different number of age groups considered by the different diary-based contact surveys). When considering the three locations used as out-of-sample validation, we estimate a slightly larger average distance of 0.29 (range: 0.21–0.37), suggesting the adequacy of the employed methodology. Finally, Fig. 3b shows a visual comparison between the synthetic and survey matrices, which highlights that the synthetic contact matrices are able to capture the specific features of each location such as contact patterns at school and the relative intensity of the main diagonals.

Figure 4a shows the synthetic overall contact matrices for China, the United States, and India. The contact matrices for all locations share many similarities: bands of increased contact along the main and off diagonals reflect the familiar household contact patterns, increased contact between adults age 20 and ~65 years old account for the interactions between the population's workforce, and the dominant contact patterns in the lower left of the contact matrices reflect the high number of interactions between school-aged individuals. Depending on the age structure of the population, the intensity of interactions occurring in the school setting can vary; however, this feature consistently dominates the contact matrix for all locations in our study.

To quantify the similarity between the overall contact matrices in different locations, we use a hierarchical clustering algorithm based on the Canberra distance to identify clusters of locations (dis)similar to each other[23]. We find that locations tend to cluster together by country (Fig. 5a), indicating that overall the contact patterns within a single country are more similar to each other than to the patterns observed in other countries. Strikingly, though not surprisingly, locations within developed countries such as Australia, Canada, and the United States are similar to each other and are clustered together, while at the same time locations throughout India, South Africa, and the North Caucasus region of Russia also cluster together, indicating a similarity in patterns between locations in the developing and transition world. Interestingly, a few territories of Canada, Russia, and India are outliers, indicating that the contact patterns in these locations

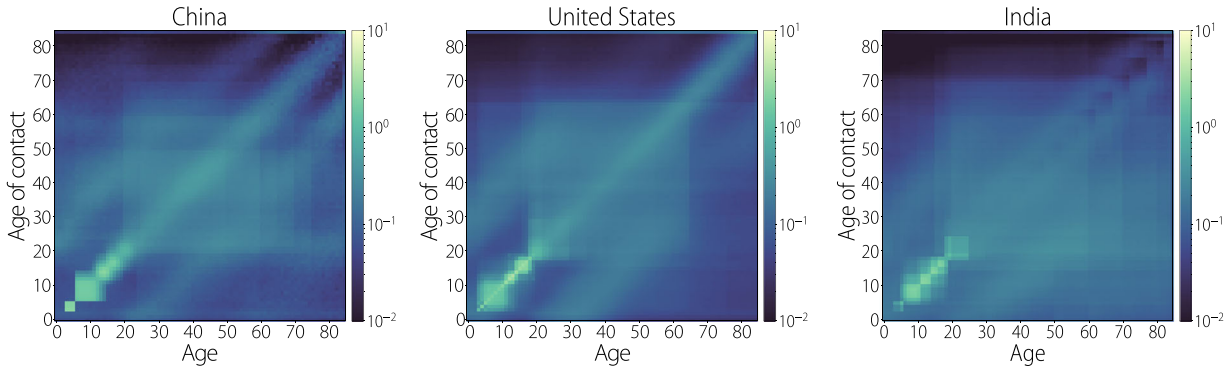

Fig. 4 Overall contact matrices. Each heatmap represents the overall average number of contacts relevant for airborne infectious disease transmission by age at the national level for China, the United States, and India.

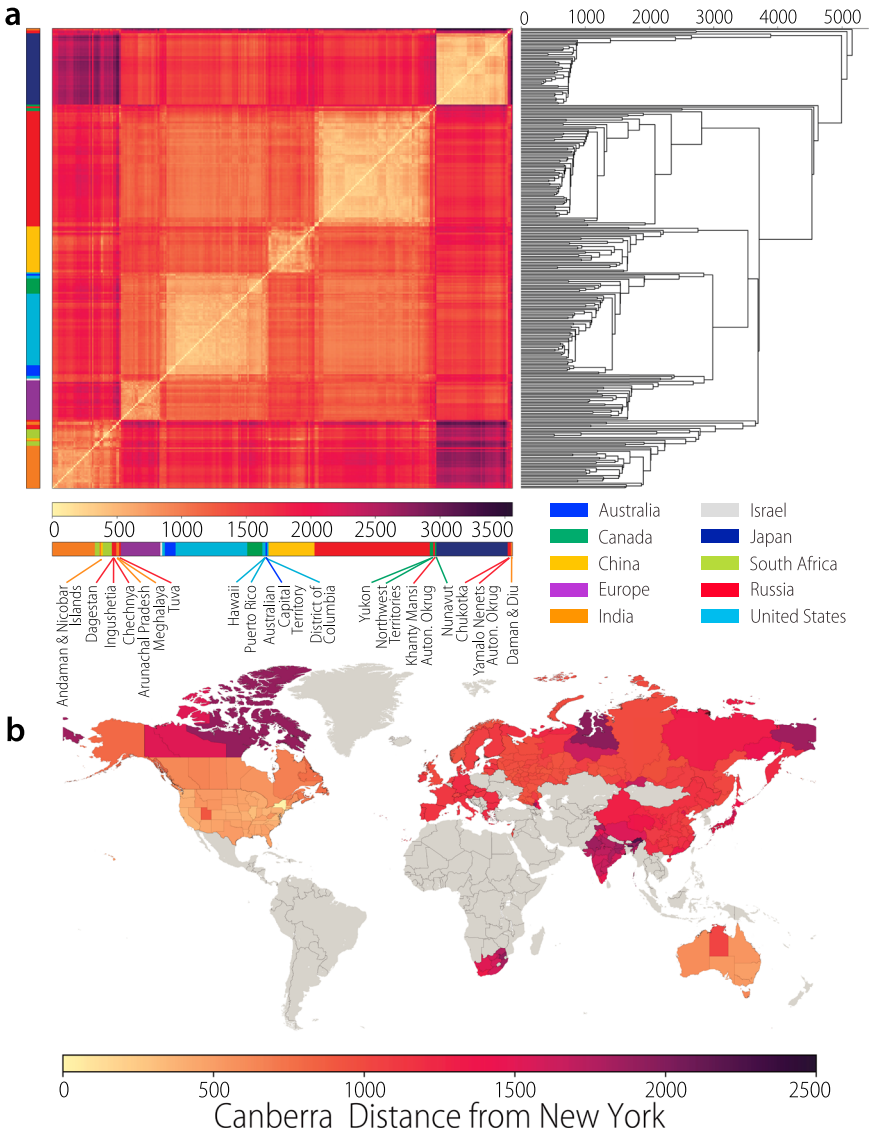

Fig. 5 Clustering of contact matrices. a Clustered matrix of the Canberra distance between subnational contact matrices and associated dendrogram using hierarchical clustering to organize subnational locations. Lighter colors indicate locations more similar to each other (distance closer to 0). **b** World map of the subnational level where colors represent the Canberra distance between each subnational location and the US state of New York (used as a reference point). The gray color means that no data is available. Note that the country of Israel is treated at the national level, rather than the subnational level, due to both its relatively small population and area, and the resolution of data available for reconstruction.

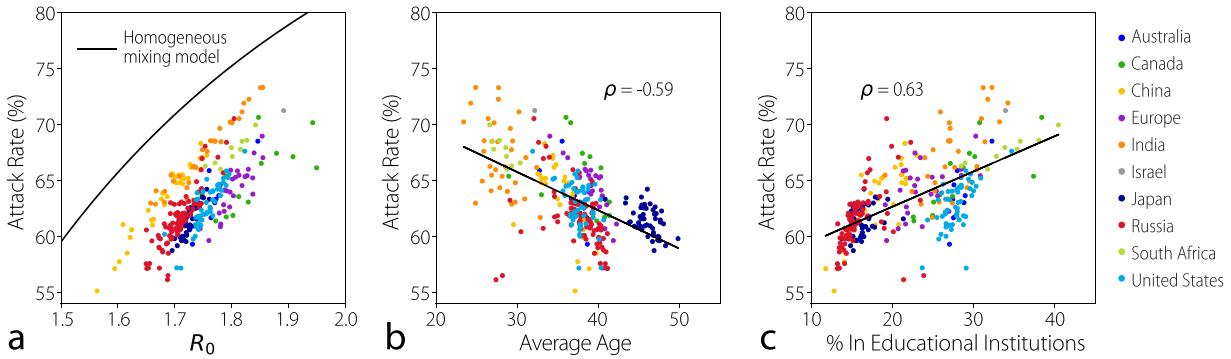

Fig. 6 Epidemic impact. a Scatter plot of the attack rate and the reproduction number R_0 from an age-structured SIR model using the contact matrix for each subnational location. European countries are included. The black line shows the results of the classic homogeneous mixing SIR model (no age groups). **b** Scatter plot of attack rates and the average age in each location. The black line represents the best-fitting linear model demonstrating a negative linear correlation between attack rates and the average age of the population. **c** Scatter plot of attack rates and percentage of the population attending educational institutions in each location. The black line represents the best-fitting linear model.

are different from what is observed in all other locations (including their respective countries). A more detailed discussion is reported in Supplementary Information. If we consider the US state of New York as a reference and compute the distance from all other locations to it, a geographical pattern clearly emerges (Fig. 5b). Indeed, the contact patterns in most states of the US, and the urbanized areas of Canada and Australia appear to be very closely related to the one inferred for New York. In contrast, most of India, South Africa, and of the territories in Canada, Russia, and Australia have contact patterns noticeably different from those obtained for the state of New York.

Epidemiological relevance. To investigate the effect of the computed contact matrices on infection transmission dynamics, we develop an age-structured SIR model to describe influenza transmission dynamics in the sites considered. The SIR model describes the spread of influenza in terms of the transition of individuals between different epidemiological compartments. Susceptible individuals (i.e., those at risk of acquiring the infection—S) can become infectious (i.e., capable to transmit the infection—I) after coming into contact with infectious individuals. Subsequently, infectious individuals recover from the infection and become removed (R) after a certain amount of time (the infectious period). In an age-structured implementation of the model, individuals are now identified also by their age, and the contact matrix is introduced to describe the number of contacts between susceptible individuals of age i and all of their possible infectious contacts of age j[2,13] (see "Methods" for details). More specifically, we considered a transmission model with identical disease parameters across geographical locations considered in the study. The contact matrices are thus the only factor driving the difference in dynamics and attack rate (total number of infected individuals) of the simulated epidemic.

Compared to the case of homogeneous mixing, where all individuals are assumed to be in contact with each other in equal proportions, the inclusion of the contact matrices in the epidemic model consistently yields a lower overall attack rate for all locations (Fig. 6a). This difference is also reflected in the strong variability of the basic reproduction number R_0, representing the number of cases generated by a typical index case in a fully susceptible population, which depends on the spectral radius of the matrix M as well as population structure (see Supplementary Information). To provide further validation of the adequacy of the matrices in characterizing the specific dynamics of influenza transmission in the Supplementary Information, we report the simulations of the age-structured SIR model calibrated on real data from the H1N1 influenza pandemic in multiple locations. The model adequately reproduces the age-specific seroprevalence profiles in Israel, Italy, Japan, UK, and USA[40–44].

To understand the underlying factors of the observed heterogeneities across geographical locations, we use a linear regression model to compare the attack rates and various socio-demographic features of each location (see Supplementary Information). We identified two socio-demographic features that correlate strongly with the attack rate: the average age of the population (Fig. 6b) and the fraction of the population in the educational system including instructors (Fig. 6c). Indeed, if we examine the attack rates by age and setting (see Supplementary Information), we observe that the greatest proportion of infections occur as a result of contact due to the school setting, and that attack rates, in general, are highest for school-aged individuals. Going further, an inspection of the incidence profile by age (see Supplementary Information) also clearly shows that individuals with high contact frequencies with others in the school setting are infected earlier in higher proportions. These results mirror well-known influenza spreading trends/patterns observed in the real world[14,23]. The observed results are robust (although with quantitative differences) to changes in transmissibility patterns and susceptibility to infection by age (see Supplementary Information). Taken together, our results suggest that developing countries with younger populations, and thus more school-aged individuals, are likely to experience higher overall attack rates when compared to older, developed countries.

We can also investigate how the attack rate and R_0 for each location would differ if we only had knowledge of the contact patterns at the national level. In this scenario, we use the country-level influenza transmission contact matrices in each location (note that each location is still characterized by its own specific age structure) and compare the results with those obtained by using the location-specific contact matrices everything else kept identical for the disease transmission model (Fig. 7a–c). By using the country-level matrix, we observe a much lower variability than by using location-specific mixing patterns. Moreover, location-specific attack rates and R_0 show a nonlinear relation with the results obtained using country-level contact patterns. Interestingly, we can observe clear geographical trends in the percent difference in attack rate using location-specific contact patterns in comparison to the corresponding country-level ones. For instance in much of the western area of China where most of the nation's ethnic minorities live, using the average matrix would lead to underestimating the final impact of an epidemic, while we

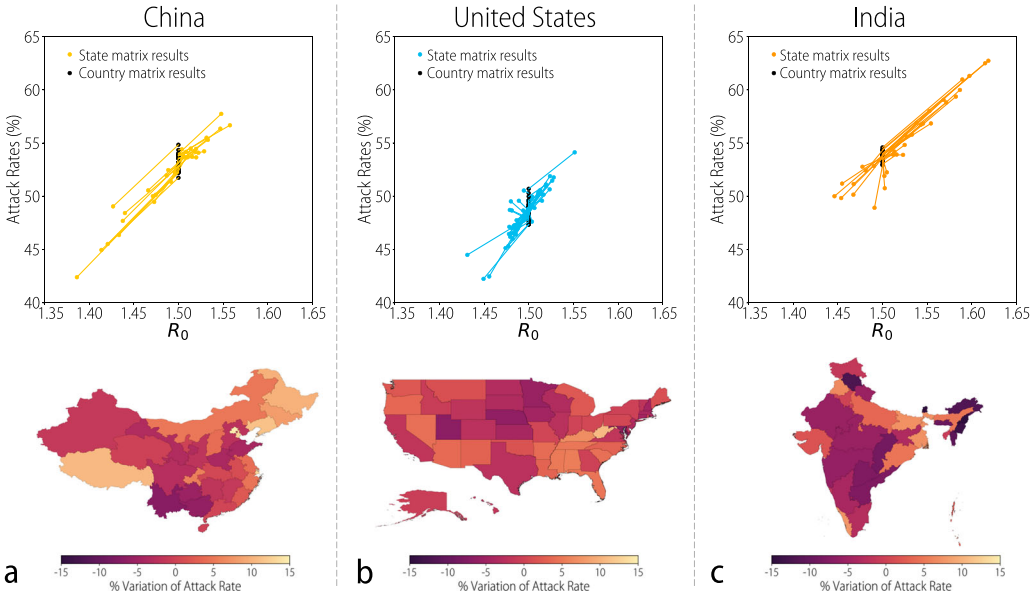

Fig. 7 Subnational heterogeneity. a The black dots represent the estimated attack rates in each province of China by using the country-level contact matrix and the location-specific age structure of the population. Colored dots represent the estimated attack rates in each location by using both the location-specific contact matrix and the age structure of the population. The colored lines connect the two estimated values of attack rate for each location. The transmission rate is set such that $R_0 = 1.5$ when using the country-level matrix. Each map shows the percentage variation of the attack rate using the location-specific contact matrix with respect to using the national contact matrix as a proxy for the subnational contact patterns (i.e., $(AR_c - AR_l)/AR_c$, where AR_c is the attack rate estimated by using the country-level contact matrix, and AR_l is that estimated by using the location-specific matrix). Colors toward Astra in the color scale indicate an overestimation of the attack rate in the location when using the country-level contact matrix as a proxy for the subnational contact patterns. Conversely, colors towards grape in the color scale indicate an underestimation of the attack rate in the location when using the country-level matrix as a proxy for the subnational contact patterns. **b** Same as **a**, but for the USA. **c** Same as **a**, but for India.

would overestimate it in the more traditionally urbanized/industrialized areas in the north-east of the country, such as Beijing and Shanghai (Fig. 7a).

Discussion

We have presented a general framework for the synthetic generation of age-stratified mixing patterns in key social settings (the household, school, workplace) for the transmission of airborne infectious diseases. The contact patterns we derived are not directly measured via survey or other direct methods (e.g., wearable sensors). Rather, we infer these age-based relationships between individuals by measuring them in synthetic populations developed using a novel approach that combines macro- and microdata available from public sources. While this is a limitation as, in general, a direct measure is preferred with respect to a derived one, this approach allows us to: (i) be flexible in the definition of effective contacts and thus to adapt our methodology to the study of different infectious diseases which require alternative definitions of "effective contact for transmission"; and (ii) focus on broad arrays of countries for which a direct measure is not available, especially at the subnational scale.

The use of age-mixing patterns in age-structured epidemic models provides insight into the epidemiology and dynamics of infectious diseases both within and between different countries around the world, as we have shown for the case of influenza. Our approach allows the integration of contact patterns that vary according to the geographical scale, the disease under consideration, and the detailed socioeconomic and demographic characteristics of the population. The developed method can be adapted to different geographic scales, conditional on the presence of sufficient data on age-specific intra-household, school, and work interdependencies. However, it is important to remark that, even if data availability allows the development of micro-level (e.g., zip-code, census block) synthetic populations, focusing on the geographic units smaller than commuting distances would break down the representativeness of age-mixing patterns to model the spread of an epidemic.

The use of data-driven heterogeneous mixing patterns, especially at the subnational level, opens up the door to potential applications in the more realistic modeling of the worldwide circulation of pathogens with epidemic/pandemic potential. The developed contact matrices also allow the study of the impact on the epidemiology of infectious diseases of socioeconomic disparities and demographic peculiarities (e.g., one-child policy). Eventually, by making all of the derived mixing patterns (in the form of readily usable contact matrices by age) publicly available, the presented results may benefit the research community actively working on the development of infectious disease forecasting approaches and mathematical models in support of the public health decision-making processes.

Methods

Development of the synthetic populations. To construct synthetic populations in different countries, we made use of a wide array of data sources (see Supplementary Information). These data provide distributions of key socio-demographic characteristics, such as the age structure, household size, age of the head of the household, age gaps between household members, household composition, employment rates, the educational system, and enrollment rates, etc. Distributions such as these are typically available either as macro-level data from census databases and other governmental sources, or as micro-level data coming from surveys conducted on a sample of the population. Census databases routinely provide information at a broader scope such as the age structure of a population, or the fertility rates; however, they often lack more detailed information related to the household composition and age relationships between household members. For this, we rely on micro-level surveys which collect the data at the household and individual level and ask participants for information in regards to their health, household condition and composition, economic conditions, and more. The kind of data available also varies by country and even at the subnational level, thus necessitating the development of adaptive algorithms that can take in the available data and accommodate for variability in data organization to produce a faithful

reconstruction of each population. With this in mind, the procedure implemented can be summarized as follows.

The first step in the reconstruction of a real-world population is the generation of households. In this process, we use two types of multinomial sampling. The first is based on the probability distribution $\mathcal{M}(y)$ of an independent socio-demographic characteristic y. For instance, such characteristic y can be the household size or composition, depending on the data available. The second type of multinomial sampling is based on the probability $\mathcal{M}(x|y_1 = i_1, y_2 = i_2, ..., y_n = i_n)$ of characteristic x conditional on the value i of a previously determined variable(s) y. In this case, x and y are assumed (when supported by available data) to have bivariate or multivariate joint distributions. Typically, the larger the number of joint distributions incorporated, the more precise the reconstruction of the real-world population. The precision of such a reconstruction is, however, often limited by the scope of the data (such as the survey sample size for each characteristic y) and its availability. For example, of the multinomial joint distributions used here, one is the distribution of the age of the head of the household by the size of the household and the household composition (whether a couple, a single parent with children, siblings, multigenerational families, etc.). The bivariate joint distributions incorporated is considerably long and includes (but is not limited to) distributions of the age of household members by the age of the head of the household, the age gap between couples living together by the age of one in the pair, the mother's age at childbirth by the age of the child, the number of household members by their relation to the age of the household head (such as a spouse, parent, child, grandchild, sibling, in-law, etc.) by the age of the household head and the household composition. These joint distributions were either found in the macro data or estimated from the micro survey data. Characteristics of the resulting synthetic households are compared to the distributions of the summary statistics available from the macro-level data using a goodness-of-fit test at the desired level of significance (generally 5%).

A similar procedure is used to assign those individuals to their respective schools and workplaces based on enrollment and employment records. These records detail the enrollment and employment rates by age, institutional sizes, and their age structures, as well as the student-to-teacher ratios in the case of schools. A more detailed explanation of the construction of the synthetic population can be found in Supplementary Information together with the results of the comparison between the synthetic and actual population statistics.

Construction of age-based contact matrices. We use synthetic contact networks to infer average age-based contact patterns within each social setting. For each location, these age-based contact patterns are encoded in a contact matrix F^k, whose elements F^k_{ij} describes the average frequency of contact between a given individual of age i and individuals of age j in setting k. We focus on 4 social settings: the households (H), schools (S), workplaces (W), and the general community (C). Specifically, here we adopt the frequency-dependent (mass action) transmission model, with the implicit assumption that an increased population density has no effect on the per capita contact rate between individuals[45]. This choice of modeling mechanism was already proved to represent a good approximation for the description of the transmission patterns of several infectious diseases[2]. Moreover, it allows us to readily compare epidemiological parameters between social settings and locations with disparate population density, and thus makes for an appropriate framework when modeling the transmission dynamics of heterogeneous populations around the world. The calculation of the contact matrices can be described as follows.

First, we compute the relative abundance of contacts between individuals of age i and individuals of age j in each configuration s of the setting k, $\Gamma^{k(s)}_{ij}$.

$$\Gamma^{k(s)}_{ij} = \frac{\phi^{k(s)}_i(\phi^{k(s)}_j - \delta_{ij})}{\nu^{k(s)} - 1}, \tag{2}$$

where $\phi^{k(s)}_i$ is the number of individuals of age i in the configuration s (i.e., a specific household, school, or workplace) of setting k; δ_{ij} is the Kronecker delta function, which we use to omit the individual i from their own set of contacts; $\nu^{k(s)}$ is the number of individuals (of all ages) in instance s of setting k. Note that to compute $\Gamma^{k(s)}_{ij}$, we assume homogeneous mixing within each configuration of the setting, i.e., each individual can be in contact with other individuals, and as a result the matrix $\Gamma^{k(s)}_{ij}$ has the expected symmetric property $\Gamma^{k(s)}_{ij} = \Gamma^{k(s)}_{ji}$.

Second, we compute the per capita probability of contact of an individual of age i with an individual of age j in setting k as F^k_{ij}.

$$F^k_{ij} = \sum_{\{s:\nu^{k(s)}>1\}} \Gamma^{k(s)}_{ij}/N_i, \tag{3}$$

where N_i is the total number of individuals of age i. Note that matrix F^k (i.e., the matrix of elements F^k_{ij}) is not symmetric.

Third, we combine the setting-specific contact matrices by age F^k to derive a matrix of the overall contacts by age M. We propose a weighted linear combination of the derived matrices in the four focus settings, calibrated to match the empirically estimated contact matrices from two contact diary survey studies in seven locations throughout Western Europe and Russia[14,18]. We perform a multiple linear regression to calibrate the weights of the synthetic setting contact matrices such that their linear combination matches the overall contact matrix for all seven locations coming from the survey studies (see Supplementary Information for details and for a comparison between the empirical and synthetic contact matrices).

Following this approach, we are also able to evaluate the uncertainty of point estimates of the contact matrices. While the absolute level of uncertainty results to be negligible if compared to the differences between the age groups, if synthetic individuals are sampled from the synthetic population, the level of introduced uncertainty becomes comparable with the one for diary-based contact studies (see Supplementary Information).

The average number of contacts. The average number of contacts $\langle c\rangle$ can be computed as

$$\langle c\rangle = \frac{1}{N}\sum_i N_i \sum_j M_{ij} \tag{4}$$

where $N = \sum_i N_i$ is the total number of individuals in the population.

Therefore,

$$\begin{aligned}
\langle c\rangle &= \frac{1}{N}\sum_i N_i \sum_j \sum_k \omega_k F^k_{ij} \\
&= \frac{1}{N}\sum_i N_i \sum_j \sum_k \omega_k \sum_s \Gamma^{k(s)}_{ij}/N_i \\
&= \frac{1}{N}\sum_k \omega_k \sum_i \sum_j \sum_s N_i \Gamma^{k(s)}_{ij}/N_i \\
&= \frac{1}{N}\sum_k \omega_k \sum_i \sum_j \sum_s \Gamma^{k(s)}_{ij} \\
&= \sum_k \omega_k \sum_s \sum_i \sum_j \Gamma^{k(s)}_{ij}/N \\
&= \sum_k \omega_k Z_k/N,
\end{aligned} \tag{5}$$

where Z_k is the number of individuals having at least one contact in setting k. Note that in the calculation, we used the symmetric property of matrix $\Gamma^{k(s)}_{ij}$. This expression provides a relation between the parameters ω_k and the overall per capita contact of relevance in epidemiological studies.

Canberra distance. To make side-by-side comparisons of the inferred contact matrices by age, we use the Canberra distance[23]. Specifically, each matrix is treated as a vector on which the Canberra distance is defined as

$$d(x,y)_{Canberra} = \sum_i \begin{cases} \frac{|x_i - y_i|}{|x_i| + |y_i|} & \text{for } x_i, y_i \neq 0 \\ 1 & \text{for } x_i, y_i = 0 \end{cases} \tag{6}$$

This yields a distance value of 0 for two locations with identical contact matrices, and increasingly larger distance values for two locations with increasingly different contact matrices.

Age-structured disease transmission model. For each location l, the transmission dynamics of influenza are modeled through an age-structured SIR model, where the mixing patterns are defined by the contact matrix previously introduced, M_{ij}.

The model is defined by the following set of equations:

$$\begin{aligned}
\dot{S}_i &= -\lambda_i S_i \\
\dot{I}_i &= \lambda_i S_i - \gamma I_i \\
\dot{R}_i &= \gamma I_i,
\end{aligned} \tag{7}$$

where S_i is the number of susceptible individuals of age i, I_i is the number of infected individuals of age i, R_i is the number of recovered or removed individuals of age i; γ^{-1} is the infectious periods (which corresponds to the generation time in the simple SIR model[46,47]), which is set to 2.6 days[48]; and λ_i represents the force of infection to which an individual of age i is exposed to other infected individuals and expressed as

$$\lambda_i = \beta \sum_j M_{ij} \frac{I_j}{N_j}, \tag{8}$$

where β is the transmissibility of the infection, N_i is the total number of individuals of age i, and M_{ij} measures the average number of contacts for an individual of age i with all of their contacts of age j.

The basic reproduction number R_0, representing the number of cases generated by a typical index case in a fully susceptible population, can be defined for this

Table 1 Data sources for each country and the country code.

Country	Country code	Sources
Australia	AUS	Australian Bureau of Statistics[51]
Canada	CAN	Statistics Canada[52]
		BC Stats[53]
		Finding Quality Childcare: A guide for parents in Canada[54]
China	CHN	China Health and Nutrition Survey[55]
		China Census 2010[56]
		China Statistical Yearbook[57]
India	IND	The 15th Indian Census[58]
		Demographic and Health Surveys (2005)[59]
		Unified District Information System for Education[60,61]
		All India Survey on Higher Education[62]
Israel	ISR	Israel Census 2008[63]
Japan	JPN	Official Statistics of Japan[64]
Russia	RUS	Russia Longitudinal Monitoring Survey[65]
		2010 All-Russian Population Census[66]
		Federal State Statistics Service[67]
South Africa	ZAF	Statistics South Africa[68]
		Statistics on Post-School Education and Training in South Africa[69]
		World Health Survey (2003)[70]
		South African Revenue Service[71]
United States of America	USA	Decennial Census of Population and Housing[72]
		Current Population Survey[73]
		American Community Survey[74]
		IPUMS USA[75]

model as

$$R_0 = \frac{\beta}{\gamma}\rho(M), \tag{9}$$

where $\rho(M)$ is the dominant eigenvalue of the matrix M[49].

To build the synthetics populations, we use publicly available databases listed in Table 1.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A database containing the inferred setting-specific matrices as well as the contact matrices for influenza transmission for all locations (and countries) is publicly available on the dedicated online repository: https://github.com/mobs-lab/mixing-patterns[50]. Python and R routines to work with the contact matrices and examples of how to use them in age-structured compartmental models are also available.

Code availability

The code can be publicly accessed at the dedicated online repository https://github.com/mobs-lab/mixing-patterns[50].

Received: 19 February 2020; Accepted: 8 December 2020;
Published online: 12 January 2021

References

1. Van Kerkhove, M. D. & Ferguson, N. M. Epidemic and intervention modelling: a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic. *Bull World Health Organ* **90**, 306–310 (2012).
2. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, UK, 1991).
3. Metcalf, C. J. E., Edmunds, W. J. & Lessler, J. Six challenges in modelling for public health policy. *Epidemics* **10**, 93–96 (2015).
4. Keeling, M., Woolhouse, M., May, R., Davies, G. & Grenfell, B. T. Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142 (2003).
5. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J. & Vespignani, A. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Med.* **4**, e13 (2007).
6. Longini, I. M. et al. Containing pandemic influenza at the source. *Science* **309**, 1083–1087 (2005).
7. Ferguson, N. M. et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**, 209–214 (2005).
8. Merler, S. & Ajelli, M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc. R Soc. B* **277**, 557–565 (2010).
9. Fenton, K. A. et al. Sexual behaviour in Britain: reported sexually transmitted infections and prevalent genital *Chlamydia trachomatis* infection. *The Lancet* **358**, 1851–1854 (2001).
10. Dodd, P. J. et al. Age-and sex-specific social contact patterns and incidence of *Mycobacterium tuberculosis* infection. *Am. J. Epidemiol.* **183**, 156–166 (2016).
11. Zhang, J. et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481–1486 (2020).
12. Jarvis, C. I. et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med.* **18**, 124 (2020).
13. Wallinga, J., Teunis, P. & Kretzschmar, M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epidemiol.* **164**, 936–944 (2006).
14. Mossong, J. et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74 (2008).
15. Hens, N. et al. Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infect. Dis.* **9**, 5 (2009).
16. Horby, P. et al. Social contact patterns in Vietnam and implications for the control of infectious diseases. *PLoS ONE* **6**, e16965 (2011).
17. Read, J. M. et al. Social mixing patterns in rural and urban areas of southern China. *Proc. R Soc. B* **281**, 20140268 (2014).
18. Ajelli, M. & Litvinova, M. Estimating contact patterns relevant to the spread of infectious diseases in Russia. *J. Theor. Biol.* **419**, 1–7 (2017).
19. Melegaro, A. et al. Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe. *PLoS ONE* **12**, e0170459 (2017).
20. Cattuto, C. et al. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* **5**, e11596 (2010).
21. Kiti, M. C. et al. Quantifying social contacts in a household setting of rural kenya using wearable proximity sensors. *EPJ Data Sci.* **5**, 21 (2016).
22. Zagheni, E. et al. Using time-use data to parameterize models for the spread of close-contact infectious diseases. *Am. J. Epidemiol.* **168**, 1082–1090 (2008).
23. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A. & Merler, S. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Comput. Biol.* **8**, e1002673 (2012).
24. Grefenstette, J. J. et al. FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC Public Health* **13**, 940 (2013).

25. Gallagher, S., Richardson, L. F., Ventura, S. L. & Eddy, W. F. SPEW: synthetic populations and ecosystems of the world. *J. Comput. Graph Stat.* **27**, 773–784 (2018).
26. Iozzi, F. et al. Little Italy: an agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. *PLoS Comput. Biol.* **6**, e1001021 (2010).
27. De Cao, E., Zagheni, E., Manfredi, P. & Melegaro, A. The relative importance of frequency of contacts and duration of exposure for the spread of directly transmitted infections. *Biostatistics* **15**, 470–483 (2014).
28. Prem, K., Cook, A. R. & Jit, M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Comput. Biol.* **13**, e1005697 (2017).
29. Litvinova, M., Liu, Q.-H., Kulikov, E. S. & Ajelli, M. Reactive school closure weakens the network of social interactions and reduces the spread of influenza. *Proc. Natl Acad. Sci. USA* **116**, 13174–13181 (2019).
30. Béraud, G. et al. The French connection: the first large population-based contact survey in France relevant for the spread of infectious diseases. *PLoS ONE* **10**, e0133203 (2015).
31. Munasinghe, L., Asai, Y. & Nishiura, H. Quantifying heterogeneous contact patterns in Japan: a social contact survey. *Theor. Biol. Med. Model* **16**, 6 (2019).
32. Zhang, J. et al. Patterns of human social contact and contact with animals in Shanghai, China. *Sci. Rep.* **9**, 1–11 (2019).
33. Ajelli, M., Poletti, P., Melegaro, A. & Merler, S. The role of different social contexts in shaping influenza transmission during the 2009 pandemic. *Sci. Rep.* **4**, 7218 (2014).
34. Ajelli, M. & Merler, S. The impact of the unstructured contacts component in influenza pandemic modeling. *PLoS ONE* **3**, e1519 (2008).
35. Kretzschmar, M., Teunis, P. F. & Pebody, R. G. Incidence and reproduction numbers of pertussis: estimates from serological and social contact data in five European countries. *PLoS Med.* **7**, e1000291 (2010).
36. Kucharski, A. J. & Gog, J. R. The role of social contacts and original antigenic sin in shaping the age pattern of immunity to seasonal influenza. *PLoS Comput. Biol.* **8**, e1002741 (2012).
37. Poletti, P. et al. Perspectives on the impact of varicella immunization on herpes zoster. A model-based evaluation from three European countries. *PLoS ONE* **8**, e60732 (2013).
38. Merler, S. & Ajelli, M. Deciphering the relative weights of demographic transition and vaccination in the decrease of measles incidence in Italy. *P. Roy. Soc. B* **281**, 20132676 (2014).
39. Kretzschmar, M. E. et al. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health* **5**, e452–e459 (2020).
40. Weil, M. et al. The dynamics of infection and the persistence of immunity to A (H1N1) pdm09 virus in Israel. *Influenza Other Respir Viruses* **7**, 838–846 (2013).
41. Merler, S. et al. Pandemic influenza A/H1N1pdm in Italy: age, risk and population susceptibility. *PLoS ONE* **8**, e74785 (2013).
42. Japanese Infectious Disease Surveillance Center. Influenza antibody holding status survey in FY 2010 - First Report (2010). https://www.niid.go.jp/niid/en/idsc-e.html (2018).
43. Hardelid, P. et al. Assessment of baseline age-specific antibody prevalence and incidence of infection to novel influenza A/H1N1 2009. *Health Technol Assess* **14**, 115–92 (2010).
44. Reed, C., Katz, J. M., Hancock, K., Balish, A. & Fry, A. M. Prevalence of seropositivity to pandemic influenza A/H1N1 virus in the United States following the 2009 pandemic. *PLoS ONE* **7**, e48187 (2012).
45. Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
46. Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R Soc. B* **274**, 599–604 (2006).
47. Liu, Q.-H. et al. Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl Acad. Sci. USA* **115**, 12680–12685 (2018).
48. Vink, M. A., Bootsma, M. C. J. & Wallinga, J. Serial intervals of respiratory infectious diseases: a systematic review and analysis. *Am. J. Epidemiol.* **180**, 865–875 (2014).
49. Diekmann, O., Heesterbeek, J. A. P. & Metz, J. A. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J. Math Biol.* **28**, 365–382 (1990).
50. Mistry, D. et al. Inferring high-resolution human mixing patterns for disease modeling. https://doi.org/10.5281/zenodo.4287574 (2020).
51. Australian Bureau of Statistics. Australian Bureau of Statistics. http://www.abs.gov.au/ (2011).
52. Statistics Canada. Statistics Canada. https://www.statcan.gc.ca/eng/start (2011).
53. Government of British Columbia. BC Stats. https://www2.gov.bc.ca/ (2011).
54. Childcare Resource and Research Unit, Canadian Union of Postal Workers. Finding Quality Child Care: A guide for parents in Canada. https://findingqualitychildcare.ca/ (2014).
55. National Institute for Nutrition and Health, China Center for Disease Control and Prevention, Carolina Population Center, University of North Carolina at Chapel Hill. China Health and Nutrition Survey (CHNS). https://www.cpc.unc.edu/projects/china/data/datasets (2009).
56. China Statistics Press. Census 2010. http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm (2010).
57. China Statistics Press. China Statistical Yearbook 2010. http://www.stats.gov.cn/tjsj/ndsj/2010/indexeh.htm (2010).
58. Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India. 2011 Census Data. http://www.censusindia.gov.in/pca/Searchdata.aspx (2011).
59. Demographic and Health Surveys (2005–2006). India: Standard DHS 2005-06. https://dhsprogram.com/data/dataset/India_Standard-DHS_2006.cfm?flag=0 (2016).
60. Unified District Information System for Education (UDISE), National Institute of Educational Planning and Adminstration (2011–2012) Elementary Education in India. http://udise.in/src.htm (2016).
61. Unified District Information System for Education (UDISE), National Institute of Educational Planning and Adminstration (2012–2013) Secondary Education in India: State Report Cards 2012–13. http://udise.in/src.htm (2016).
62. Department of Higher Education, Ministry of Human Resource Development, Government of India. All India Survey on Higher Education. http://mhrd.gov.in/sites/upload_files/mhrd/files/statistics/AISHE2011-12P_1.pdf (2013).
63. Israel Central Bureau of Statistics. Israel Census 2008. http://www.cbs.gov.il/census/census/pnimi_page_e.html?id_topic=2 (2008).
64. Japanese Government Statistics. e-Stat, Portal Site of Official Statistics of Japan. https://www.e-stat.go.jp/en/ (2010).
65. Popkin, B. M. et al. The Russia Longitudinal Monitoring Survey (RLMS). https://www.cpc.unc.edu/projects/china/data/datasets (2010).
66. Federal State Statistics Service. 2010 All-Russian Population Census. http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm (2010).
67. Federal State Statistics Service. Labor market, employment and wages. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/wages/labour_force/# (2010).
68. Statistics South Africa. Census 2011. http://www.statssa.gov.za/ (2011).
69. Department: Higher Education and Training, Republic of South Africa. Statistics on Post-School Education and Training in South Africa: 2011. http://www.cbs.gov.il/census/census/pnimi_page_e.html?id_topic=2 (2008).
70. World Health Organization (WHO) (2003-2005) World Health Survey. http://apps.who.int/healthinfo/systems/surveydata/index.php/catalog (2017).
71. South African Revenue Service, National Treasury. 2013 Tax Statistics. http://www.sars.gov.za/About/SATaxSystem/Pages/Tax-Statistics.aspx (2013).
72. United States Census Bureau. Decennial Census of Population and Housing. https://www.census.gov/programs-surveys/decennial-census/decade.2010.html (2010).
73. United States Census Bureau. Current Population Survey. https://www.census.gov/programs-surveys/cps/data-detail.html (2010).
74. United States Census Bureau. American Community Survey. https://www.census.gov/programs-surveys/acs/data.html (2010).
75. Ruggles, Steven and Flood, Sarah and Goeken, Ronald and Grover, Josiah and Meyer, Erin and Pacas, Jose and Sobek, Matthew. IPUMS USA: Version 8.0 [dataset]. https://doi.org/10.18128/D010.V8.0 (2010).

Acknowledgements

A.P.y.P., M.C., K.M., X.X., M.E.H., I.M.L., and A.V. acknowledge funding from Models of Infectious Disease Agent Study, National Institute of General Medical Sciences Grant U54GM111274 and the CDC contract 75D30119C05765. Q.-H.L. has received funding from the National Natural Science Foundation of China (No. 62003230), the Fundamental Research Funds for the Central Universities (No. 1082204112289). The authors would like to thank Nicole Samay for her assistance in preparing the figures.

Author contributions

M.A. and A.V. designed the experiment. D.M., M.L., A.P.P., M.C., L.F., M.F.C.G., S.A.H., Q.-H.L., K.M., and X.X. collected and integrated the data. D.M., M.L., and M.A. carried out the analysis. D.M., M.L., A.P.P., M.C., L.F., M.F.C.G., S.A.H., Q.-H.L., K.M., X.X., M.E.H., I.M.L., S.M., M.A., and A.V. contributed to the writing and discussion of the paper.

Competing interests

A.P.y.P., M.C., and A.V. report grants from Metabiota Inc and M.A. reports research funding from Seqirus, outside the submitted work. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20544-y.

Correspondence and requests for materials should be addressed to M.A. or A.V.

Peer review information *Nature Communications* thanks Piero Manfredi and Lander Willem for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.