medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the COVID-19 era

Kiesha Prem[1,2], Kevin van Zandvoort[1], Petra Klepac[1], Rosalind M Eggo[1], Nicholas G Davies[1], Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Alex R Cook[2], Mark Jit[1]*

[1]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine
[2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore

*Corresponding author

Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group: Christopher I Jarvis, Quentin J Leclerc, Jon C Emery, Gwenan M Knight, Amy Gimma, Simon R Procter, Kathleen O'Reilly, Sophie R Meakin, Charlie Diamond, Stefan Flasche, Billy J Quilty, Anna M Foss, Thibaut Jombart, Katherine E. Atkins, Georgia R Gore-Langton, Adam J Kucharski, James W Rudge, Matthew Quaife, Arminder K Deol, Carl A B Pearson, C Julian Villabona-Arenas, Graham Medley, Alicia Rosello, Hamish P Gibbs, Samuel Clifford, Rein M G J Houben, David Simons, James D Munday, Megan Auzenbergs, Rachel Lowe, Joel Hellewell, Sam Abbott, Damien C Tully, Stéphane Hué, W John Edmunds, Yang Liu, Fiona Yueqian Sun, Oliver Brady, Sebastian Funk, Nikos I Bosse, Akira Endo, Timothy W Russell, Emily S Nightingale.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Abstract

Mathematical models have played a key role in understanding the spread of directly-transmissible infectious diseases such as Coronavirus Disease 2019 (COVID-19), as well as the effectiveness of public health responses. As the risk of contracting directly-transmitted infections depends on who interacts with whom, mathematical models often use contact matrices to characterise the spread of infectious pathogens. These contact matrices are usually generated from diary-based contact surveys. However, the majority of places in the world do not have representative empirical contact studies, so synthetic contact matrices have been constructed using more widely available setting-specific survey data on household, school, classroom, and workplace composition combined with empirical data on contact patterns in Europe. In 2017, the largest set of synthetic contact matrices to date were published for 152 geographical locations. In this study, we update these matrices with the most recent data and extend our analysis to 177 geographical locations. Due to the observed geographic differences within countries, we also quantify contact patterns in rural and urban settings where data is available. Further, we compare both the 2017 and 2020 synthetic matrices to out-of-sample empirically-constructed contact matrices, and explore the effects of using both the empirical and synthetic contact matrices when modelling physical distancing interventions for the COVID-19 pandemic. We found that the synthetic contact matrices reproduce the main traits of the contact patterns in the empirically-constructed contact matrices. Models parameterised with the empirical and synthetic matrices generated similar findings with few differences observed in age groups where the empirical matrices have missing or aggregated age groups. This finding means that synthetic contact matrices may be used in modelling outbreaks in settings for which empirical studies have yet to be conducted.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Author summary

The risk of contracting a directly transmitted infectious disease such as the Coronavirus Disease 2019 (COVID-19) depends on who interacts with whom. Such person-to-person interactions vary by age and locations—e.g., at home, at work, at school, or in the community—due to the different social structures. These social structures, and thus contact patterns, vary across and within countries. Although social contact patterns can be measured using contact surveys, the majority of countries around the world, particularly low- and middle-income countries, lack nationally representative contact surveys. A simple way to present contact data is to use matrices where the elements represent the rate of contact between subgroups such as age groups represented by the columns and rows. In 2017, we generated age- and location-specific synthetic contact matrices for 152 geographical regions by adapting contact pattern data from eight European countries using country-specific data on household size, school and workplace composition. We have now updated these matrices with the most recent data (Demographic Household Surveys, World Bank, UN Population Division) extending the coverage to 177 geographical locations, covering 97.2% of the world's population. We also quantified contact patterns in rural and urban settings. When compared to out-of-sample empirically-measured contact patterns, we found that the synthetic matrices reproduce the main features of these contact patterns.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Introduction

The emergence of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) responsible for causing Coronavirus Disease 2019 (COVID-19) has affected the lives of billions worldwide (1). SARS-CoV-2 is predominantly transmitted between people via respiratory droplets and, as such, the transmission dynamics are strongly influenced by the number and type of close contacts between infectious and susceptible individuals (2–7).

Mathematical models have played a key role in understanding both the spread of directly-transmissible infectious diseases such as COVID-19 (8–10) and the effectiveness of public health responses (11–16). Since transmission events can rarely be directly observed and measured, most transmission models are based on the *social contact hypothesis* (17) which implies the risk of transmission between a susceptible and an infected individual be proportional to the rate of contact between them (18). Rates of contact are known to differ according to characteristics such as the age, of both individuals, and the setting in which the contact takes place, such as the home, school or workplace; they are also commonly assortative, and infection may be concentrated in demographic segments as a result (17,19,20).

Age-structured models often define the rate of mixing between age groups through a *mixing matrix* where the elements represent the proclivity of contact between two individuals from subgroups (such as age groups) represented by the columns and rows. Mixing matrices can be generated from surveys that record the number and type of contacts between people, such as the respondent-completed diaries used in the landmark POLYMOD contact pattern study, which measured social contact patterns in eight European countries (20). However, the majority of countries around the world lack data from contact surveys that can be used to inform the mixing matrix. This problem is particularly acute in low- and lower-middle-income countries (LMICs), where only 4 studies are available, compared to 54 in high-income countries (21). Our previous work (22) used country-specific data on household size, school, and workplace composition plus empirical contact data from the POLYMOD survey to generate age- and location-specific contact matrices (synthetic contact

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

matrices) to use in settings where social contact patterns had not yet been directly measured.

These synthetic contact matrices have been widely used in models of SARS-CoV-2 spread and the impact of interventions such as physical distancing which alter the pattern of contacts (e.g. (13)). Following the publication of our previous work, new empirical contact surveys have been conducted in LMICs (reviewed in (21)), full demographic data are now available for more countries for older age groups, which is particularly salient given the age-gradient in the severity of COVID-19 (23,24), and more recent household composition data are now available for more countries than before. Updating the matrices is particularly important since public health interventions during the pandemic, such as shielding, are often age-structured (25).

Geographic differences within countries have also been observed, with large early outbreaks in urban population centres such as Wuhan, New York, London and Madrid (26,27) spreading into more rural areas, which in many countries may lack the healthcare infrastructure to handle surges in severe cases. Tailored public health response (e.g. for shielding policy) in rural and urban settings may thus be called for to minimise unnecessary economic and social impacts. Assessing such policies requires differences between contact patterns in rural and urban environments to be quantified, which has previously been done only for a few countries (28,29). In these studies, individuals in rural settings documented more contacts at home than their urban counterparts (28,29).

In this paper, we update the synthetic contact matrices with the most recent data, comparing them to measured contact matrices, and develop customised contact matrices for rural and urban settings. We use these to explore the effects of physical distancing interventions for the COVID-19 pandemic in a transmission model.

Materials and methods

Updating country-specific demography and setting parameters

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

As in Prem et al. (22), we employed a Bayesian hierarchical modelling framework to estimate the age- and location-specific contact rates in each of the POLYMOD countries (Belgium, Germany, Finland, United Kingdom, Italy, Luxembourg, the Netherlands, Poland), accounting for repeat measurements of contacts made in different locations by the same individual. We model the number of contacts documented by individual at a particular location with an individual in age group , as where the mean parameter varies for each individual , by 's age, , and by location, i.e: . The parameter characterises differences in social activity levels between individuals i.e., the random effect belonging to individual i. The parameter denotes the proclivity of contact between individuals from two age groups, and , at location and it is the key estimand. Because the number of contacts should be comparable for individuals of similar ages, we imposed spatial smoothness for the parameter. Noninformative prior distributions were assumed for all parameters in the model, as detailed in (22).

We updated the synthetic contact matrices (22) with more recent data on population age structure, household age structure of 43 countries with recent Demographic Household Surveys (DHS) (30) and socio-demographic factors for 177 geographical regions, including countries and some subnational regions such as the Hong Kong and Macau Special Administrative Regions (SARs) of the People's Republic of China. We include 14 country characteristics from the World Bank and United Nations Educational, Scientific and Cultural Organization Institute for Statistics (UIS) databases: gross domestic product per capita, total fertility rate and adolescent fertility rate, population density, population growth rate, internet penetration rate, secondary school education attainment levels, as proxies of development, and under-five mortality rate, the life expectancy of males and females, mortality rates of males, risk of maternal death, mortality from road traffic injury, and the incidence of tuberculosis, as proxies for overall health in the country. The DHS provides nationally-representative household surveys with the largest dataset, from India, containing information on ~ 3 million individuals from about 600 000 households (**Supplementary Table 1**). To project the household age structure for a geographical location with no available household data, we use a weighted mean of the population-adjusted household age structures of the POLYMOD and DHS countries

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

as described in the **Supplementary Materials**. Because the household age structures vary across countries in different stages of development and with different demographics, we use the updated 14 indicators, all standardized by z-scoring, to quantify the similarities between countries with and without household data to derive these weights. We internally validated the household age matrices using leave-one-out validation to verify these matrices describing household structure could be reverse-engineered for countries (POLYMOD and DHS) for which empirical household age matrices were available, as described in Prem et al. (22) and in the **Supplementary Materials**.

By accounting for the demographic structure, household structure (where known), and a variety of metrics including workforce participation and school enrolment, we then estimated contact patterns at home, work, school and other locations for non-POLYMOD countries. Specifically, the population age compositions for 177 geographical regions were obtained from the United Nations Population Division (31). To derive the working population matrices for each geographical location c, we use the labour force participation rate by sex and 5-year age groups, w_a^c, for the 177 geographical regions from the International Labour Organization (ILO) (32). We derive the working population matrix of a country $W_{a,\alpha}^c$ from the cross product of w_a^c and w_α^c, and the elements describe the probability of encounter between individuals from two age groups, a and α, in the workplace.

When constructing the school-going population matrices, we use the country-specific pupil-to-teacher ratio in schools at various level of education (i.e., pre-primary, primary, secondary and tertiary), enrolment rates of students at various level of education, starting ages and number of years of schooling at various level of education from UIS (33) and the distribution of teachers by age from the Organisation for Economic Co-operation and Development (OECD) (34). Using the country-specific data, we first estimate the number of students in each age group by education level. Together with the country-specific pupil-to-teacher ratio at each education level, distribution of teachers and workforce by age, we then project the number of teachers in each age group. Both students and teachers form the school-going population. Similar to the formulation of the working population matrix, the

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

school-going population matrix estimates the probability of an encounter between two ages. The steps to construct both the working and school-going populations are detailed in the **Supplementary Materials**.

After projecting populations at home, work and school for the 177 geographical regions, we infer the synthetic age- and location-specific contact matrices (**Supplementary Materials**). For contacts in other locations (not home, work or school), we adjusted the POLYMOD contact matrices with the country-specific population. We also compare the proportion of contacts at other locations measured from the empirical contact studies.

Stratifying contacts by rural and urban areas

We stratified the age- and location-specific contact matrices according to rural and urban areas by the rural and urban population age compositions for all geographical regions of the world from the United Nations Population Division (35) (see (36) for urban and rural classification). The nationally-representative DHS household surveys additionally provide data for rural and urban areas (30), allowing us to derive rural-urban household age matrices. We compare the population age compositions and household age matrices in rural and urban settings of countries with stratified household data (**Supplementary Materials**).

We assessed the age-specific labour force participation rates by rural and urban regions from ILO (37). Using the differences in rural and urban schools' pupil-to-teacher ratio from OECD (38), we construct rural and urban school population matrices. These differences were available for 36 countries, and we assumed the OECD average for the regions without data. We also compare the mean total number of contacts among children (0–9-year-olds) and older adults (60–69-year-olds), as well as the basic reproduction number in rural and urban settings.

Comparing synthetic matrices to empirical contact matrices

We extracted data from all contact surveys listed in the Zenodo social contact database (39) and directly from the published studies (40), and used them to

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

construct empirical contact matrices using the socialmixr R package (41). Data were available for 11 geographical locations: Shanghai and Hong Kong SAR, China (42,43), France (44), Kenya (45), Peru (46), the Russian Federation (47), South Africa (29,48), Vietnam (49), Zambia (29) and Zimbabwe (50). We then compared each element of the empirical matrices with our synthetic matrices. We also compared the proportion of contacts in "Other" locations, since this was the only setting not directly informed by local data (other than population age structure) in the synthetic matrices. To understand potential sources of differences between the empirical and synthetic matrices as well as between empirical matrices between different regions, we extracted details of how each survey was conducted from the original publications.

Table 1 summarises the changes between the construction of the 2017 and 2020 synthetic matrices. Analyses were done in R version 3.6.2 (51), and the codes are deposited in https://github.com/kieshaprem/synthetic-contact-matrices.

Impact on modelling of interventions

We compare the difference in relative reduction of COVID-19 cases between using the empirical and synthetic matrices in models of COVID-19 epidemics in ten geographical regions—China, France, Hong Kong SAR, Kenya, Peru, the Russian Federation, South Africa, Uganda, Vietnam and Zimbabwe—using an age-stratified compartmental model (13,25). We model an unmitigated epidemic and three intervention scenarios: 20% physical distancing, 50% physical distancing, and shielding. In all intervention scenarios, we assume a 50% reduction in transmission from individuals with clinical symptoms through self-isolation. In addition, we assume, the following: (i) 20% physical distancing: 20% reduction in transmission outside of the household, (ii) 50% physical distancing: 50% reduction in transmission outside of the household, (iii) shielding: older individuals (≥ 60-years-old) do not mix with other members of the population. Shielding is implemented as having 80% of the older individuals (≥ 60-years-old) shielded, and reducing their contact rates with unshielded and other shielded individuals by 80%. We considered six contact matrices when modelling the interventions to the COVID-19 pandemic: the empirically-constructed contact matrices at the study-year and adjusted for the 2020

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

population, the 2017 synthetic matrices, and the updated synthetic matrices at the national, rural, or urban settings.

Results

Twenty-five geographical regions were added to this study compared to the 2017 study. We also updated the population demographic data used for all countries including Namibia, Syrian Arab Republic, Republic of South Sudan, Kuwait, and Vanuatu where the proportion of individuals aged > 70 years was previously not recorded.

There were varied methods adopted in 11 contact surveys conducted to generate the empirical contact matrices covering 11 geographical locations (**Table 2**). The surveys differed substantially from each other and the original POLYMOD survey in sampling frames and survey methodology. Dodd et al. (29) measured social contacts among adults in South Africa and Zambia. Three surveys were conducted in exclusively rural regions (40,46,48), (including one in a remote highlands region (46)), three other surveys were conducted only in urban regions (42,43,47), and the remaining five surveys were conducted in a variety of urban and rural settings (29,44,45,49,50). Although most studies adopted random or stratified sampling to recruit their respondents, a handful included convenience (42,46,47) and quota (43,44) sampling methods in their recruitment. In most contact diary approaches, contacts are categorised as physical contacts (e.g., skin-to-skin contacts) and nonphysical contacts (e.g. two-way conversations with three or more words in the physical presence of another person) (20). They were equally split between studies that asked respondents to fill in surveys retrospectively (29,40,43,46,49) and prospectively (42,44,45,47,48,50).

The estimated proportions of contacts in other locations from POLYMOD contact survey largely match analogous figures in empirical contact studies from five geographical locations which report this—Shanghai and Hong Kong SAR, China; the Russian Federation; Peru; and Zimbabwe—but are higher than those from France for most ages (**Fig 1**). It is slightly higher in the synthetic matrices in adults (i.e., 20–40-year-olds) in Shanghai, Hong Kong and the Russian Federation, and slightly

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

lower in older individuals (i.e., >60-year-olds) in Peru, but all other ages match closely.

The pronounced diagonals observed in all contact matrices are matched in the synthetic matrices (**Fig 2A and 2B**), as are the secondary diagonals indicating the occurrence of intergenerational mixing. The updated synthetic contact matrices show close similarities to empirical matrices (median correlation between normalised synthetic and empirical matrices 0.83, interquartile range 0.67–0.85).

We reconstructed the empirical household age structures for the POLYMOD and DHS countries with high fidelity (median correlation between the observed and modelled HAM 0.92, with an interquartile range 0.85–0.95) (See **Supplementary Materials** Section 2.1 for details). The differences in the population and households age composition by rural and urban settings are presented in the **Supplementary Materials**.

For many of the low-income countries (LIC), a larger mean number of contacts among children (i.e., 0–9-years-old) were observed in rural settings than urban settings. However, in high-income countries (HIC), urban settings had a larger mean number of contacts among children (**Fig 3**). Among HIC, the basic reproduction number in rural and urban settings are positively correlated (r=0.64, 95% confidence interval: 0.44–0.78).

The choice of using synthetic or empirical matrices did not make a large difference to the overall number of severe COVID-19 cases predicted in a mathematical model of SARS-CoV-2 transmission and disease across the three physical distancing interventions (**Fig 4**). The relative magnitude of this discrepancy differed between countries. Differences were more marked in specific age groups (which varied by country). The largest age-related differences could potentially be attributed to particular features of empirical survey design such as missing (Peru, Russia) or aggregated (Kenya, South Africa, Uganda, Vietnam) age groups, mode of questionnaire chosen by participants (Hong Kong SAR) and survey administration during school holidays (Zimbabwe) (See **Supplementary Materials** Section 2.5 for details).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Discussion

Social mixing patterns have not been directly measured in most countries or regions within countries, particularly in low- and lower-middle-income settings. Synthetic contact matrices provide alternative age- and location-specific social mixing patterns for countries in different stages of sociodemographic and economic development (22). The synthetic contact matrices presented here were derived by the amalgamation of several data sources and methods: (i) integration into a Bayesian hierarchical framework of age- and location-specific contact rates from eight European countries from the POLYMOD contact study; (ii) construction of age-structured populations at home, work, and school in many non-POLYMOD countries by combining household age-structure data from the POLYMOD study and DHS (which include mostly data from lower-income countries), socio-demographic factors from the UN Population Division and various international indicators; and (iii) projection of age-structured populations at home, work, and school and age- and location-specific contact matrices to other non-POLYMOD and non-DHS countries. Both empirical and synthetic contact matrices capture age-assortativity in mixing patterns; the pronounced primary diagonal highlights that individuals interact with others of similar age. Both also show secondary diagonals, approximately one generation apart, indicating parent-child interactions.

This paper provides a substantial update and improvement to previous synthetic matrices published in 2017 (**Table 1**). Improvements in the availability of demographic data globally have enabled us to provide validated approximations to age- and location-specific contact rates for 177 geographical regions covering 97.2% of the world's population, compared to 152 geographical regions covering 95.9% previously. Household data from 34 additional LMICs were included in the revision. We have also used the most recent data to build the working and school-going populations. We have extended the method to project contact patterns in rural and urban settings using country-specific urban and rural data. We find a higher positive correlation in mean contact rates and basic reproduction number in rural and urban settings of HIC, owing to the smaller rural-urban differences in these countries. Moreover, when assessing the consistency of results under different mixing

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

assumptions (empirical and synthetic), we observed small differences in the modelled relative reduction in number of cases across the three physical distancing interventions for the COVID-19 pandemic.

The synthetic matrices provide consistency for inter-country comparisons since they are based on common datasets. This is challenging to achieve through empirical data collection (see Table 2). For such studies, surveying across the whole population poses several challenges. Establishing a sampling frame and obtaining a sample representative of an entire country's population is expensive and in some regions logistically challenging, so researchers often restrict studies to a particular subpopulation. For instance, many recent empirical contact studies only represent certain subregions of countries rather than entire countries. Sometimes surveys rely on nonprobability sampling techniques (42–44,46,47), e.g., convenience and quota sampling, when probability sampling techniques are not feasible. Paper or online self-reported contact diaries are largely used in social contact surveys. Compared to less common face-to-face interviews, respondent-filled contact diaries have a less demanding data collection procedure but may report a lower response rate (21,52). Zhang et al. (42) found significantly higher contacts documented by telephone interview than by self-reporting in Shanghai, China. In addition, contact diaries can be administered prospectively or retrospectively (**Table 2**). Prospective surveys have been shown to be less prone to recall bias compared to their retrospective counterpart (43), but it is often more challenging to find willing participants for prospective surveys. Other methods, e.g., proximity sensors and phone-based GPS trackers or Bluetooth scanners, have also been employed to measure mixing patterns between individuals (53–56) and are forming part of many countries' contact tracing efforts during the COVID-19 pandemic (57), though most have been implemented to protect users' privacy by storing data with the user rather than centrally. When we compared our synthetic matrices with empirical contact matrices from 11 studies using contact diaries, we found broad consistencies between findings from the two approaches. However, there were also differences which might reflect the heterogeneity in methods used to collect empirical data.

Another consideration affecting both synthetic and empirical matrices is that they change over time. Estimating synthetic matrices relies on the POLYMOD contact

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

survey administered more than a decade ago. Another larger contact survey, BBC Pandemic (53,58) conducted in the UK used mobile phone-based GPS tracking instead of diary-based surveys, reported a decrease in contacts among adolescents compared to POLYMOD, which may reflect substitution of face-to-face contacts with electronic communication in this age group. More differences are expected as countries implement physical distancing measures to mitigate the COVID-19 pandemic. The COVID-19 pandemic has brought about changes in human behaviour, in particular, how we come into contact with one another. The contact matrices serve as a baseline for comparison with "new normal" interventions implemented to control COVID-19.

Both synthetic and empirical matrices have complementary strengths and limitations. Empirical contact patterns are dependent on the study design and study population, and when the survey is administered. The synthetic contact matrices are constructed using proxies of contacts such as population and household age structures and country characteristics. However, the datasets used to develop these proxy measures (notably population age structure and DHS data) are generally much larger and more nationally representative than most empirical contact studies. To assess the robustness or consistency of the results under different mixing patterns, modellers should consider using multiple contact matrices constructed using different methods for sensitivity analyses.

Conclusion

In this study, we provide synthetic contact matrices for 177 geographical regions by updating our previous matrices with larger and more recent datasets on population age structure, household, school and workplace composition. The synthetic contact matrices reproduce the main features of the contact patterns in the out-of-sample empirically collected contact matrices.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

1. World Health Organization. Coronavirus Disease (COVID-19) Situation Reports [Internet]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
2. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Euro Surveill. 2020 Jan 1;25(4).
3. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020 Feb 15;395(10223):514–23.
4. WHO. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) [Internet]. 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf
5. World Health Organization. Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations [Internet]. Available from: https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations
6. Liu J, Liao X, Qian S, Yuan J, Wang F, Liu Y, et al. Community Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, Shenzhen, China, 2020. Emerg Infect Dis. 2020 Jun 1;26(6).
7. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med. 2020 Jan 29;
8. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis [Internet]. 2020 Mar [cited 2020 Mar 15];0(0). Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309920301444
9. Abbott S, Hellewell J, Thompson R, al. et. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020.
10. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet. 2020 Feb 29;395(10225):689–97.

11. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Heal [Internet]. 2020 Apr 1 [cited 2020 Jun 26];8(4):e488–96. Available from: http://www.thelancet.com/article/S2214109X20300747/fulltext
12. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Heal. 2020 Mar;0(0).
13. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ, Jombart T, et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. Lancet Public Heal [Internet]. 2020 Jun [cited 2020 Jun 26];0(0). Available from: http://www.thelancet.com/article/S246826672030133X/fulltext
14. Clark A, Jit M, Warren-gash C, Guthrie B, Wang HHX, Mercer SW, et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. Lancet Glob Heal [Internet]. 2020 [cited 2020 Jun 26];0(20):1–15. Available from: http://www.thelancet.com/article/S2214109X20302643/fulltext
15. Peak CM, Kahn R, Grad YH, Childs LM, Li R, Lipsitch M, et al. Individual quarantine versus active monitoring of contacts for the mitigation of COVID-19: a modelling study. Lancet Infect Dis. 2020;
16. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against COVID-19 and influenza in Hong Kong: an observational study. medRxiv. 2020 Mar 16;2020.03.12.20034660.
17. 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 [Internet]. 2006 Sep [cited 2015 May 28];164(10):936–44. Available from: http://aje.oxfordjournals.org/cgi/doi/10.1093/aje/kwj317
18. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DAT. Close encounters

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

of the infectious kind: Methods to measure social mixing behaviour [Internet]. Vol. 140, Epidemiology and Infection. 2012 [cited 2020 Apr 24]. p. 2117–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22687447

19. Wallinga J, Edmunds WJJ, Kretzschmar M. Perspective: human contact patterns and the spread of airborne infectious diseases. Trends Microbiol [Internet]. 1999 Sep [cited 2018 Jan 15];7(9):372–7. Available from: http://www.sciencedirect.com/science/article/pii/S0966842X99015462
20. 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. Riley S, editor. PLoS Med [Internet]. 2008 Mar 25 [cited 2020 Mar 3];5(3):e74. Available from: https://dx.plos.org/10.1371/journal.pmed.0050074
21. 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. Vol. 30, Epidemiology. Lippincott Williams and Wilkins; 2019. p. 723–36.
22. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol. 2017 Sep 1;13(9):e1005697.
23. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. medRxiv [Internet]. 2020 Mar 27 [cited 2020 Apr 9];2020.03.24.20043018. Available from: http://medrxiv.org/content/early/2020/03/27/2020.03.24.20043018.abstract
24. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis [Internet]. 2020 Jun 1 [cited 2020 Jun 26];20(6):669–77. Available from: http://www.thelancet.com/article/S1473309920302437/fulltext
25. Zandvoort K van, Jarvis CI, Pearson C, Davies NG, group CC-19 working, Russell TW, et al. Response strategies for COVID-19 epidemics in African settings: a mathematical modelling study. medRxiv. 2020 May 3;2020.04.27.20081711.
26. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020 Feb 15;395(10223):497–506.
27. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA [Internet]. 2020 May 26 [cited 2020 Jun 26];323(20):2052. Available from: https://jamanetwork.com/journals/jama/fullarticle/2765184

28. Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, et al. Social mixing patterns in rural and urban areas of southern China. Proc R Soc B Biol Sci [Internet]. 2014 Apr [cited 2015 Jun 9];281(1785):20140268. Available from: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2014.0268
29. Dodd PJ, Looker C, Plumb ID, Bond V, Schaap A, Shanaube K, et al. Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. Am J Epidemiol. 2016;183(2).
30. Demographic and Health Surveys. The DHS Program - Demographic and Health Survey (DHS) [Internet]. Available from: https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm
31. United Nations Department of Economic and Social Affairs Population Division. World Population Prospects [Internet]. 2019. Available from: https://population.un.org/wpp/
32. International Labour Organization. Labour force by sex and age [Internet]. Available from: https://www.ilo.org/shinyapps/bulkexplorer45/?lang=en&segment=indicator&id=EAP_2EAP_SEX_AGE_NB_A
33. UNESCO Institute for Statistics. UIS Statistics [Internet]. Available from: http://data.uis.unesco.org/
34. OECD. Teachers by age (indicator). 2020;
35. United Nations Department of Economic and Social Affairs Population Division. Urban and Rural Population by Age and Sex, 1980-2015 [Internet]. Available from: https://www.un.org/en/development/desa/population/publications/dataset/urban/urbanAndRuralPopulationByAgeAndSex.asp
36. United Nations Statistics Division. United Nations Statistics Division - Demographic and Social Statistics [Internet]. [cited 2020 Jul 21]. Available from: https://unstats.un.org/unsd/demographic/sconcerns/densurb/densurbmethods.htm#D

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

37. International Labour Organization. Labour force participation rate by sex, age and rural / urban areas [Internet]. Available from: https://www.ilo.org/shinyapps/bulkexplorer11/?lang=en&segment=indicator&id=EAP_2WAP_SEX_AGE_GEO_RT_A
38. OECD. Differences in rural and urban schools'; student-teacher ratio and class size, 2015. OECD Publishing; 2018.
39. Social contact data | Zenodo [Internet]. Available from: https://zenodo.org/communities/social_contact_data/search?page=1&size=20
40. le Polain de Waroux O, Cohuet S, Ndazima D, Kucharski AJ, Juan-Giner A, Flasche S, et al. Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: A survey in Southwest Uganda. BMC Infect Dis [Internet]. 2018 Apr 11;18(1):172. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-018-3073-1
41. Funk S. socialmixr: Social Mixing Matrices for Infectious Disease Modelling. The Comprehensive R Archive Network [Internet]. Available from: https://cran.r-project.org/web/packages/socialmixr/index.html
42. Zhang J, Klepac P, Read JM, Rosello A, Wang X, Lai S, et al. Patterns of human social contact and contact with animals in Shanghai, China. Sci Rep. 2019 Dec 1;9(1):1–11.
43. Leung K, Jit M, Lau EHY, Wu JT. Social contact patterns relevant to the spread of respiratory infectious diseases in Hong Kong. Sci Rep. 2017 Aug 11;7(1):1–12.
44. Béraud G, Kazmercziak S, Beutels P, Levy-Bruhl D, Lenne X, Mielcarek N, et al. The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases. Chuang J-H, editor. PLoS One [Internet]. 2015 Jul [cited 2015 Jul 27];10(7):e0133203. Available from: http://dx.plos.org/10.1371/journal.pone.0133203
45. Kiti MC, Kinyanjui TM, Koech DC, Munywoki PK, Medley GF, Nokes DJ. Quantifying Age-Related Rates of Social Contact Using Diaries in a Rural Coastal Population of Kenya. Borrmann S, editor. PLoS One [Internet]. 2014 Aug [cited 2015 Jul 16];9(8):e104786. Available from: http://dx.plos.org/10.1371/journal.pone.0104786
46. Grijalva CG, Goeyvaerts N, Verastegui H, Edwards KM, Gil AI, Lanata CF, et

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

al. A Household-Based Study of Contact Networks Relevant for the Spread of Infectious Diseases in the Highlands of Peru. Chowell G, editor. PLoS One [Internet]. 2015 Mar [cited 2015 Jul 27];10(3):e0118457. Available from: http://dx.plos.org/10.1371/journal.pone.0118457

47. Ajelli M, Litvinova M. Estimating contact patterns relevant to the spread of infectious diseases in Russia. J Theor Biol [Internet]. 2017 Apr [cited 2018 Jul 17];419:1–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022519317300504
48. Johnstone-Robertson SP, Mark D, Morrow C, Middelkoop K, Chiswell M, Aquino LDH, et al. Social Mixing Patterns Within a South African Township Community: Implications for Respiratory Disease Transmission and Control. Am J Epidemiol [Internet]. 2011 Dec [cited 2015 Jul 23];174(11):1246–55. Available from: http://aje.oxfordjournals.org/cgi/doi/10.1093/aje/kwr251
49. Horby P, Thai PQ, Hens N, Yen NTT, Mai LQ, Thoang DD, et al. Social Contact Patterns in Vietnam and Implications for the Control of Infectious Diseases. Munayco C, editor. PLoS One [Internet]. 2011 Feb [cited 2015 Mar 10];6(2):e16965. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016965
50. Melegaro A, Fava E Del, Poletti P, Merler S, Nyamukapa C, Williams J, et al. Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe. PLoS One [Internet]. 2017 Jan [cited 2018 Sep 10];12(1):e0170459. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0170459
51. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2018. Available from: https://www.r-project.org/
52. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Vol. 15, International Journal for Quality in Health Care. 2003. p. 261–6.
53. Klepac P, Kissler S, Gog J. Contagion! The BBC Four Pandemic – The model behind the documentary. Epidemics. 2018 Sep 1;24:49–59.
54. Smieszek T, Castell S, Barrat A, Cattuto C, White PJ, Krause G. Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes. BMC Infect Dis [Internet]. 2016 Jul [cited 2018 Sep 10];16(1):341. Available from:

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

https://doi.org/10.1186/s12879-016-1676-y

55. Ozella L, Gesualdo F, Tizzoni M, Rizzo C, Pandolfi E, Campagna I, et al. Close encounters between infants and household members measured through wearable proximity sensors. PLoS One [Internet]. 2018 Jun [cited 2018 Sep 10];13(6):e0198733. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198733
56. Cattuto C, den Broeck W, Barrat A, Colizza V, Pinton J-F, Vespignani A. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. Neylon C, editor. PLoS One [Internet]. 2010 Jul [cited 2015 Jun 3];5(7):e11596. Available from: http://dx.plos.org/10.1371/journal.pone.0011596
57. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science (80-). 2020 May 8;368(6491).
58. Klepac P, Kucharski AJ, Conlan AJ, Kissler S, Tang M, Fry H, et al. Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project. medRxiv [Internet]. 2020 Mar 5 [cited 2020 Mar 27];2020.02.16.20023754. Available from: https://www.medrxiv.org/content/10.1101/2020.02.16.20023754v2

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Contributors

KP: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing
KvZ: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review & editing
PK: Investigation, Methodology, Software, Supervision, Writing – review & editing
RME: Investigation, Methodology, Software, Writing – review & editing
NGD: Investigation, Methodology, Software, Writing – review & editing
ARC: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing
MJ: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing

The following authors were part of the Centre for Mathematical Modelling of Infectious Disease 2019-nCoV working group. Each contributed in processing, cleaning and interpretation of data, interpreted findings, contributed to the manuscript, and approved the work for publication: Christopher I Jarvis, Quentin J Leclerc, Jon C Emery, Gwenan M Knight, Amy Gimma, Simon R Procter, Kathleen O'Reilly, Sophie R Meakin, Charlie Diamond, Stefan Flasche, Billy J Quilty, Anna M Foss, Thibaut Jombart, Katherine E. Atkins, Georgia R Gore-Langton, Adam J Kucharski, James W Rudge, Matthew Quaife, Arminder K Deol, Carl A B Pearson, C Julian Villabona-Arenas, Graham Medley, Alicia Rosello, Hamish P Gibbs, Samuel Clifford, Rein M G J Houben, David Simons, James D Munday, Megan Auzenbergs, Rachel Lowe, Joel Hellewell, Sam Abbott, Damien C Tully, Stéphane Hué, W John Edmunds, Yang Liu, Fiona Yueqian Sun, Oliver Brady, Sebastian Funk, Nikos I Bosse, Akira Endo, Timothy W Russell, Emily S Nightingale.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Declaration of interests

We declare no competing interests.

Data availability

All data used in this study can be downloaded from the cited references. The codes used to generate these analyses and the updated synthetic matrices are available at https://github.com/kieshaprem/synthetic-contact-matrices.

Funding

The following funding sources are acknowledged as providing funding for the named authors. KP, PK and MJ were partly funded by the Bill & Melinda Gates Foundation (INV-003174). MJ and NGD were partly funded by the National Institute for Health Research (NIHR; NIHR200929) MJ was partly funded by NIHR using UK aid from the UK Government to support global health research (16/137/109). The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care. KvZ was partly funded by DfID/Wellcome Trust (Epidemic Preparedness Coronavirus research programme 221303/Z/20/Z) and DfID/Wellcome Trust/NIHR (Elrha R2HC/UK DFID/Wellcome Trust/NIHR). KP, PK and MJ were partly funded by the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688). RME was partly funded by HDR UK (MR/S003975/1) and UK MRC (MC_PC 19065). PK was partly funded by the Royal Society (RP\EA\180004). ARC was partly funded by the Singapore's National Medical Research Council (COVID19RF-004, NMRC/CG/C026/2017_NUHS).

The following funding sources are acknowledged as providing funding for the working group authors. Alan Turing Institute (AE). BBSRC LIDP (BB/M009513/1: DS). This research was partly funded by the Bill & Melinda Gates Foundation (INV-001754: MQ; INV-003174: YL; NTD Modelling Consortium OPP1184344: CABP, GM; OPP1180644: SRP; OPP1183986: ESN; OPP1191821: KO'R, MA). DFID/Wellcome Trust (Epidemic Preparedness Coronavirus research programme

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

221303/Z/20/Z: CABP). DTRA (HDTRA1-18-1-0051: JWR). ERC Starting Grant (#757688: CJVA, KEA; #757699: JCE, RMGJH; 757699: MQ). This project has received funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688: WJE, YL). This research was partly funded by the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1: AG, CIJ, TJ). Nakajima Foundation (AE). This research was partly funded by the National Institute for Health Research (NIHR) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care (16/137/109: CD, FYS, YL; 16/137/109 & 16/136/46: BJQ; Health Protection Research Unit for Modelling Methodology HPRU-2012-10096: TJ; PR-OD-1017-20002: AR). Royal Society (Dorothy Hodgkin Fellowship: RL). UK DHSC/UK Aid/NIHR (ITCRZ 03010: HPG). UK MRC (LID DTP MR/N013638/1: GRGL, QJL; MR/P014658/1: GMK). Authors of this research receive funding from UK Public Health Rapid Support Team funded by the United Kingdom Department of Health and Social Care (TJ). Wellcome Trust (206250/Z/17/Z: AJK, TWR; 206471/Z/17/Z: OJB; 208812/Z/17/Z: SC, SFlasche; 210758/Z/18/Z: JDM, JH, NIB, SA, SFunk, SRM). No funding (AKD, AMF, DCT, SH).

Acknowledgments

We thank Patrick Walker, Oliver Watson and Azra Ghani from Imperial College London, as well as other modellers from around the world, who provided feedback on the 2017 synthetic contact matrices.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Tables

Table 1. Summary of the changes between the 2017 and 2020 synthetic matrices

	2020 synthetic matrices	2017 synthetic matrices
Overall coverage	177 geographical locations covering 97.2% of the world's population, including rural and urban settings	152 geographical locations covering 95.9% of the world's population
Population age composition data	2020 UN Population Division population demographic data for 177 geographical regions	2015 UN Population Division population demographic data for 152 geographical regions
Urban and rural population age composition data	Stratified population demographic data by urban and rural settings for 177 geographical regions	Not considered
Country-specific household data	51 countries: 34 additional low- and lower-middle-income countries	17 countries
Country-specific urban and rural household data	Stratified household data by urban and rural settings (for 43 DHS countries)	Not considered
Country-specific labour force participation rate data by age	177 geographical locations and stratified by urban and rural settings	152 geographical locations

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Country-specific school data	School data were curated at various levels of education: pre-primary, primary, secondary and tertiary. The updated school data included enrolment rates, average starting ages of schooling, number of years of schooling, pupil-to-teacher ratio at each education level from UNESCO Institute for Statistics for 177 geographical regions.	School data were curated at various levels of education: pre-primary, primary, secondary and tertiary. The country-specific school data included enrolment rates, pupil-to-teacher ratio at each education level from UNESCO Institute for Statistics for 152 geographical regions.
Country-specific urban and rural school data	Rural and urban differences in schools' pupil-to-teacher ratio for 36 countries from OECD	Not considered
Comparison between the synthetic matrices and out-of-sample empirically-constructed contact matrices	Comparison of mean contacts by age, full contact matrices, the proportion of contacts at other locations, and relative reduction in cases and age-specific infection attack rate using an age-stratified compartmental model of COVID-19 Social contact data from contact surveys were extracted from the Zenodo social contact database and directly from the published studies for 10 geographical locations.	Comparison of mean contacts by age Contact matrices were assessed directly from the published studies in 5 geographical locations: Kenya (digitised image), Peru (digitised image), Russia, South Africa, and Vietnam.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 2. Description of empirical contact survey studies used to construct contact matrices.

Region, Country	Authors (year)	Study design	Study population	Sampling methods	Data collection method (mode)	Rural / Urban	Casual contacts (i.e., short-term)
Shanghai, China	Zhang et al. (2019) (42)	Prospective	General population (n = 965)	Convenience sample	Paper-diary (self-report and interview-based)	Urban	Participants were allowed to include group contacts
France	Béraud et al. (2015) (44)	Prospective	General population (n = 2033)	Quota sampling	Paper-diary (self-report)	Rural and urban	Participants were allowed to provide open-ended notes
Hong Kong SAR, China	Leung et al. (2017) (43)	Retrospective	General population (n = 1149)	Quota sampling	Paper-diary and online survey (self-	Urban	Participants were allowed to include group contacts

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

					report)		
Kilifi, Kenya	Kiti et al. (2014) (45)	Prospective	General population (n = 568)	Stratified random sample	Paper-diary (self-report)	Rural and semi-urban	Undocumented
Highlands San Marcos, Cajamarca-Peru	Grijalva et al. (2015) (46)	Retrospective	General population (n = 588)	Convenience sample	Paper-diary (interview-based)	Rural	Participants were allowed to report unregistered contacts
Tomsk, Russia	Ajelli et al. (2017) (47)	Prospective	General population (n = 559)	Random and convenience samples	Paper-diary (self-report)	Urban	Participants were allowed to report unregistered contacts.
South Africa	Johnstone et al. (2011) (48)	Prospective	General population (n = 571)	Random sample	Paper-diary (self-report)	Rural	Undocumented
Sheema North Sub-	le Polain de	Retrospective	General population	Random sample	Paper-diary	Rural	Participants were allowed

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

District, Uganda	Waroux et al. (2018) (40)		(n = 566)		(interview-based)		to report unregistered contacts
Red River Delta, North Vietnam	Horby et al. (2011) (49)	Retrospective	General population (n = 865)	Random sample	Paper-diary (interview-based)	Semi-rural	Participants were allowed to report unregistered contacts.
Zambia and South Africa	Dodd et al. (2016) (29)	Retrospective	Adults (>18 years) (n = 3582)	Random sample	Paper-diary (interview-based)	Rural and urban	Participants were asked by the interviewer
Manicaland, Zimbabwe	Melegaro et al. (2017) (50)	Prospective	General population (n = 2490)	Stratified random sample	Paper-diary (self-report)	Rural and peri-urban	Participants were asked by the interviewer

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure captions

Fig 1. Comparison of the estimated proportion of contacts at other locations for the empirical contact studies from six geographical regions and POLYMOD survey. The estimated age-specific proportion of all contacts at other locations—transport, leisure, other locations—matrices from contact surveys at the country or geographical region (in black) are compared against that observed in the POLYMOD countries (in grey).

Fig 2A. Comparison of the normalised empirical and synthetic age-specific contact matrices in five geographical regions. The empirical matrices collected from contact surveys, modelled synthetic contact matrices, and the scatter plots of the entries in the observed (x-axis) and modelled (y-axis) contact matrices are presented. The correlation between the empirical and synthetic matrices are shown. The matrices are normalised such that its dominant eigenvalue is 1. To match the population surveyed in the empirical studies, the contact matrices from rural settings of Kenya and Peru are presented; and the contact matrix from urban settings of China is presented. No data are available in the grey regions.

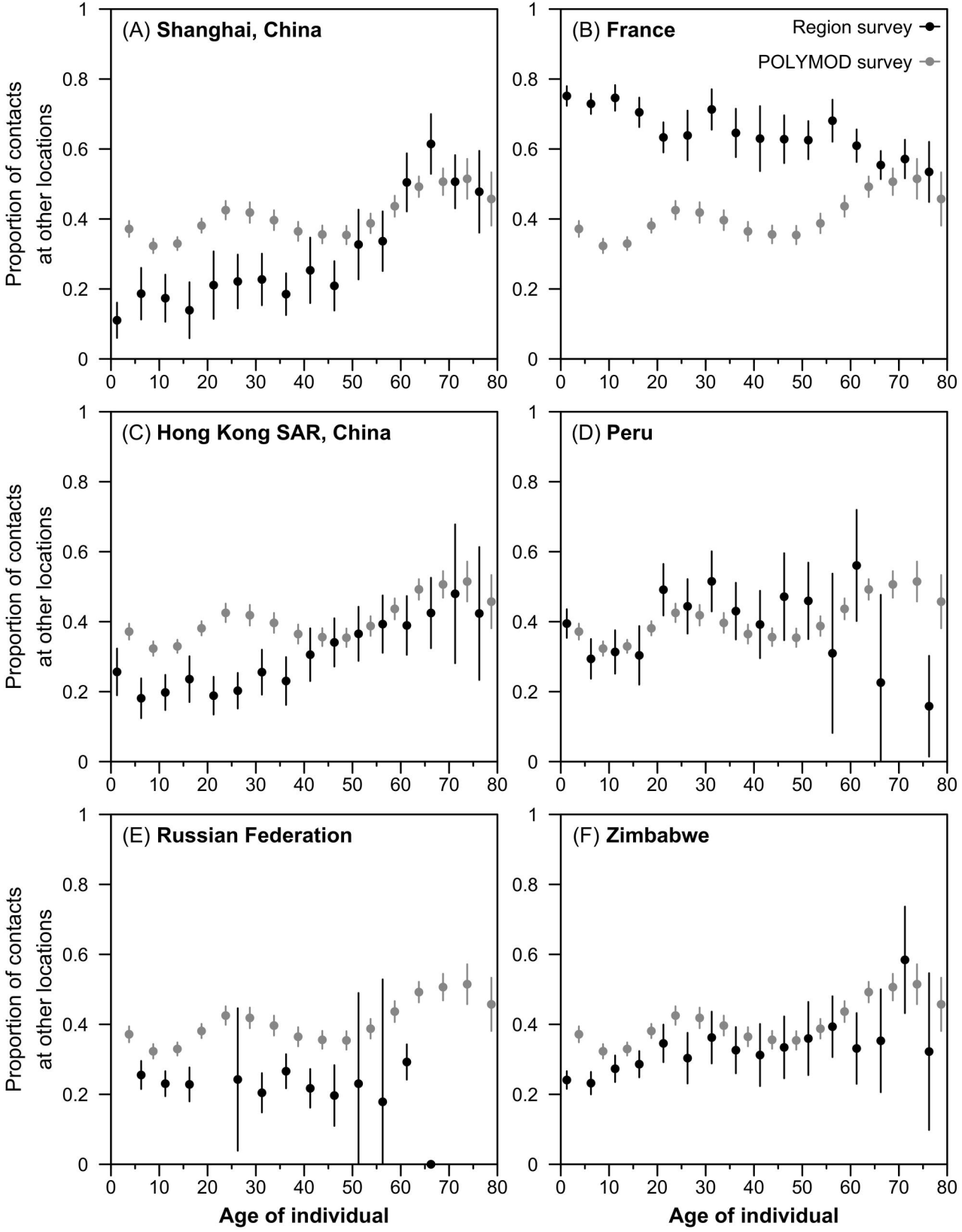
Fig 2B. Comparison of the normalised empirical and synthetic age-specific contact matrices in five geographical regions. The empirical matrices collected from contact surveys, modelled synthetic contact matrices, and the scatter plots of the entries in the observed (x-axis) and modelled (y-axis) contact matrices are presented. The correlation between the empirical and synthetic matrices are shown. The matrices are normalised such that its dominant eigenvalue is 1. To match the population surveyed in the empirical studies, the contact matrices from rural settings of South Africa, Uganda, Vietnam, and Zimbabwe are presented; and the contact matrices from urban settings of the Russian Federation are presented. No data are available in the grey regions.

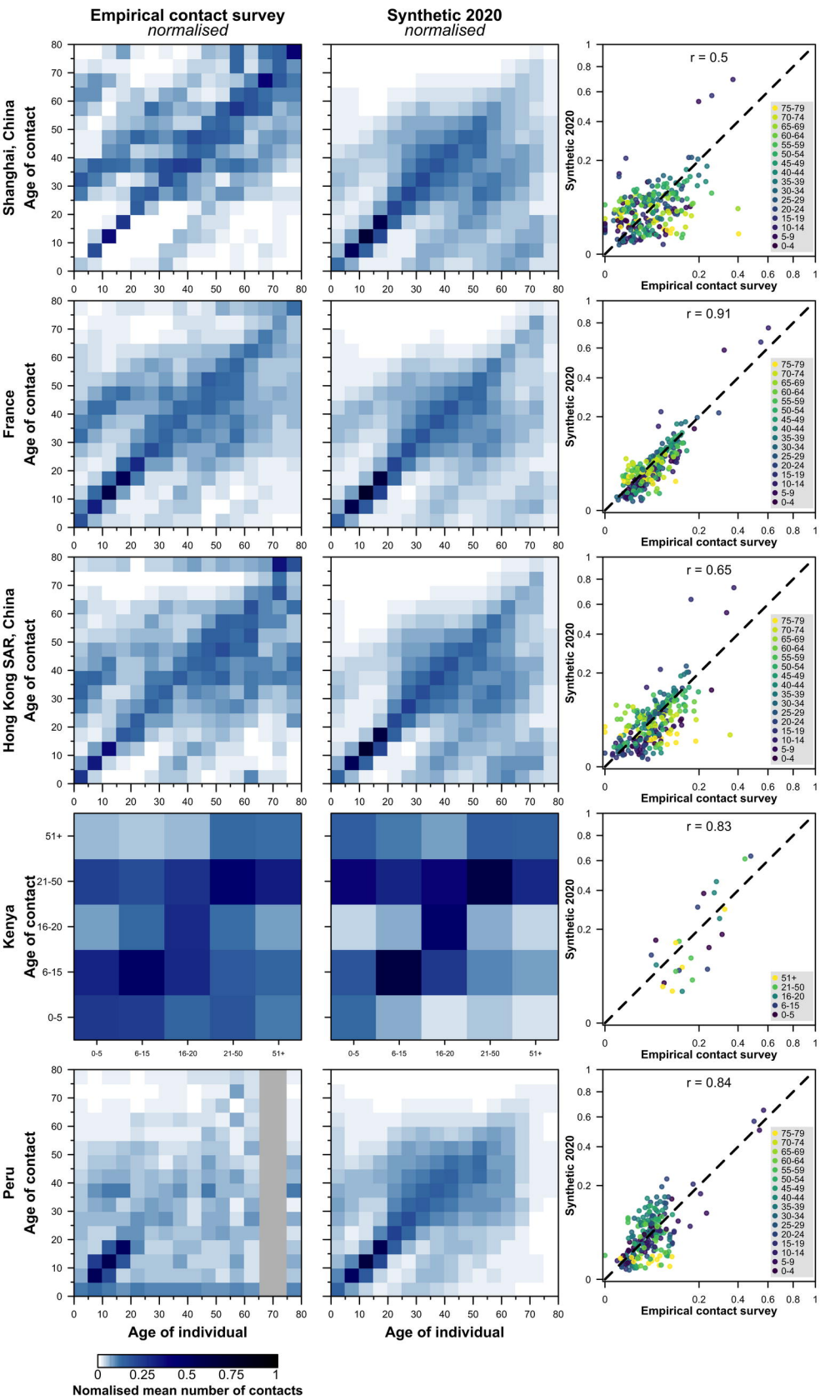
Fig 3. Mean number of contacts and basic reproduction number between rural and urban settings. Panels a and b present the scatter plots of the mean number of contacts in younger and older individuals, respectively, in rural (x-axis) and urban (y-

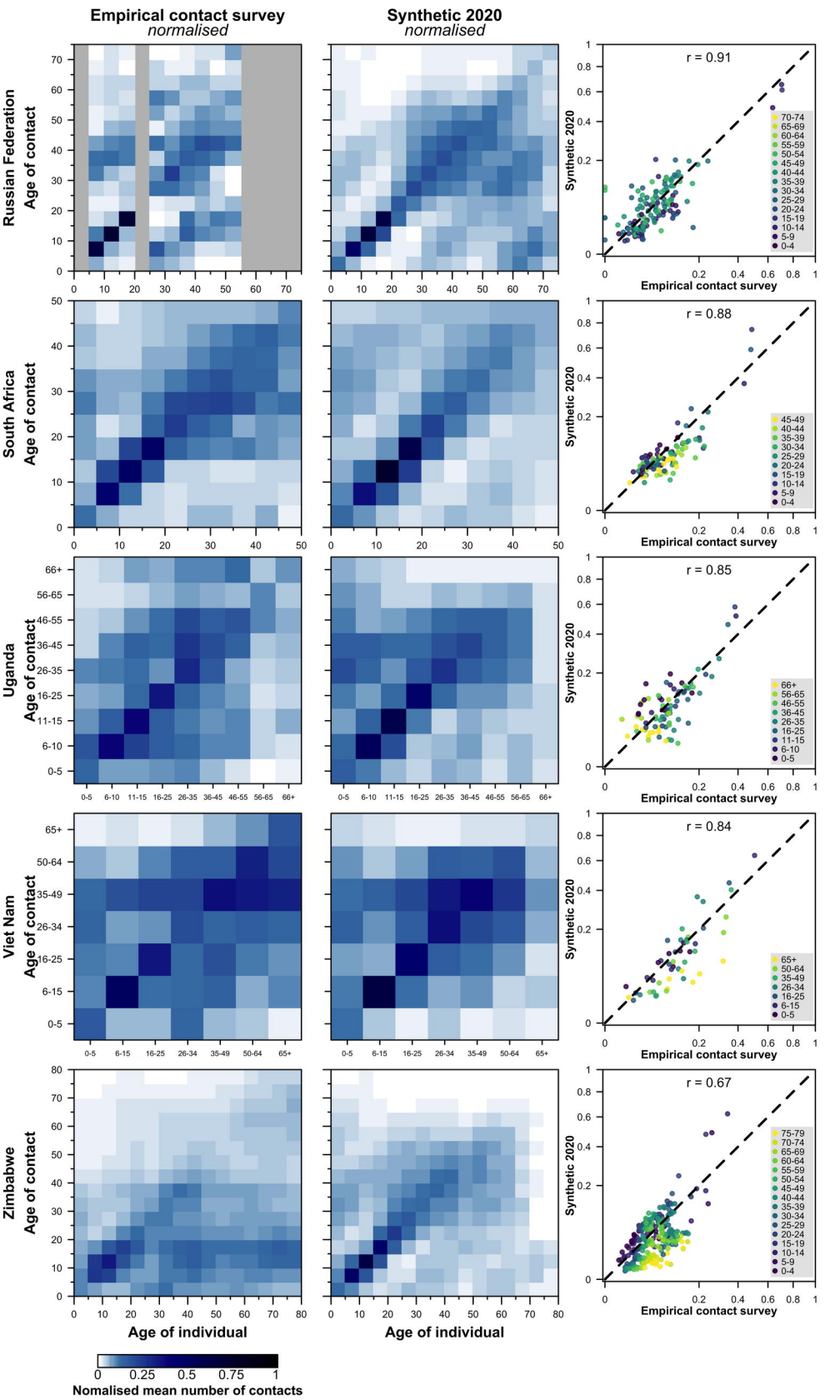
medRxiv preprint doi: https://doi.org/10.1101/2020.07.22.20159772; this version posted July 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

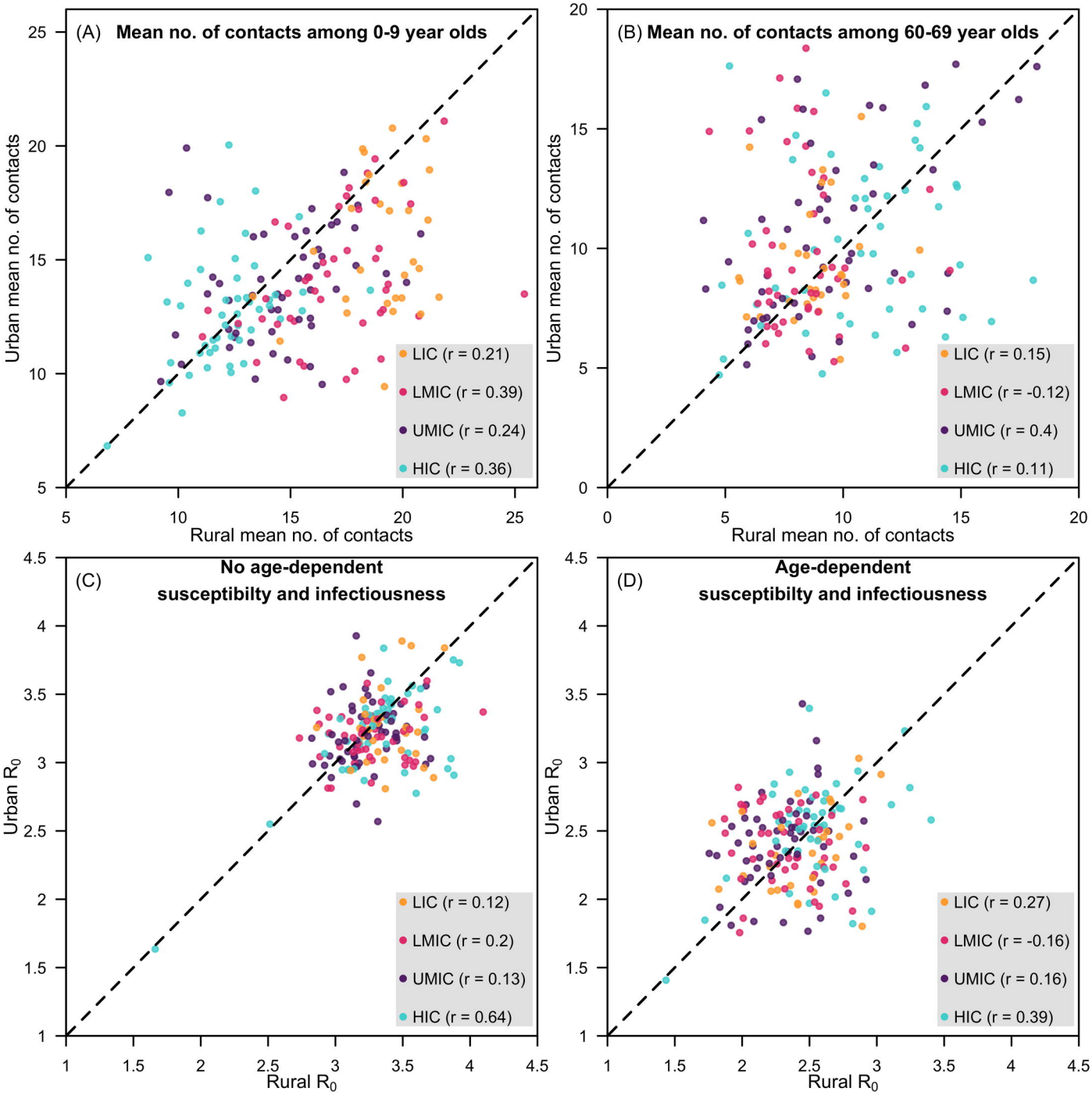
axis) settings of a country. Panels c and d present the scatter plots of the basic reproduction number in rural (x-axis) and urban (y-axis) settings of a country without and with age-dependent susceptibility and infectiousness. Geographical regions are grouped as low-income countries (LIC), lower-middle-income countries (LMIC), upper-middle-income countries (UMIC), and high-income countries (HIC), as designated by the World Bank in 2019. Within income group correlations of rural and urban values are presented in the accompanying parentheses.

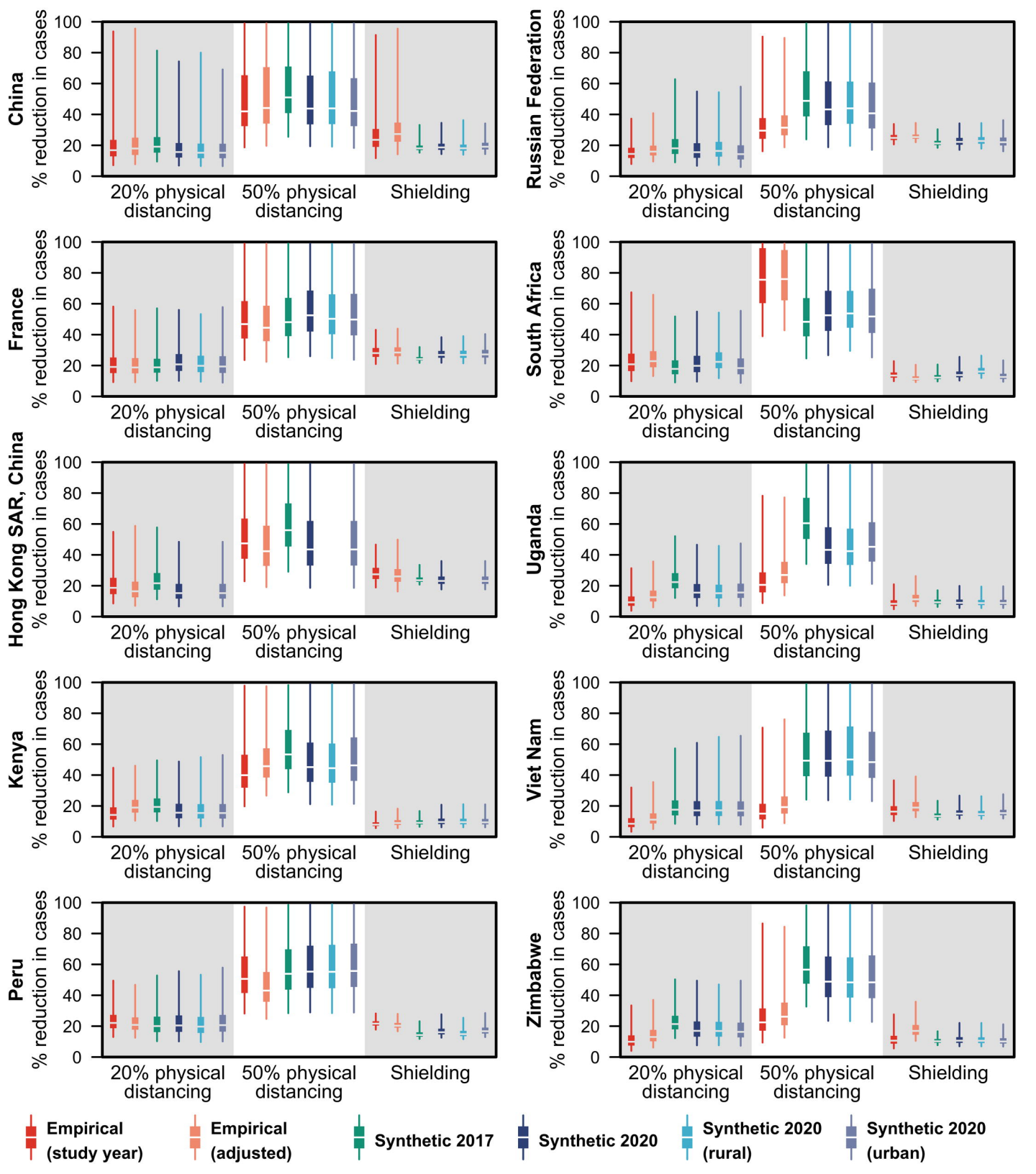
Fig 4. Relative reduction in cases due to interventions in models of COVID-19 epidemics under three intervention scenarios in ten geographical regions using the empirical and synthetic matrices. The percentage reduction in cases in each of the three intervention scenario—20% physical distancing, 50% physical distancing, and shielding—against the unmitigated epidemic under different contact matrices is shown in the boxplots with boxes bounded by the interquartile range (25th and 75th percentiles), median in white and, whiskers spanning the 2.5–97.5^{th} percentiles. Six contact matrices were considered in the COVID-19 modelling: the empirically-constructed contact matrices at the study-year and adjusted for the 2020 population, the 2017 synthetic matrices, and the updated synthetic matrices at the national, rural, or urban settings.


(A) Shanghai, China
(B) France
Region survey
POLYMOD survey
(C) Hong Kong SAR, China
(D) Peru
(E) Russian Federation
(F) Zimbabwe
Proportion of contacts at other locations
Age of individual
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50
60
70
80


Empirical contact survey
normalised
Synthetic 2020
normalised
Shanghai, China
France
Hong Kong SAR, China
Kenya
Peru
Age of contact
Age of individual
Synthetic 2020
Empirical contact survey
r = 0.5
r = 0.91
r = 0.65
r = 0.83
r = 0.84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
51+
21-50
16-20
6-15
0-5
Nomalised mean number of contacts


Empirical contact survey
normalised
Synthetic 2020
normalised
Russian Federation
South Africa
Uganda
Viet Nam
Zimbabwe
Age of contact
Age of individual
Synthetic 2020
Empirical contact survey
r = 0.91
r = 0.88
r = 0.85
r = 0.84
r = 0.67
Nomalised mean number of contacts


(A) Mean no. of contacts among 0-9 year olds
Urban mean no. of contacts
Rural mean no. of contacts
LIC (r = 0.21)
LMIC (r = 0.39)
UMIC (r = 0.24)
HIC (r = 0.36)
(B) Mean no. of contacts among 60-69 year olds
Urban mean no. of contacts
Rural mean no. of contacts
LIC (r = 0.15)
LMIC (r = -0.12)
UMIC (r = 0.4)
HIC (r = 0.11)
(C) No age-dependent susceptibilty and infectiousness
Urban R_0
Rural R_0
LIC (r = 0.12)
LMIC (r = 0.2)
UMIC (r = 0.13)
HIC (r = 0.64)
(D) Age-dependent susceptibilty and infectiousness
Urban R_0
Rural R_0
LIC (r = 0.27)
LMIC (r = -0.16)
UMIC (r = 0.16)
HIC (r = 0.39)


China
Russian Federation
France
South Africa
Hong Kong SAR, China
Uganda
Kenya
Viet Nam
Peru
Zimbabwe
% reduction in cases
0
20
40
60
80
100
20% physical distancing
50% physical distancing
Shielding
Empirical (study year)
Empirical (adjusted)
Synthetic 2017
Synthetic 2020
Synthetic 2020 (rural)
Synthetic 2020 (urban)