

The interplay between social inequalities and the spread of epidemic processes

PhD Thesis

by

Adriana Manna

Submitted to
Central European University
Department of Network and Data Science

*In partial fulfilment of the requirements for the degree of Doctor of Philosophy in
Network Science*

Supervisor: Prof. Márton Karsai

Vienna, Austria

2024

Ai miei nonni

Abstract

Mathematical models of infectious disease transmission are essential for predicting and analyzing epidemic outbreaks. During the COVID-19 pandemic, these models have played a critical role in identifying the mechanisms behind virus spread and evaluating control measures such as school closures, lockdowns, and mobility restrictions.

In recent years, traditional models have evolved to incorporate real-life heterogeneities, recognizing that the spread of diseases is influenced not only by pathogen-specific factors but also by social, demographic, and economic variables. Social scientists have shown that individuals' socio-demographic and economic characteristics not only shape the spread of diseases but also affect the severity of outcomes among those infected.

Despite extensive research on health inequalities, there has been limited focus on how these pre-existing inequalities influence the epidemic process itself. Current epidemiological models typically stratify populations by age but often overlook other critical heterogeneities within socio-economic strata. For example, contact patterns are usually simplified into age-stratified matrices (C_{ij}), which encode the average number of interactions between age groups. Factors like vaccination rates, susceptibility, and fatality rates are also primarily considered by age, neglecting other socio-economic dimensions that could alter the course of an epidemic.

This thesis aims to fill this gap by investigating the interplay between social inequalities and epidemic spread. The analysis focuses on two main areas. First, from a computational social science perspective, we identify key socio-economic characteristics—such as education and employment—that influence contact rates and vaccination uptake. Using data from the MASZK study in Hungary, we demonstrate that these socio-economic factors, like education and employment strongly shape individuals' contact patterns and their likelihood of getting vaccinated. Second, we propose an extended mathematical framework incorporating generalized contact matrices, which account for multiple socio-economic dimensions beyond age. Through simulations based on both synthetic and real-world data, we show that neglecting socio-economic status (SES) leads to underestimating the basic reproductive number (R_0) and misrepresenting epidemic dynamics. Our results reveal that SES-based segregation in social networks, along with varying activity levels and vaccination rates, can result in localized outbreaks and disproportionate disease burdens in lower-SES groups. Additionally, the effectiveness of non-pharmaceutical interventions (NPIs) is shown to vary significantly across different socio-economic strata.

This work contributes to a deeper understanding of how social inequalities shape epidemic outcomes, emphasizing the need for public health policies that address these disparities in both prevention and response.

Acknowledgements

Anyone who has known me in recent years has surely had the opportunity to hear me go on and on about how Vienna was “not my city”, “beautiful to visit but not to live in, “too grey,” “too cold,” even “too organized.” It may seem strange that my acknowledgements begin here. Thank you, Vienna, for being, despite it all, a place I can call “home”, for serving as the backdrop to these years of my PhD journey, and for gifting me those grey days that I could share with people who quickly became some of the most important in my life.

I am deeply grateful to my supervisor, Márton Karsai, for his unwavering guidance and support throughout this journey. Thank you for helping me find a clear path forward in every situation, even when I felt lost or overwhelmed. Your ability to see beyond and your scientific way of thinking have been crucial in putting together the pieces of this work. I feel fortunate to have had the opportunity to work with you and to have found in you not only a mentor but also a constant source of insight, encouragement, and, above all, friendship. A sincere thank you to my collaborators, especially Jùlia Koltai and Nicola Perra, for their invaluable contributions and for all that I learned through our work together. I am also grateful to all the professors of DNDS for their kindness and friendliness, which helped make the office atmosphere so welcoming, particularly to Elisa, Fede, and Petra.

I thank Alessia Melegaro, none of this would have been possible without you. Thank you for seeing potential in me, for supporting and guiding me through research and life’s interrogatives, and for being by my side from the very beginning as a professor, a mentor, and a friend; and Filippo Trentini, with whom I immediately found a genuine understanding, thank you for your sincere friendship and for always being available to help, and sharing your experience and knowledge.

I am grateful to Leo R., who one day led his PhD cohort on a bike trip to admire the beautiful flowers of Tulln, to Elsa, who generously swapped bikes with me and rode mine with a broken handlebar and a flat tyre for 30 kilometres on the way back, as Leo G. immediately recognized, it was the start of a great friendship. We may not have found flowers, but I can’t imagine a better beginning to these wonderful friendships. Thank you, Elsa, for your boundless generosity, for sharing countless evenings and days over these years, and for being my best friend from the very first of those days, no words can express how grateful I am to have shared these years with you. Thank you, Leo D.G., for your smiles, our talks, and for helping me see things through your eyes that I couldn’t see on my own. And thank you, Leo R., for sharing your enthusiasm and the brilliant idea to escape to Tenerife for ten days in the heart of winter.

I am grateful to every student who, for long or short, became part of the DNDS family. Each of you, in your own way, has left an indelible mark, filling these years with laughter and joy. Gabri, thank you for your warm welcome and for looking after all of us; Helciño, for your unique way of simply being you; Juli, for laughing so loud during

work hours; Timur, for making the pride parade unforgettable; Basti, for being my very German friend and housemate; Onkar, for always having all the answers; Clarina, for that amazing “*curri curri adri*” sign and sharing with me the passion for “art and craft”; Silviña, for filling our evenings with the sweets notes of *Skinny Love*; Gale, thank you for being my constant distraction at work and for teaching me it’s always the right time for a biliardino match; Thomas, for patiently supporting my countless requests for help and being my kind friend *Fromage*; Martí, for your warm hugs; Felipe, for cooking that unforgettable salmon; Lorenzo, for teaching Helcio the most important Italian words, for *sjoerd* to Sjoerd, Marti C, Luka, Sandeep, Sina, Jan, Bojan, Yijing, Vale, Greg, Iacopo, Filippo, Jun, Omar, Onur, Nelson, Rémi and Lisette. And thanks to all of you for joining the *salad team* from time to time!

I thank my family, *i miei nonni*, for their unconditional love; Mamma and Papà, for always being there for me; my sisters, Ludo and Emma, who have been extraordinary lifelong companions; my cousins, especially Irma, for being like an older sister, and Stella, for being the best *conqui* ever, and all my uncles and aunts. I am grateful to all my friends, especially Adri, Polline, Lele, Marghe, Paola, and the *Banditos*: Franci, Cla, Elly, Benny, Lalla, Bruni, Sabri and Tata, my “football team” and my dearest friends with whom I grew up and I hope I will grow old.

And thank you, Ludo; meeting your shy gaze was the best unexpected twist I could have imagined.

Thank you to all of you, without your help and affection, this achievement would not have been possible.

Declaration of Authorship

I, the undersigned, Adriana Manna, candidate for the PhD degree in Network Science declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, April 5, 2025



Signature

Copyright Notice

Copyright © Adriana Manna, 2024. *The interplay between social inequalities and the spread of an epidemic process* - This work is licensed under [Creative Commons Attribution-NonCommercial-ShareAlike \(CC BY-NC-SA\) 4.0](#) International license.



¹

¹Icon by [Font Awesome](#)

List of Publications

A Manna, J Koltai, M Karsai. (2024) Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics. Nat Commun 15, 4137.

A Manna, L Dall’Amico, M Tizzoni, M Karsai, N Perra. (2024) Generalized contact matrices allow integrating socioeconomic variables into epidemic models. Science Advances 10.41 : eadk4606.

A Manna, M Karsai, N Perra. (2024) Social inequalities in vaccine coverage and their effects on epidemic spreading. medRxiv, 2024.11.01.24316556

Contents

1	Background	1
1.1	Networks observation for epidemics	2
1.1.1	Homogeneous mixing	3
1.1.2	Social structure	3
1.1.3	Contact network models	4
1.1.4	Multi-scale models	5
1.1.5	Agent-based models	5
1.2	Mathematical epidemiology	5
1.2.1	Compartmental models in epidemiology	6
1.2.2	The basic reproductive number	9
1.2.3	Modeling control measures	10
1.3	Data on proxy interactions	12
1.3.1	Contact diaries	12
1.3.2	Other contact data	14
1.4	Social inequalities in epidemics	14
1.4.1	Contacts and adaption capacity	15
1.4.2	Pre-existing health inequalities	16
1.4.3	Vaccination and opinion	16
1.5	Research Goals	17
1.6	Structure of the Thesis	17
2	Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics	19
2.1	Hungary's Pandemic Experience	20
2.2	The main determinants of human contact patterns	21
2.2.1	Sociodemographic dimensions	22
2.2.2	Statistical analysis	22
2.3	Beyond age-stratification: <i>the extended SEIR model</i>	25
2.3.1	Decoupled age-contact matrices	25
2.3.2	The <i>extended</i> SEIR model	27
2.4	Vaccination uptake differences	29
2.4.1	Statistical analysis	29
2.4.2	Investigating difference in vaccination within the <i>extended</i> SEIR model	30
2.5	Stratified modelling of the Hungarian scenario	31
2.6	Discussion	32
3	Generalized contact matrices allow integrating socio-economic variables into epidemic models	35
3.1	Generalized contact matrices	36

3.2	Epidemic model	37
3.3	Spectral properties of generalized contact matrices and the basic reproductive number	37
3.4	Numerical simulations	39
3.4.1	Synthetic generalized contact matrices	39
3.4.2	Modeling the adoption of non-pharmaceutical interventions	43
3.4.3	Empirical generalized contact matrices	44
3.5	Discussion	46
4	Social inequalities in vaccine coverage and their effects on epidemic spreading	49
4.1	Epidemiological model with generalized contact matrices and vaccination	50
4.1.1	Modeling vaccination uptake among the second dimension	51
4.2	Inferring epidemic burden for subgroups	51
4.3	Scenario analysis	52
4.3.1	Scenario 1: concurrent vaccinations and epidemics	53
4.3.2	Scenario 2: vaccinations completed before epidemics	56
4.4	Hungarian contact data	58
4.5	Discussion	59
5	Conclusion	61
A	Appendix chapter 1	64
A.1	The next-generation matrix approach	64
A.1.1	SEIR with homogeneous mixing	64
A.1.2	SEIR with structured population	65
B	Appendix chapter 2	67
B.1	MASZK dataset	67
B.1.1	Socio-demographic dimensions	67
B.1.2	Sample sizes	68
B.1.3	Data pre-processing	75
B.2	Robustness checks on the statistical model	75
B.2.1	Model on contacts	75
B.2.2	Model on vaccination	83
B.3	Bootstrapping	83
B.4	Contacts	84
B.4.1	Average number of contacts in the community and in the work layer by sub-groups	84
B.4.2	Average number of community contacts by sub-groups and age groups	86
B.4.3	Contact Matrices	87
B.5	Vaccination uptake	92
B.6	<i>Extended SEIRD</i> with vaccination	92
B.7	Epidemic Simulations	93
B.7.1	Impact of different contact patterns	94
B.7.2	Impact of different vaccination uptake	100
B.8	Model Calibration	103
C	Appendix chapter 3	105

C.1	Synthetic generalized contact matrices	105
C.1.1	Random mixing regime	108
C.2	Epidemic models with generalized contact matrices	109
C.2.1	Numerical simulations	109
C.2.2	Derivation of R_0	110
C.3	Other contacts data	112
C.3.1	Age-contact matrices	112
C.4	Parameters and implementation	112
C.5	Impact of generalized contact matrices on epidemic modeling	114
C.6	Robustness analysis	115
C.6.1	Sensitivity analysis on the spectral radius	115
C.6.2	Testing the model with three dimensions	117
C.7	Empirical generalized contact data	117
D	Appendix chapter 4	120
D.1	Vaccination data	120
D.2	Computing attack rates	120
D.3	Synthetic population and mixing patterns	122
D.4	Robustness analysis for Figures 1 and 3	122
D.4.1	Vaccination stratified only by SES	122
D.4.2	Vaccination stratified by age and SES	128
D.5	Non pharmaceutical interventions	131
D.6	Hungarian contact data	135

List of Figures

1.1	Different scale structures used in epidemic modelling	3
1.2	Flow diagram of the SEIR model	6
1.3	Flow diagram of the SEIRD model with vaccination	10
2.1	COVID-19 trends in Hungary: cases and contacts	21
2.2	Maximum confidence level computation pipeline	23
2.3	Key contact determinants during COVID-19 in Hungary	24
2.4	Contact dynamics by population subgroups	25
2.5	Age-contact dynamics by population subgroups	27
2.6	Differences in attack rates in extended SEIR models	28
2.7	Key vaccination determinants during COVID-19 in Hungary	30
2.8	Disparities in COVID-19 vaccination and averted attack rates	30
2.9	Modelled epidemic dynamics in Hungary during the 4th COVID-19 epidemic wave	32
3.1	Generalized contact matrices	37
3.2	Impact of generalized contact matrices on epidemic spreading	40
3.3	Modeling the adoption of non-pharmaceutical interventions (NPIs)	43
3.4	Empirical generalized contact matrices-June 2022	45
4.1	Scenario 1: Epidemic outcomes.	54
4.2	Scenario 1: The impact of NPIs under different vaccination distributions.	55
4.3	Scenario 2: Epidemic outcomes.	56
4.4	Scenario 2: The impact of NPIs under different vaccination distributions.	57
4.5	Epidemic outcomes under different vaccination distributions and real contact data.	58
B.1	Maximum confidence level: Accounting for the high number of zeros in contact	76
B.2	Maximum confidence level: Accounting for the high number of zeros in community contact	77
B.3	Control measures for COVID-19 in Hungary	78
B.4	Maximum confidence level: Accounting for the effect of control measures on contacts	79
B.5	Maximum confidence level: Accounting for the effect of control measures on community contacts	80
B.6	Maximum confidence level: Accounting for the effect of other control variable in contacts	81
B.7	Maximum confidence level: Accounting for the effect of other control variable in community contacts	82
B.8	Maximum confidence level: Accounting for the effect of NPIs on vaccination	83

B.9	Contact dynamics by population subgroups: income level and settlement type	84
B.10	level	85
B.11	Relative contact dynamics by population subgroups: income level and settlement type	85
B.12	Contact dynamics by population subgroups and age	86
B.13	Hungarian age contact matrices	87
B.14	Hungarian decoupled age contact matrices by employment situation	88
B.15	Hungarian decoupled age contact matrices by education level	89
B.16	Hungarian decoupled age contact matrices by settlement type	90
B.17	Hungarian decoupled age contact matrices by income level	91
B.18	Vaccination uptake by population sub-group and age groups	92
B.19	Extended SEIRD vs conventional SEIRD	94
B.20	Attack rate by age and population subgroup, $R_0 = 2.5$	95
B.21	Extended SEIRD vs conventional SEIRD: Sensitivity analysis to different R_0 values	97
B.22	Attack rate by age and population subgroup, $R_0 = 1.5$	98
B.23	Attack rate by age and population subgroup, $R_0 = 3.5$	99
B.24	Difference in attack rate due to vaccination by age and population subgroup, $g_1 = 0.6$	100
B.25	Difference in attack rate due to vaccination by age and population subgroup, $g_1 = 0.2$	101
B.26	Difference in attack rate due to vaccination by age and population subgroup, $g_1 = 0.9$	102
B.27	Posterior distribution of the calibrated parameters	104
B.28	New deaths and infections during the 4th COVID-19 wave in Hungary	104
C.1	Population distribution by age and age-contact matrices for Hungary and Zimbabwe	112
C.2	Impact of generalized contact matrices on epidemic spreading	115
C.3	Sensitivity analysis on $\rho(G_{ab})$	116
C.4	Three dimensions generalized contact matrices	118
C.5	Empirical generalized contact matrices - April 2021	118
C.6	Empirical generalized contact matrices - November 2021	119
D.1	Hungary COVID-19 vaccination	120
D.2	Scenario 1. Epidemic outcomes and predictability	123
D.3	Scenario 2. Epidemic outcomes and predictability	123
D.4	Scenario 1. Epidemic outcomes and predictability	124
D.5	Scenario 2. Epidemic outcomes and predictability	125
D.6	Sensitivity analysis	127
D.7	Scenario 1. Epidemic outcomes and predictability	128
D.8	Scenario 2. Epidemic outcomes and predictability	129
D.9	Sensitivity analysis	130
D.10	Scenario 1: The impact of NPIs on SES groups under different vaccination distributions.	131
D.11	Scenario 1: The impact of NPIs under different vaccination distributions.	132
D.12	Scenario 2: The impact of NPIs on SES groups under different vaccination distributions.	133
D.13	Scenario 2: The impact of NPIs under different vaccination distributions.	134

D.14 Epidemic outcomes under different vaccination distributions and real contact data.	135
D.15 Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the entire population.	136
D.16 Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the vaccinated population.	137
D.17 Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the non-vaccinated population.	138

List of Tables

B.1	Number of survey participants in the different aggregated periods by age group and the education level	68
B.2	Number of survey participants in the different aggregated periods by age group and the education level	69
B.3	Number of survey participants in the different aggregated periods by age group and the income level.	70
B.4	Number of survey participants in the different aggregated periods by age group and the settlement.	71
B.5	Number of survey participants in the different aggregated periods by age group and gender.	72
B.6	Number of survey participants in the different aggregated periods by age group and chronic disease.	73
B.7	Number of survey participants in the different aggregated periods by age group and acute disease.	74
B.8	Number of survey participants in the different aggregated periods by age group and smoking behaviour.	74
B.9	Infection Fatality Rate by age group	104
C.1	List of parameters used to generate Figure 2 in the main text.	113
C.2	List of parameters used to generate Figure 3 in the main text, for (i) baseline and, (ii) NPI 2 scenarios.	114
D.1	List of parameters used to generate Figure 2 in the main text of Chapter 4.	122
D.2	List of parameters used to simulate the NPIs	131

Chapter 1

Background

Mathematical models of infectious disease transmission are crucial to provide predictions and scenario analysis of epidemics outbreaks. During various historical pandemics and the most recent COVID-19 one, epidemiological models have been demonstrated to be essential to identify the mechanisms underlying the diffusion of the virus and to evaluate effective control measures (lock-downs, restrictions to mobility or immunization strategies). In recent years, the traditional modelling approaches [1, 2] have developed toward a more accurate representation of the reality, progressively incorporating real-life heterogeneities which are driven not only by specific pathogen characteristics but also by different social, demographic, and economic factors of the single individuals [3, 4].

Simultaneously, social scientists have investigated historical and contemporary evidence of inequalities in pandemics, arguing that the sociodemographic and economic characteristics of individuals have not only shaped the spreading process of viruses in populations, but also the course of the disease among those infected [5]. There is a widespread agreement that pandemics disproportionately affect certain population groups rather than others [6–11].

Individual characteristics related to socioeconomic status (SES), such as wealth, race, ethnicity, occupation, and education, among others, have been shown to play a key role in the transmission of infectious diseases [12–15]. From the 1918 and 2009 influenza pandemics [6, 16] to the West African Ebola outbreak [17] and the COVID-19 pandemic [15, 18–28] individuals experiencing lower SES have been consistently associated with higher rates of infections and deaths.

Health-related inequalities in the burden of an epidemic partly arise from differences in the level of exposure to viruses and bacteria. These are associated with differences in social interactions, mobility patterns and work-related conditions, which are aggravated by disparities in the ability to be compliant with non-pharmaceutical interventions (NPIs), such as self-isolation, home-office and avoiding crowded places [21, 25, 26, 29–31]. At the same time, inequalities in the severity and fatality of a disease can account for the heterogeneity in pre-existing individual health conditions, protection attitudes, and access to medical care, which are themselves related to sociodemographic and economic factors [32, 33].

Although it is widely recognized that socioeconomic inequalities play a crucial role in the transmission dynamics of diseases, traditional mathematical approaches have often overlooked these factors. Indeed, the state-of-the-art framework of modelling infectious diseases incorporates stratification of the population according to age groups [34], while discarding other potential relevant heterogeneities between groups of individuals belonging to different socio-economic strata. They commonly ignore the mechanisms through

which these heterogeneities come into play, both directly and indirectly, in the different phases of an epidemic process.

In traditional epidemiological models, contact patterns are usually represented in the aggregate form of an *age contact matrix* (C_{ij}), which encodes information on the average number of contacts that individuals of different age groups have with each other [35–40]. Moreover, not only the description of contact patterns is limited to an age structure but also other epidemiological-relevant factors, such as vaccination uptake, infection fatality rates [41] or susceptibility [42] are usually described only by considering differences between different age groups.

Socioeconomic status is often considered only *a posteriori* in analyses of models’ output (e.g., number of deaths or cases) but rarely enters the core of their formulation as age. The roots of this shortcoming can be traced back to the lack of empirical surveys of social contact data including characteristics of the respondents and their peers to account for SES [43, 44]. The lack of data has also hampered the development and exploration of general modeling frameworks designed to include these features in their core [15].

While age is unarguably one of the most important determinants of these characters, the current literature falls short of understanding the role of other social, demographic, and economic factors in shaping human behaviour that are relevant to the epidemic’s spreading. In recent years, researchers have advocated including social aspects in infectious disease modelling, arguing that the epidemic modelling community lacks a deep understanding of the mechanisms through which the socio-economic divide translates into heterogeneities in the spread of infectious diseases [12–15].

With this in mind, this thesis aims to shed light on the interplay between social inequalities and the spread of epidemic diseases. The focus is on two main areas. First, from a computational social science perspective, we seek to identify the most important individual characteristics and corresponding subgroups within the population that significantly influence their epidemic-relevant behaviors. Second, from a mathematical epidemiology perspective, we aim to extend models to explicitly account for different social groups defined beyond age, and examine how the epidemiological outcomes of these proposed models differ from traditional models that overlook these dimensions. These questions will be explored through the lens of network science, focusing on the different structures of contact networks that may emerge in diverse socio-economic groups, with particular emphasis on analyzing the *ego networks* of individuals within these groups.

1.1 Networks observation for epidemics

Human contact patterns represent the routes through which many infectious diseases spread by shaping the underlying transmission chain within a susceptible population. The nature of the required contact for disease spread depends on the type of disease. For instance, influenza-like illnesses can be transmitted through simple face-to-face interactions, while sexually transmitted infections require intimate contact. Regardless of the type of contact, the most intuitive way to understand human-to-human interactions is to describe them as a *network*, where nodes represent individuals, and their interactions are represented by the edges connecting them. Consequently, all paths connecting pairs of nodes in the network represent the potential routes through which the disease can propagate within the population.

Modeling the dynamics of epidemics that propagate through contact networks requires encoding this information into a mathematical framework. The main challenge

lies in the heterogeneity of contact patterns among individuals. These patterns depend on various factors, such as the type of occupation (for example, office job or not), travel habits, personal attitudes, and more. Therefore, accurately tracking and modeling individual movements across different scales, from local interactions to global mobility patterns, is crucial for the accuracy and reliability of epidemic models and forecasts.

In this context, mathematical and statistical epidemiology has evolved from simple models to more structured approaches, where the heterogeneities and specific details of the population and system under study have become increasingly important. As the level of complexity increases, it is possible to define five broad classes of modelling approaches in computational epidemiology [45](see Fig. 1.1).

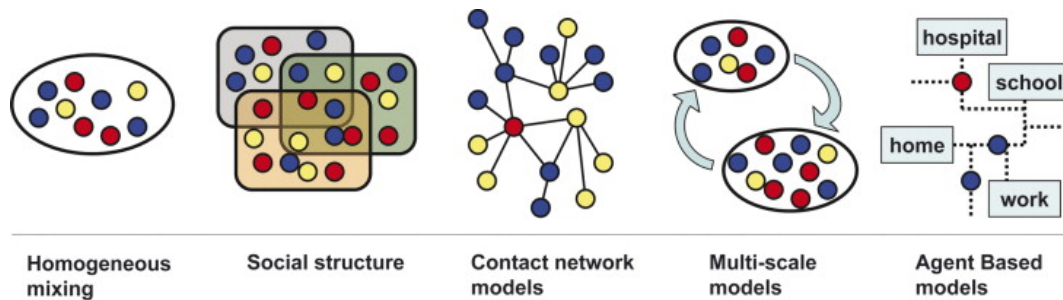


Figure 1.1: Different scale structures used in epidemic modelling Circles represent individuals and each colour corresponds to a specific stage of the disease. *Source:* Ref. [45]

1.1.1 Homogeneous mixing

The most straightforward assumption in epidemiological modeling is homogeneous mixing [1]. This framework is based on a simple hypothesis: each individual of a given population is equally likely to interact with any other individual at any time. This equal to assume no distinction between individuals and no structure of interactions. Although homogeneous mixing models offer simplicity and ease of computation, they often fail to capture the nuanced and heterogeneous nature of real-world social interactions, which can significantly impact disease spread and outcomes.

1.1.2 Social structure

A first step to overcome the homogeneous mixing assumption is to subdivide the population into discrete compartments based on demographic characteristics such as age, gender, etc. [34] This stratification allows for a more accurate representation of interaction structures between different subgroups of the population. From a mathematical point of view, this stratification can be encoded in *contact matrices*, which provide detailed information on the average number of contacts individuals from different population groups have with each other. In this framework, individuals within the same subgroup are assumed to be identical in terms of their contact patterns and any other epidemic-relevant variables, such as susceptibility to infection, mortality rate, or vaccine uptake. This assumption simplifies the modeling process while still capturing key heterogeneities in the population's mixing patterns.

Age contact matrices Standard approaches often recognize that contact patterns and disease transmission dynamics vary significantly with age, leading to the development of social structure models that account for the stratification of contacts across

different age brackets. To this end, the *age contact matrices* $\mathbf{C}$ are introduced. The element C_{ij} quantifies the average number of contacts that an individual in age-bracket i has with individuals in age-bracket j within a certain time window. The population hence, it is divided into age brackets so that $N = \sum_{i=1}^K N_i$. The variables N_i indicate the number of individual in age group i while K indicates the number of different age-groups. Given the definition of contact matrices, we can define $R_{ij} = C_{ij}N_i$ as the total number of contacts that individuals in i have with those in j . In matrix form we can write $\mathbf{R} = \mathbf{NC}$, where $\mathbf{N}$ is diagonal and its elements describe the population in each age group.

Reciprocity The nature of social interaction poses the crucial constraint of *reciprocity* to the calculation of the age-contact matrix, that is, the total number of contacts that individuals in age-group i with individuals in age-group j must be equal to the total number of contacts that individuals in age-group j form with individuals in age-group i , such that $R_{ji} = R_{ij}$. However, due to limitations of surveys particularly due to errors in the reporting process, age contact matrices will not fulfil reciprocity perfectly and symmetrization procedures are needed to impose such conditions [46–48]. In the context of traditional age contact matrices, these are typically straightforward. Usually, the age contact matrix of the total number of contacts is symmetrized as follows:

$$C_{ij}^{sym} = \frac{R_{ij} + R_{ji}}{2N_i} \quad (1.1)$$

Where C_{ij}^{sym} is the symmetrized age contact matrix, R_{ij} indicates the total number of contacts from individuals in age class i with individuals in age class j and N_i is the total number of individuals in age class i .

Context-stratification As the number and type of contacts considerably vary by location, the setting (that is, location l) where interactions take place (i.e., household, workplace, school, community) is another popular choice for stratification [38]. The entries C_{ij} are then expressed as

$$C_{ij} = \sum_l \omega_l C_{ij}^{(l)} \quad (1.2)$$

where ω_l are weights capturing possible heterogeneities in the relevance of contacts in each context in the transmission of the disease.

1.1.3 Contact network models

In many real-world scenarios, individuals exhibit significant heterogeneity in contact rates; therefore, assumptions of homogeneous mixing or simplified social structures may be inadequate. In this context, network theory [49] offers a comprehensive framework for detailing interactions among individuals.

From a mathematical point of view, a network can be represented by a graph

$$G = (V, E) \quad (1.3)$$

where $V = \{v_1 \dots v_n\}$ denotes the set of vertices and $E = \{e_1 \dots e_n\}$ denotes the set of edges between these vertices. Each edge represents a relationship or interaction between the nodes it connects. Edges can be bidirectional, forming *undirected networks*, or have

a specific direction, forming *directed networks*. Additionally, edges can be weighted to represent the strength or frequency of contacts. For example, w_{ij} could denote the weight of the edge between individuals i and j , representing the frequency of their interactions.

Building on this framework, network science has notably advanced epidemiology, providing deeper insights into disease dynamics in complex networks [50].

1.1.4 Multi-scale models

Multi-scale models incorporate different levels of population granularity [51–53]. These models can range from local interactions within small communities to global movements across countries. A common approach in multi-scale modeling frameworks is the use of *Metapopulation Models*. These models consider nodes as entities where multiple individuals reside and move along the edges connecting these nodes. At the macroscopic level, the system is composed of a heterogeneous network of sub-populations. At this level, the contagion process in one sub-population can spread to other sub-populations because of individuals moving across sub-populations. At the microscopic level, each sub-population represents a group of individuals, and the contagion process evolves based on how these individuals mix within their sub-populations. Typically, their mixing patterns are described as homogeneous or defined by specific sociodemographic dimensions [54]. In this way, this class of models allows to consider different levels of heterogeneity. They can incorporate not only geographical structures, such as the mobility flows of individuals between locations, but also varying social mixing patterns within each sub-population.

1.1.5 Agent-based models

In recent years, epidemic models have advanced into large-scale micro-simulations and data-driven approaches that offer highly detailed spatial resolutions. One such approach is represented by the *Agent-Based Models* (ABMs), which are spatially structured and consider the discrete nature and mobility of individuals, incorporating the stochasticity of their interactions [55–58]. These models are based on the construction of synthetic populations that characterize individual mobility patterns down to households, schools, and workplaces. The resulting ensemble of potential epidemic scenarios provides quantitative indicators such as median, mean, and reference ranges for epidemic observables, including new cases, seeding events, and infection arrival times. An example of this approach is the Global Epidemic and Mobility (GLEAM) model, which integrates census and mobility data into a fully stochastic metapopulation network model, allowing for detailed simulation of the global spread of influenza-like illnesses [59].

1.2 Mathematical epidemiology

The history of mathematical modeling in epidemiology has its roots in the 18th century, beginning with Daniel Bernoulli’s pioneering work on smallpox inoculation [60, 61]. Bernoulli aimed to demonstrate the benefits of preventive ”vaccination” through mathematical arguments, marking the inception of mathematical epidemiology. The field advanced significantly in the early 20th century when the increasingly popular dynamical systems approaches were applied to epidemiology. A pivotal moment occurred around 1920, during the Golden Age of Theoretical Biology, with Ronald Ross’s mathematical modeling of malaria and the introduction of the SIR model by Kermack and McKendrick

in 1927. This general compartmental mathematical model remains the foundation of contemporary epidemic modeling, influencing both historical and modern studies of disease dynamics [62].

1.2.1 Compartmental models in epidemiology

In the mathematical representation of the diffusion of infectious diseases, the dynamics within the host are typically reduced to changes between *the states of the diseases*, often referred to as *compartments*. Consequently, mathematical models for epidemics are called *compartmental models*. They aim to describe within a mathematical framework the dynamic process through which individuals flow from one compartment to another.

In the most simple case, the so-called *SI model*, individuals can only be in two states: (i) *susceptible* (S), which are healthy individuals who can contract the diseases if they contact an infected person and (ii) *infected* (I) individuals who have the disease. In addition to these, many other *compartments* can be added to the model, such as *exposed* (E), individuals who are infected but not yet infectious, or *recovered* (R), individuals who recovered from the disease and became immune. Depending on the type of compartments and the transition rules, compartmental models can effectively describe a wide range of infectious diseases [2, 62–64]. By incorporating various factors such as latency periods, varying transmission rates, and population heterogeneity, these models have been successfully applied to diseases like influenza, measles, and COVID-19.

The SEIR model

Throughout the thesis, we will primarily work with the SEIR model, which is introduced here and further elaborated on in the following sections.

In the SEIR model, individuals can ever only be in one of four states: *susceptible*, *exposed*, *infectious/infected*, and *recovered*. In this model, there are three possible transitions. The first transition, denoted $S \rightarrow E$, occurs when a susceptible individual interacts with an infectious one and becomes infected but not yet infectious, thus entering the exposed state. The second transition, denoted $E \rightarrow I$, happens when the exposed individual becomes infectious. Finally, the third transition, denoted $I \rightarrow R$, takes place when the infectious individual recovers from the disease and is assumed to have acquired permanent immunity. These transitions are represented in the flow diagram in Figure 1.2.

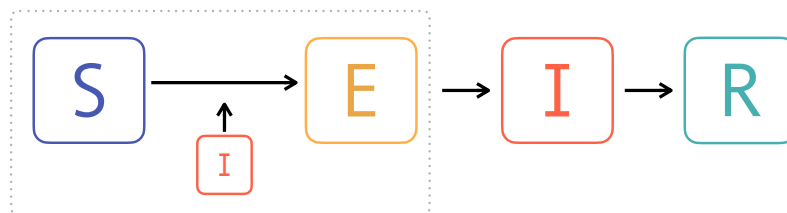


Figure 1.2: Flow diagram of the SEIR model The model allows the $S \rightarrow E$, $E \rightarrow I$ and $I \rightarrow R$ transitions, denoted by arrows from one compartment to another. Since the $S \rightarrow E$ transition occurs only if there is an interaction with an infectious individual (in compartment I), there is a call-out line to the infectious compartment.

The $E \rightarrow I$ and $I \rightarrow R$ transitions are the simplest to model. The probability that an individual will move from E to I and from I to R depends on how much time they have spent in the E and I classes, respectively. The transition from E to I occurs when an

exposed individual completes the incubation period and becomes infectious. The transition from I to R occurs when an infectious individual completes the infectious period and recovers. The distribution of incubation and infectious periods, which determine the probabilities of transition, can generally be estimated from clinical data. However, for simplicity in modeling, the probabilities of transition are often assumed to be constant. Thus, it is possible to define transition probabilities per unit of time. The rate at which individuals become infectious is denoted by ϵ , while the recovery probability is denoted by γ . Since we are dealing with probabilities per unit time, the average duration of the incubation and infectious periods are given by ϵ^{-1} and γ^{-1} , respectively.

The definition of the transition probability for the $S \rightarrow I$ transition is more complicated. The probability that a susceptible individual moves into the infectious compartment depends on two main factors: (i) the number of contacts with infectious people, and (ii) the probability that the disease is transmitted to the susceptible individual given contact with an infectious individual.

The number of contacts with infectious individuals, in turn, depends on the per-capita number of contacts per unit time with other individuals and the total number of infectious people present in the population. In the mathematical framework, this probability is represented by the *force of infection* λ , which expresses the probability per unit time at which susceptible individuals may contract the infection.

To express these relations explicitly, we define $S(t)$, $E(t)$, $I(t)$ and $R(t)$ as the numbers of susceptible, exposed, infectious, and recovered individuals at time t , respectively. The differential equations governing the epidemic's dynamics depend on how interactions among individuals are defined, as discussed in Section 1.1.

From a statistical perspective, the goal is often to group individuals who share similar characteristics thus exhibit statistically similar behaviour. For instance, in analyzing dynamical processes on complex networks, Degree-based Mean Field (DBMF) theory was one of the first theoretical approaches proposed [65, 66]. The DBMF approximation for dynamical processes on networks assumes that all nodes with the same degree k are statistically equivalent, meaning that their state dynamics can be averaged and described collectively. This simplification allows for the reduction of complex networks into a more tractable form, where the dynamics are governed by the degree distribution of the network rather than by individual node interactions.

Similarly, when social interactions are represented using age contact matrices, it is assumed that all individuals within the same age group are statistically equivalent. In particular, in an SEIR model where contacts are encoded as age contact matrices, each compartment of the epidemic model is repeated for each age group. Specifically, if the age contact matrix has K age groups, the model employs K sets of equations, each corresponding to a specific age group. In these cases, the differential equations governing the epidemic can be written as:

$$\begin{aligned}
 \partial_t S_i(t) &= -\lambda_i(t) S_i(t), \\
 \partial_t E_i(t) &= +\lambda_i(t) S_i(t) - \epsilon E_i(t), \\
 \partial_t I_i(t) &= +\epsilon E_i(t) - \gamma I_i(t), \\
 \partial_t R_i(t) &= +\gamma I_i(t)
 \end{aligned}
 \tag{1.4}$$

In this set of equations, the subscript i refers to the different age groups. The force of infection $\lambda_i(t)$ can be expressed as follows:

$$\lambda_i(t) = \phi \sum_{j=1}^K C_{ij} \frac{I_j(t)}{N_j} \quad (1.5)$$

where ϕ represents the transmissibility of the disease — the average number of transmissions per unit time — and C_{ij} represents the ij -th entry of the age-contact matrix. $I_j(t)$ represents the total number of infectious individuals at time t in age class j and $I_j(t)/N_j$ is the density of infected individuals in age class j at time t . Here, the temporal dependence is modelled by the variation in the number of infected at different time steps. In writing these equations, we also assume (i) a constant (closed) population size, N , such that $S + E + R + I = N$, and (ii) constant rates.

When heterogeneities among individuals are completely ignored, a more straightforward assumption is *homogeneous mixing*, where all individuals are considered as belonging to a single group. The epidemic dynamics can be described similarly to Equation (1.4). However, unlike the age-stratified model, which uses K sets of equations, this framework employs a single set of equations for the entire population. In this case, the force of infection can be expressed as:

$$\lambda = \frac{\phi I(t)}{N} \quad (1.6)$$

Here, $I(t)$ represents the total number of infectious individuals at time t and $I(t)/N$ is the density of infected individuals in the overall population at time t .

Modelling mortality

To model the number of individuals who die from the infection, a useful extension of the SEIR model is the addition of a death compartment. This extension modifies the standard SEIR equations by incorporating an additional transition from the infected I (infected) to D (deceased). This transition is governed by the infection fatality rate (IFR), which represents the proportion of infected individuals who ultimately die from the disease. In the age-stratified framework, starting from Equation (1.4), we need to adjust the equation describing the recovered compartment and introduce a new equation for the death compartment for each age group i . The Infection Fatality Rate (IFR) can be made dependent on the age group, thereby accounting for the impact of age-related factors on mortality probability. The modified equations are as follows:

$$\begin{aligned} \partial_t R_i(t) &= \gamma(1 - IFR_i)I_i(t) \\ \partial_t D_i(t) &= \gamma IFR_i I_i(t) \end{aligned} \quad (1.7)$$

Interestingly in this framework, the infection fatality rate splits the outflow from the infected compartment I_i into two pathways: one leading to recovery and the other to death. The logic behind these dynamics is that the total rate at which infected individuals leave the I_i compartment is still governed by the recovery rate γ . However, instead of all individuals moving to the recovered compartment R_i , a fraction determined by $(1 - IFR_i)$ recovers, while the remaining fraction, determined by IFR_i , dies.

1.2.2 The basic reproductive number

Arguably the most important quantity in disease modelling is the *basic reproduction number*, denoted as R_0 , which is defined as the average number of secondary cases arising from a typical primary case in an entirely susceptible population [67]. Intuitively, this quantity implies the strength of the epidemic and whether an outbreak will occur or not.

In epidemiology, models typically have two main equilibrium states: the disease-free equilibrium (DFE) and the endemic equilibrium. The DFE occurs when there are no infected individuals in the population, which in the SEIR model means $E^* = 0$ and $I^* = 0$. The endemic equilibrium, on the other hand, exists when infected individuals persist indefinitely, meaning $E^* > 0$ and $I^* > 0$.

The standard DFE for the SEIR model is represented by $(S^*, E^*, I^*, R^*) = (N, 0, 0, 0)$, indicating that the entire population is susceptible. This model does not have an endemic equilibrium because the number of infected individuals eventually returns to zero after the epidemic has run its course. An endemic equilibrium requires a continuous supply of susceptible individuals, often through births or waning immunity, for the infection to persist. Understanding the long-term behavior of these equilibria involves assessing their stability using *linear stability analysis*. This stability is closely linked to the basic reproductive number, R_0 .

We can use linear stability analysis, i.e. when the eigenvalues of the Jacobian matrix evaluated at the DFE have a negative real part, to show that the DFE is linearly stable if $R_0 < 1$, thus the epidemic will not be able to spread, and when $R_0 > 1$ it is unstable, meaning the epidemic will invade.

The next-generation matrix approach

As a rule, several traits of individuals are epidemiologically relevant for infectious agents, with age being the most commonly used. As seen before for the definition of the *age-contact matrices*, if age is considered as a discrete variable, individuals can be subdivided into a finite number of discrete categories based on their age groups. One can then define a matrix that relates the number of newly infected individuals in the various categories (age groups) in consecutive generations. This matrix is called the next-generation matrix (NGM) and it was introduced by Diekmann et al. [67], who proposed defining R_0 as the dominant eigenvalue of this matrix [68, 69]. This method analyzes the dynamics near the disease-free equilibrium (DFE). Below, we summarize the NGM approach for the SEIR model with an age-structured population. However, an extensive explanation of the mathematical details is presented in Appendix A.1, together with the mathematical details for the SEIR model with homogeneous mixing.

SEIR with social structure To find R_0 we focus on the compartments that describe any stages of the infection, that for an age-structured population are the exposed and the infectious compartments for each age group: E_i and I_i where $i \in [1, K]$. It is convenient to rewrite the differential equations for these compartments as $d_t \mathbf{x} = f(\mathbf{x}) - w(\mathbf{x})$, where $\mathbf{x} \in [E_i, I_i]$, $f(\mathbf{x})$ encodes all terms that lead to new infections and $w(\mathbf{x})$ all other transitions in and out of the compartments. The $\mathbf{F}$ and $\mathbf{W}$ matrices account for age-specific contacts and transitions. By linearizing the early epidemic dynamics around the DFE and using the Jacobian matrix, the system can be expressed as $d_t \mathbf{x} = (\mathbf{F} - \mathbf{W})\mathbf{x}$ where the matrix $\mathbf{F}$ is Jacobian applied to $f(\mathbf{x})$ and similarly the matrix $\mathbf{W}$ is the Jacobian applied to $w(\mathbf{x})$. The basic reproductive number R_0 is given by the spectral

radius of $\mathbf{FW}^{-1}$ which simplifies to $R_0 = \rho(\mathbf{FW}^{-1})$, where ρ denotes the spectral radius and we finally can write $R_0 = \frac{\phi}{\gamma} \rho(\tilde{\mathbf{C}})$ where $\tilde{C}_{ij} = \frac{C_{ij}}{N_j} N_i$.

1.2.3 Modeling control measures

One of the primary strengths of compartmental models is their versatility. Researchers can extend these models to incorporate additional factors and complexities, allowing them to describe a variety of different infection dynamics and control measures. For example one can use these models to account for non-pharmaceutical interventions and vaccination and thereby providing a comprehensive framework for understanding and predicting the spread of infectious diseases.

Vaccination Vaccination is a critical tool in controlling the spread of infectious diseases and has been of great interest both theoretically and empirically. While mathematical epidemiology has developed a comprehensive range of models to describe vaccination dynamics within compartmental frameworks [70, 71], this section will provide an introduction to the fundamental concepts underlying these models.

From a mathematical point of view, the most common approach is to add a new compartment, V , representing the vaccinated individuals. In this framework, individuals flow into the V compartment at a rate determined by the vaccination rate ν . These vaccinated individuals are typically considered to have become immune. However, this is a simple assumption. In many cases, the efficacy of vaccination against infection and death is not 100%.

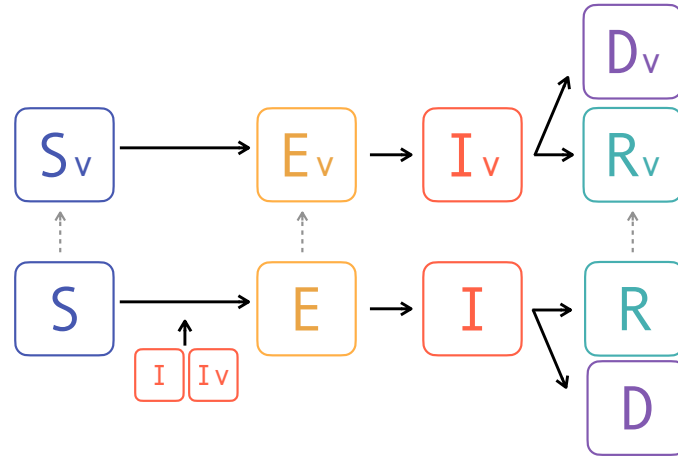


Figure 1.3: Flow diagram of the SEIRD model with vaccination The model depicts the transitions between compartments for both non-vaccinated and vaccinated individuals. For non-vaccinated individuals, the transitions are $S \rightarrow E$, $E \rightarrow I$, $I \rightarrow R$, and $I \rightarrow D$. For vaccinated individuals, the transitions are $S_v \rightarrow E_v$, $E_v \rightarrow I_v$, $I_v \rightarrow R_v$, and $I_v \rightarrow D_v$. These transitions are represented by solid black arrows. Transitions due to vaccination, denoted by dashed grey arrows, include $S \rightarrow S_v$, $E \rightarrow E_v$, and $R \rightarrow R_v$.

To account for this, compartmental models can be customized to split each compartment into *vaccinated* and *non-vaccinated* subgroups. For example, in the case of an SEIRD model in the homogeneous mixing, one would extend it as follows:

$$\begin{aligned}
\partial_t S(t) &= -\lambda_t S(t) - \nu S(t) \\
\partial_t S_v(t) &= -(1 - g_1)\lambda_t S_v(t) + \nu S(t) \\
\partial_t E(t) &= \lambda_t S(t) - \epsilon E(t) - \nu E(t) \\
\partial_t E_v(t) &= (1 - g_1)\lambda_t S_v(t) - \epsilon E_v(t) + \nu E(t), \\
\partial_t I(t) &= \epsilon E(t) - \gamma I(t) \\
\partial_t I_v(t) &= \epsilon E_v(t) - \gamma I_v(t) \\
\partial_t R(t) &= \gamma(1 - IFR)I(t) - \nu R(t) \\
\partial_t R_v(t) &= \gamma(1 - (1 - g_2)I_v(t) - \nu R(t) \\
\partial_t D(t) &= \gamma IFRI(t)_v \\
\partial_t D_v(t) &= \gamma(1 - g_2)IFRI(t)_v
\end{aligned} \tag{1.8}$$

In these equations, the subscript v denotes vaccinated individuals, ν is the vaccination rate, while g_1 and g_2 represent the efficacy against infection and the efficacy against death, respectively. In this model, individuals in all compartments can get vaccinated except the one infected or deaths. These transitions are represented in the flow diagram in Figure 1.3. This framework can also be extended to an age-structured model, where the vaccination rate ν can be specified for different age groups, thereby accounting for variations in vaccination coverage across age groups.

Non-pharmaceutical interventions Non-pharmaceutical interventions (NPIs) are measures that can help control the spread of infectious diseases. Social distancing, face masks, and increased hygiene are examples. Compartmental models can effectively incorporate NPIs to simulate their effects on disease dynamics. By adjusting parameters and modifying the flow of individuals among compartments, these models can reflect changes such as reduced contact rates, increased isolation of infected individuals, and decreased transmissibility due to measures like mask-wearing. Consequently, these models can project the potential impact of various NPIs on the course of an epidemic, providing valuable insights for public health decision-making.

This capability was crucial during the COVID-19 pandemic, where models were used to evaluate the effectiveness of various interventions in mitigating the spread of the virus [18, 72–75].

Non-pharmaceutical interventions aimed at reducing the number of contacts are particularly effective in controlling the spread of infectious diseases. These interventions can be incorporated into compartmental models by adjusting the contact rates between individuals. In structured population models stratified by age, the age contact matrix, defined as in equation (1.2), represents the weighted sum of contacts that individuals have in different locations such as work, school, home, and other places. To model NPIs that directly reduce the number of contacts individuals have, we can reduce or eliminate the corresponding elements of the contact matrix. Specifically, we can apply a reduction factor ρ to each specific setting and rewrite C_{ij} as follows:

$$C_{ij}^{\text{NPI}} = \sum_l \rho_l \omega_l C_{ij}^{(l)} \tag{1.9}$$

where ρ_l ranges from 0 to 1, with 1 representing no intervention and 0 representing

maximum stringency of the NPI. Depending on the NPI we want to model, we can tune these parameters accordingly. For example, to model the closure of offices and schools, we could drastically reduce ρ_{school} and ρ_{work} .

This approach can incorporate age heterogeneity considering, for instance, contact reductions as a function of age and location [76]. Thus, it allows us to model the impact of targeted NPIs more precisely. By incorporating these modified contact matrices into the model equations, we can simulate the effects of different NPIs on disease dynamics.

1.3 Data on proxy interactions

Mathematical models for epidemic spreading heavily rely on the specific mathematical encoding of the network of human-to-human interactions. However, empirical data describing proxy interactions between individuals are by nature difficult to gather. Various techniques have been developed to this aim. These techniques include surveys to collect contact diaries [35, 36, 77, 78], the development of synthetic population models [38, 39] and thanks to the increased availability and use of novel technologies, data collection methods based on various types of wearable sensors and contact tracing app [79–81].

1.3.1 Contact diaries

Contact diaries are one of the primary tools used to collect empirical data on social interactions that might be relevant to epidemic spreading. This technique requires participants to record their contacts, usually from the day before. In these surveys, a contact that is relevant for disease transmission is commonly defined as *an interaction where the respondent and a peer stayed within 2 meters for more than 15 minutes* [82]. Often participants are also asked to provide some information about the person they had contact with (*contactee*); the standard practice is to ask for age (or age group) and gender. This information on the age of the *contactee* is crucial for estimating the *age-contact matrices*. Additionally, participants are asked about the contact itself: whether it was physical or not, the frequency and duration of the contact, and the context in which it occurred—such as at school, in the workplace, within a household, or during another type of activity.

One of the main disadvantages of contact diary surveys is that self-reporting procedures entail biases that are difficult to estimate [83, 84], as participants might not recall all their contacts or might incorrectly estimate their duration [85, 86]. According to [87], factors such as ambiguity, emotions, and rapid responding, together with the retrospective collection of survey data, can induce an extreme responding bias in scale-rating diaries. Furthermore, time perception can be inaccurate in retrospective analysis, as people’s recollections generally decay rapidly with time [88]; in this respect, having a precise schedule of daily activities can help assign duration to meetings [89].

On the other hand, the main advantage of this type of data collection is that the survey can be designed to be statistically representative of the population being studied along a given set of socio-demographic dimensions, such as *age*, *gender*, and *settlement*. They also provide detailed information about the participants. When it comes to mathematical models, contact diaries can be encoded in the form of *contact matrices*, making computation quick.

As age has been widely recognized as the most important dimension in determining differences in contact patterns, significant efforts have been made over the last few years to design studies to estimate such *age contact matrices*.

The POLYMOD study One of the most important data collection has been carried out in the POLYMOD study [36]. This is a population-based prospective survey of mixing patterns carried out in 2008 in eight European countries: Belgium, Germany, Finland, Great Britain, Italy, Luxembourg, The Netherlands, and Poland. The authors collected information on 7.290 participants that recorded characteristics of 97.904 contacts with different individuals during one day, including age, sex, location, duration, frequency, and occurrence of physical contact. They found that mixing patterns and contact characteristics were remarkably similar across different European countries and that contact patterns were highly assortative with age. The age contact matrices estimated in this study have long been used in numerous research efforts to analyze the diffusion of infectious diseases.

The CoMix study As the COVID-19 pandemic emerged, the implementation of non-pharmaceutical interventions (NPIs) such as curfews and lockdowns radically altered the way individuals interacted. Many studies have since attempted to estimate age-contact matrices in the context of COVID-19 [77, 90]. A significant set of these studies is represented by the CoMix studies [91], conducted across 17 countries (United Kingdom, Austria, Denmark, Estonia, France, Poland, Portugal, Spain, Italy, Croatia, Finland, Greece, Hungary, Lithuania, Malta, Slovakia, Slovenia, and Switzerland) during the pandemic. The CoMix questionnaires were based on the *POLYMOD* contact survey.

The MASZK study The main data source of this work is represented by the *MASZK questionnaire* [92]. It collects data on social mixing patterns during the COVID-19 pandemic. It was carried out in Hungary from April 2020 to July 2022 on a monthly basis. The data was collected via cross-sectional anonymous representative phone surveys using the Computer Assisted Telephone Interviewing (CATI) methodology and involved a 1000 large nationally representative sample each month.

The primary goal of the data collection effort was to follow how people changed their social contact patterns during the different intervention periods of the pandemic and during a pre-pandemic reference period. Relevant to this study, the questionnaires recorded information about the approximate contact numbers were recorded between the respondents and their peers from different age groups of 0–4, 5–14, 15–29, 30–44, 45–59, 60–69, 70–79, and 80+. Contact number data about underage children were collected by asking legal guardians to estimate daily contact patterns.

Beyond information on contacts before and during the pandemic, the MASZK dataset provided us with an extensive set of information on *socio-demographic characteristics* (gender, education level etc.), *health condition* (chronic and acute illness etc.), *financial and working situation* (income, employment status, home office etc.), and *attitude towards COVID-19 related measures and recommendations* (attitude towards vaccination, mask-wearing etc.) of the participants. In Appendix B.1 we provide a detailed explanation of the main information used in this thesis.

In addition, in April 2020, a special phone call survey has been carried out during which participants were asked to provide a diary of contacts and to report for each of them the (i) age, (ii) gender, (iii) length of the contact, (iv) frequency and (v) type of relation. Afterwards, this special phone call survey has been extended in order to collect more information about the *alters*. Namely, respondents are asked to provide also (vi) the economic status, and (vii) the education level of the contacted person. This latter data collection has been carried out in April 2021, November 2021 and January 2022.

1.3.2 Other contact data

Wearable sensors Wearable sensors are another important tool for collecting data on social interactions [79, 80, 93, 94]. These devices can automatically record interactions, providing high-resolution data on the frequency, duration, and proximity of contacts. Typically, wearable sensors such as RFID tags or Bluetooth-enabled devices are worn by participants and continuously scan for nearby devices, logging each interaction when another sensor is detected within a certain range. This technique allows for the observation of real networks of *face to face* interaction, providing the opportunity to work with actual networks without the need for approximation or aggregation. Additionally, this method reduces recall bias and provides more accurate and detailed contact information compared to self-reported data. However, the main limitation of automated sensing platforms based on sensors is that they do not register contacts with individuals not participating in the data collection (i.e., not wearing any sensor), thus providing data only on the contacts among a closed population. This means that the data collected is not representative of the entire population of a specific location, city, or country.

Contact tracing apps These mobile applications use Bluetooth technology to detect and log proximity between devices, effectively tracing contacts when individuals are near each other for a significant duration. If someone tests positive for an infection, the app can quickly alert all individuals who were in close contact, enabling them to get tested or self-isolate. They played a crucial role during the COVID-19 pandemic by helping to monitor and control the spread of the virus [95–98]. This method offers several benefits over traditional contact tracing, as it provides real-time data and allows for faster notification. However, it also comes with challenges related to privacy concerns, data security, and adoption rates. Effective implementation depends on widespread use, as the utility of these apps diminishes significantly if only a small portion of the population uses them. Moreover, issues such as data privacy, misinformation, and technological accessibility need to be addressed to enhance trust and encourage broader adoption.

Synthetic population Synthetic populations are an innovative approach used to model social interactions. These models create virtual representations of individuals within a population, incorporating demographic and social characteristics to simulate realistic contact patterns. For example, Prem et al. utilized the POLYMOD age-contact matrices to project those of 144 other countries using a Bayesian hierarchical model before the COVID-19 pandemic [39] and during [99]. Similarly, Mistry et al. [38] proposed a data-driven approach to generate population-level contact matrices by leveraging detailed macro (census) and micro (survey) data on key socio-demographic features. Their work produced age-stratified contact matrices for 35 countries. One significant advantage of synthetic populations is the ability to make inferences for countries where those data are missing. However, they have two main disadvantages: they heavily rely on the accuracy of the input data, which may be limited, and the computational complexity can be significant, requiring substantial resources to generate and analyze the models.

1.4 Social inequalities in epidemics

The impact of social and economic inequalities on epidemic outbreaks has long been studied by social scientists, and there is widespread agreement in recognizing that pandemics disproportionately affect certain population groups over others [7–10]. Abundant

evidence has shown the impact of socio-economic, ethnic, and geographical inequalities on morbidity and mortality rates during pandemics. From the Influenza pandemics of 1918 and 2009 [6, 11, 16, 100] to the West African Ebola outbreak [17] and the Covid-19 pandemic [15, 18–28], individuals experiencing a lower socioeconomic status (SES) have been consistently associated with higher rates of infections and deaths. For example, Grantz et al.[16] investigated social inequalities during the 1918 influenza pandemic and found that factors such as literacy, home ownership, and unemployment were significantly associated with mortality rates. Similarly, Zhao et al.[11] examined the 2009 influenza pandemic (A(H1N1)pdm09) and identified a higher mortality rate among non-white populations and individuals living in areas with the highest levels of deprivation.

The current Covid-19 pandemic further prove that epidemics are not socially-neutral [9, 101]. Bambra et al.[33] argue that, for the most disadvantaged communities, Covid-19 is experienced as a *syndemic*¹, that is, a co-occurring, synergistic pandemic that interacts with and exacerbates their existing noncommunicable diseases (NCDs) and social conditions. International literature highlights disparities in COVID-19 prevalence and mortality rates across high-income and low-income countries, affluent and less affluent neighborhoods [25], socio-economic groups [103], and urban and rural areas [104].

These disparities arise through various mechanisms, which can be categorized into three main topics: (i) contacts and awareness, (ii) pre-existing health inequalities, and (iii) vaccination and opinion.

1.4.1 Contacts and adaption capacity

Many studies have demonstrated that contact attitudes which determines the exposure to the virus depend on several factors related to socio-economic status. Here, many different factors can come into play such as: (i) crowded living conditions and multi-generational households, which are especially prevalent in socioeconomically marginalized communities [105, 106]; (ii) work-related exposure for those occupations that do not permit working from home and entail physical proximity to other people or direct contact with the public [107]; (iii) use of public transportation (iv) lack of appropriate personal protective equipment; and (v) the possibility of respecting the lockdown restrictions. To this concern, [21] found that the impact of NPIs on mobility patterns significantly correlates with the Human Development Index of comunas in Santiago de Chile, showing that more developed areas of the city become better isolated during the pandemic. However, if what is stated above leads to assuming higher contact rates of people with low socioeconomic status, it is important to take into account that places of higher average income are associated with a higher customary density of personal interactions, including links to external sources of infection due to more intense travelling (both for work and leisure). In the same vein, poverty is often associated with greater social isolation compared to richer people who may well be better connected in both work and leisure activities. Additionally, studies on social mobility indicate that individuals with similar socioeconomic status (SES) tend to form homophilous networks, interacting predominantly with others from the same SES group [108–110]. These assortative interaction patterns restrict social mixing between groups of different SES levels, potentially leading to the segregation of epidemic processes. This limited inter-group contact can result in biased epidemic outcomes, as the disease may spread disproportionately within

¹The concept of *syndemic* has been introduced in 2009 by Merrill Singer[102] to describe the relation between HIV/AIDS, substance use and violence in the USA during the 1990s: "A *syndemic* is a set of closely intertwined and mutual enhancing health problems that significantly affect the overall health status of a population within the context of a perpetuating configuration of noxious social conditions".

certain SES clusters, creating localized outbreaks across the network, and affecting the overall transmission dynamics.

1.4.2 Pre-existing health inequalities

Health inequities refer to disparities in health status or to imbalances in the allocation of health resources across various population groups. These inequities arise from the diverse social conditions in which individuals are born, grow, live, work, and age [111].

There is ample evidence that the risk of a severe disease and death is generally higher among individuals of poor general health and nutritional status, and among those with underlying chronic conditions such as cardiovascular diseases, lung diseases, diabetes and cancer.

In turn, the prevalence of these conditions is inversely associated with socioeconomic status [32], the lower an individual's socio-economic position, the higher their risk of poor health.

Additionally, the negative impact of unemployment on health is well known, and includes poor mental health, increased alcohol and substance use and family violence [112].

1.4.3 Vaccination and opinion

Vaccination is one of the most effective public health interventions for controlling infectious diseases and protecting both individuals and communities [113]. Beyond their efficacy at various endpoints (e.g., preventing infection or death), vaccine access and uptake are key determinants of vaccination's impact on disease burden. Although vaccines are designed to be accessible to all, limited stockpiles often necessitate prioritization strategies [114, 115], commonly focusing on essential workers, the elderly, children, or individuals with underlying health conditions.

While socioeconomic status should ideally not influence access, significant inequities in vaccine distribution persist within and across countries, as seen during the COVID-19 pandemic [28, 116–120]. Studies indicate that individuals from lower socioeconomic backgrounds are often less likely to receive COVID-19 vaccinations [121–123]. For example, during Florida's initial vaccine rollout, lower-income and minority communities encountered notable barriers to vaccine access [124]. Geographic and ethnic disparities are evident in Hungary, where socioeconomic deprivation correlates with lower vaccination coverage [123, 125, 126]. In the Greater Manchester area in the UK, COVID-19 vaccination rates were significantly lower among Black and Asian ethnic groups compared to other groups [127].

The association between socioeconomic status and vaccination uptake also varies across diseases [128], though lower SES consistently correlates with reduced vaccination rates. These disparities result from a combination of structural and behavioral factors. Structural barriers include limited access to healthcare facilities, financial constraints, and lower health literacy, all of which restrict access to accurate information about vaccination benefits and availability [129]. Logistical challenges—such as inflexible work schedules and lack of transportation—further impede vaccination among disadvantaged populations [119].

Behavioral factors, such as mistrust in healthcare systems and governmental institutions and susceptibility to misinformation, disproportionately affect lower SES groups, often leading to higher levels of vaccine hesitancy [130, 131]. Social dynamics and exposure to misinformation play an important role in shaping vaccine uptake [132, 133]. For in-

stance, studies in France have highlighted elevated COVID-19 vaccine hesitancy among lower SES groups, often driven by concerns about vaccine safety and distrust in government health initiatives [130, 131].

These trends are observed consistently across age groups, underscoring the crucial influence of socioeconomic factors on both vaccine hesitancy and access [134, 135]. Together, these structural and behavioural factors contribute to persistent inequities in vaccination uptake across socioeconomic strata.

1.5 Research Goals

Although a large body of the literature has developed in order to investigate the determinants of health inequalities and the impact that epidemics have in determining them, little effort has been made to understand the complex mechanisms that the same existing inequalities in health and other socioeconomic aspects can have on the spread of the epidemic itself. The current state-of-the-art framework for modeling infectious diseases typically incorporates population stratification primarily by age, neglecting other potentially significant heterogeneities between socio-economic groups. This oversight ignores the mechanisms by which these inequalities affect the epidemic process, both directly and indirectly, across its various phases.

Therefore, this thesis investigates the interplay between demographic health conditions and social inequalities and an epidemic process. To achieve this objective, we address the following research questions: (i) Which demographic and economic variables most significantly shape epidemic-relevant behaviors, particularly social interaction and vaccination uptake, and are thus crucial for explaining infection patterns? (ii) How can these variables be integrated into epidemiological models? (iii) What are the main differences between traditional age-structured compartmental models and multi-structured compartmental models that explicitly account for additional socio-economic dimensions?

By answering these questions, this work represents a step forward in bridging the gap between epidemiological models grounded in physics and mathematics and the extensive body of social science research that has long studied health and socio-economic inequalities within and between specific settings and countries.

1.6 Structure of the Thesis

The present thesis is structured as follows:

Chapter 1: Background This chapter presents the most relevant findings from existing literature in both epidemiology and social science. It concludes with an outline of the general research goals.

Chapter 2: Importance of Social Inequalities This chapter investigates which socio-demographic and economic characteristics influence contact patterns and vaccination uptake. It takes a step beyond age-structured models by introducing a simple extension of traditional age-structured compartmental models that explicitly accounts for socio-economic differences in epidemiologically relevant behaviors.

Chapter 3: Generalized Age Contact Matrices This chapter delves deeper into mathematical modeling, introducing a new framework that explicitly includes socio-

economic groups. It presents generalized contact matrices, an extension of the age contact matrix, which stratifies contacts across multiple dimensions.

Chapter 4: Social inequalities in vaccine coverage and their effects on epidemic spreading This chapter focuses on incorporating vaccination strategies within the framework of generalized age contact matrices.

Chapter 5: Conclusion This chapter summarizes the findings and their implications for public health strategies. It discusses how this thesis advances the integration of epidemiological modeling with social science research on health and socio-economic inequalities.

Chapter 2

Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics

Individuals' socio-demographic and economic characteristics are among the most significant factors shaping the dynamics of epidemic spreading processes. These factors not only influence the epidemic outcome in the host population but also largely determine the course and severity of the disease for those infected. There is widespread agreement that pandemics disproportionately affect certain population groups, leading to health-related inequalities (See Section 1.4).

Although socio-economic inequalities are recognized as crucial to understanding disease transmission dynamics, traditional mathematical models often neglect these factors. Current models typically stratify populations by age (see Section 1.1.2), ignoring other significant heterogeneities among socio-economic groups. They commonly ignore the mechanisms by which these socio-economic differences influence the various phases of an epidemic, both directly and indirectly. While age is unarguably one of the most important determinants of these characters, the current literature falls short of understanding the role of other social, demographic, and economic factors in shaping human behaviour that are relevant to the epidemic's spreading.

With this in mind, in this chapter we aim to shed light on these mechanisms to address the following interrogatives: which are the most important individual characters, and corresponding subgroups of the population, that differentiate the most their epidemic-relevant behaviours; and, how do these differences translate into epidemiological outcomes? Particularly, we focus on examining two important aspects of human behaviour relevant to epidemics: contact patterns and vaccination uptake.

We use the *MASZK study* (See Section 1.3.1), by aggregating the 26 data collection in six periods covering the four epidemic waves (Ws) and two interim periods (IPs), as demonstrated in Figure 2.1.

Through the analysis of contact patterns and vaccination uptake, our aim is to show existing significant differences among subgroups of different socio-demographic characters, when accounting for the effect of age. We demonstrate that dimensions such as employment situation and education level play a crucial role in determining contact numbers and vaccination uptake during a pandemic. Additionally, by proposing a new data-driven mathematical framework, which explicitly considers further social dimensions, other than age, we analyse the impact of such differences in terms of epidemic outcomes. Finally, by focusing on the Hungarian COVID-19 pandemic scenario, we re-

veal the unequal impact of the pandemic in terms of individuals belonging to different socio-economic statuses, where we differentiate individuals by their employment situation and income level. Note that although all the models have been completed on each pandemic period, for the demonstration of our findings we exclusively show results about the 4th wave in the main text. We chose this period to demonstrate our results because NPIs were not significantly changing, while the vaccination rate had saturated already during the 4th wave, this way neither of these effects could bias the observed behaviours. We report our findings concerning other periods in Appendix B.

We find that generally higher socioeconomic groups of the population have a higher number of contacts and a higher vaccination uptake with respect to disadvantaged groups. Subsequently, we propose a data-driven epidemiological model that incorporates these behavioral differences. Finally, we apply our model to analyze the fourth wave of COVID-19 in Hungary, providing valuable information on real-world scenarios. By bridging the gap between individual characteristics and epidemic spread, our research contributes to a more comprehensive understanding of disease dynamics and informs effective public health strategies.

This chapter is based on the paper [122].

2.1 Hungary's Pandemic Experience

Between March 2020 and July 2022, Hungary faced four significant waves of the COVID-19 pandemic, each bringing unique challenges and requiring distinct public health responses. Following the declaration of a state of emergency in March 2020, the government implemented a series of dynamic containment measures, including lockdowns, curfews, school closures, and mandatory mask-wearing in public spaces. These interventions were continuously adjusted in response to the evolving epidemiological situation and the severity of outbreaks. Although geographical and socioeconomic inequalities in the health status of the Hungarian population long precede the pandemic [? ?], they played a critical role in shaping the pandemic's impact across Hungary. The first wave in early 2020 was swiftly controlled through robust measures aimed at suppressing the virus and preventing the healthcare system from being overwhelmed. Hungary acted decisively when community transmission began in Italy, adopting strong restrictions as the WHO declared a global pandemic. The second wave, which occurred during late 2020 and early 2021, proved more severe due to the emergence of the Alpha variant, resulting in higher morbidity and mortality rates. Communities in more deprived areas, particularly those with higher Roma populations, experienced disproportionately high mortality rates despite reporting fewer confirmed cases [136]. Although they lack the individual data needed to determine whether the Roma minority population experienced unproportionally higher death rates, the findings of Oroszi et al. align with reports from Roma rights advocates describing multiple fatalities within single families. In Hungary, the first COVID-19 vaccine arrived on December 29, 2020. To receive the vaccine, individuals registered online through a government portal. Patients were prioritized based on chronic illness, age, and occupation, and a specific vaccine type was offered depending on their characteristics and vaccine availability. Patients could either accept or decline the proposed vaccine. By Spring 2021, as more vaccines became available, individuals were able to choose from different options via a web application. This system provided alternatives for those hesitant about certain vaccines, helping to mitigate vaccine hesitancy [?]. Although six types of vaccines were widely available, the percentage of people receiving the primary series of COVID-19 vaccinations remained below the EU

average. The arrival of vaccines introduced hope but also revealed disparities in access and uptake, reflecting disparities driven by socioeconomic status, geographical location, and access to healthcare facilities. Socio-demographic characteristics were an important determinant of the observed vaccine hesitancy, although they remained relatively stable over time. Individuals in younger age groups and those with lower socioeconomic status were more likely to decline vaccination, while those living in the capital city were the least likely [121]. These findings were confirmed during the 4th and 5th wave [121]. Inequalities were also found at the geographical level, with vaccination coverage in deprived municipalities being 38% lower than the national average, thereby amplifying their vulnerability to severe outcomes [123]. During the third wave, residents in the most deprived municipalities had a 15–24% lower relative incidence of confirmed cases but faced 17–37% higher mortality compared to the national average [123].

2.2 The main determinants of human contact patterns

Human contact patterns represent the routes of infectious disease spreading by shaping the underlying transmission chain among susceptible individuals. During the COVID-19 pandemic, many aspects of human behavior have experienced drastic interruptions in most countries around the world. This was largely due to the implementation of non-pharmaceutical interventions (NPIs) that were installed to mitigate the spread and other effects of the pandemic. They aimed at controlling the number of contacts, as well as influencing individual attitudes, to change the ways humans meet and interact with each other. Their effects are evident in Fig. 2.1, where the average number of daily contacts in Hungary is shown. This value increases during interim periods (IPs) when the numbers of daily infection cases are low, and decreases during epidemic waves (Ws) when infection risk is high, this way sensitively reflecting the adaptive behavior of people throughout the pandemic.

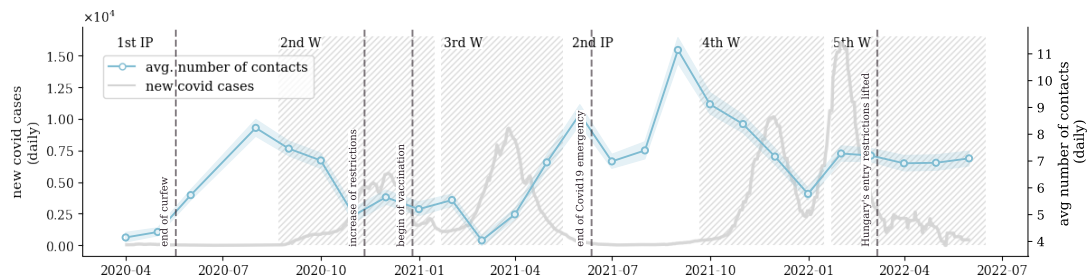


Figure 2.1: COVID-19 trends in Hungary: cases and contacts *Left axis:* number of new daily COVID-19 cases in Hungary from 2020/04 to 2022/07. *Right axis:* average number of daily contacts from 2020/04 to 2022/06 excluding household contacts. The values are shown as the median and interquartile range (*IQR*) of 1000 bootstrapped samples. The white and grey areas delimit the periods that have been aggregated in the analysis: two interim periods (*IPs*) (white areas) and four epidemic waves (*W*) (grey dashed areas).

Although at the aggregate level, these patterns are clear, there are non-trivial disparities at the level of individuals that may result in diverse contact patterns for given subgroups of the population. To investigate these disparities, we build a statistical analysis, where we focus on several socio-economic dimensions that, interacting with age, may significantly affect the number of contacts that individuals have.

2.2.1 Sociodemographic dimensions

The sociodemographic dimensions that we analyse are the following: (i) *education level*, which can have four possible levels: low, mid-low, mid-high and high; (ii) *employment situation*, which can be either employed or not-working, including students and retirees individuals; (iii) *perceived income* (called simply *income* through the manuscript) can have three possible levels: low, mid and high; (iv) *gender* refers to the biological gender and can be either female or male; (v) *settlement*, which refers to the area where individuals live and can be either capital, rural or urban; (vi) *chronic disease* is a Boolean dimension indicating if an individual is affected by any chronic disease; (vii) *acute disease* is a Boolean dimension indicating if an individual is affected by any acute disease, and (viii) *smoking* is a Boolean dimension indicating if an individual is a smoker or not. A detailed explanation of these variables is provided in Appendix B.1.1.

2.2.2 Statistical analysis

We model the expected number of contacts of respondents i using a negative binomial regression [36, 137] as defined in Eq. (2.1):

$$\mu_i = \alpha + \beta_1 \text{age_group}_i + \beta_2 X_i + \beta_3 \text{age_group}_i * X_i + \epsilon_i, \quad (2.1)$$

where age_group_i is the age class of i ; X_i is the variable of interest (e.g., education, income etc.), $\text{age_group}_i * X_i$ is the interaction term of age group and the variable of interest, and ϵ_i is the error term. Given μ_i , we define $\lambda_i = \exp(\mu_i)$ to be the expected number of contacts for respondent i . Then we model the reported number of contacts for respondent i , y_i , as

$$y_i \sim \text{Neg-Bin}(\lambda_i, \phi), \quad (2.2)$$

where $\phi \in [1, \infty)$ is a shape parameter that is inversely related to over-dispersion: the higher ϕ is estimated to be, the closest y_i 's distribution is to a Poisson distribution with rate parameter λ_i .

We build a model (2.1) for each variable of interest (X). Particularly, the interaction term between age_group_i and the variable of interest allows us to examine whether there are differences in the effect of X_i on the number of contacts in the different age groups. To be able to provide a meaningful description of the interactions, we analyse the average marginal effect (AME) [138–140] of X_i on the number of contacts for different age groups, defined as:

$$\begin{aligned} \text{AME}_{X_i} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \mu_i}{\partial \text{age_group}_i} \\ \frac{\partial \mu_i}{\partial \text{age_group}_i} &= \beta_1 + \beta_3 X_i. \end{aligned} \quad (2.3)$$

Working mainly with categorical variables (for example, education level or employment situation), we can calculate different AMEs for all categories of categorical variables in each age_group. It means, that for example for the interaction of the settlement type (with 3 categories, out of which one is a reference category) and age (with 5 categories) on the number of contacts, we have $2 \times 5 = 10$ different AME values: one for each settlement category by each age category. However, as we would like to know the overall effect of the interaction with the given variable, settlement, we have to aggregate these values into one measure. Therefore, we developed the following strategy. For each age group, we examined all the AMEs related to the category of the variable analyzed (e.g., all AMEs related to the categories of the settlement type). In each case, we calculated

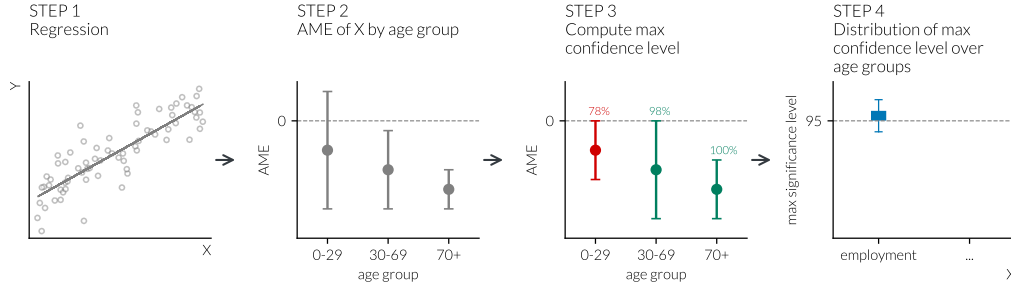


Figure 2.2: Maximum confidence level computation pipeline. Pipeline of the statistical analysis to compute the distribution of the *max confidence level* over age groups for each of the variables analyzed. The figure is made considering as an illustrative example the variable *employment*.

the confidence level [141], at which the confidence interval belonging to the AME of a given category reaches zero. The higher this calculated confidence level is, the more certain we can be that the given variable category has an effect on the number of contacts in the given age group. Equivalently, the estimated effect is significantly different from zero at a smaller type I error probability. Out of these confidence levels, we consider the maximum confidence level for each age group as that denotes the highest confidence level at which the AMEs of the given variable (e.g., settlement type) reaches zero, i.e. the smallest type I error probability [142] at which the estimated effect is significant. Finally, to summarise the results across age groups, we look at its distribution. By following this procedure for each of the variables of interest in the different waves of the pandemic, we are able to rank the variables according to their importance in driving differences in contact patterns additionally to age, in different periods of the COVID-19 pandemic. The pipeline of these analyses is illustrated in Fig. 2.2.

Results

In Fig. 2.3 we show the distribution of the *maximum confidence level* of the effects of these variables that they have on the number of contacts in interaction with age, during each period. Based on these results, employment, education and income level were found to be the three most important dimensions in determining the number of contacts. This observation stands if we consider the overall number of contacts including both work and community relationships, and it is true as well if we only consider community contacts (with results shown in Section B.2.1 of the Appendix).

Robustness analyses

Although, the data were aggregated to obtain NPIs-homogeneous periods, minor variations in restrictions persist within certain intervals (see Fig. B.3). Thus, acknowledging the importance of NPIs in shaping human contacts during the COVID-19 pandemic, we tested the validity of our results when considering this effect. Additionally, we account for the interplay of vaccination behavior and contacts, which can be significant in dynamic situations when vaccination strategies are implemented during an epidemic crisis [143]. Specifically, we have included two additional control variables in the statistical model (2.1): (i) the Oxford Stringency Index, a composite measure quantifying the government’s response strictness to the pandemic, and (ii) a dummy variable indicating individual vaccination status (vax_i). The results of these robustness analyses are pre-

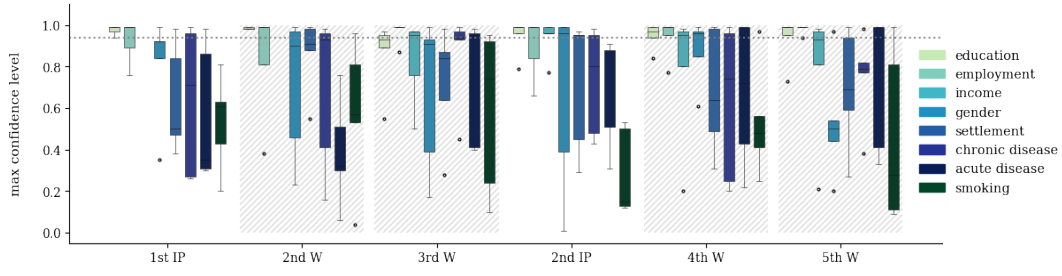


Figure 2.3: Key contact determinants during COVID-19 in Hungary. Box-plot (outliers, minimum, lower quartile, median, upper quartile and maximum) of the maximum confidence level at which the effect of the different categories of the variable on the total number of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for education level, employment situation, income level (which is present from the 2nd W, since this information was missing in the first few data collections), gender, settlement, chronic disease, acute disease, and smoking behaviour. The higher this value is, the more the variable influences people’s number of contacts given their age. The horizontal dotted line is placed at 95% and it represents the confidence level at which the AME is considered to be statistically significant.

sented in Section B.2.1 of the appendix, along with further robustness analysis where we control for all the other available individual features.

Average number of contacts by sub-groups

To further investigate the ways individuals of different characters adapt their number of contacts to the actual epidemiological situation, in Figure 2.4 we show the average number of contacts over time decoupled by education level (Fig. 2.4 *a, c*) and employment situation (Fig 2.4 *b, d*) for adult individuals older than 15 years (for the corresponding plots related to *income* and *settlement* see Section B.4.1, while for the plot decoupled by age groups, see Section B.4.2 of the Appendix).

The results in panel (*a*) suggest that high- and mid-high-educated individuals consistently have a higher number of contacts in the community layer throughout the observed period. In addition, the pattern of the average number of contacts of these groups suggests that probably they were able to better adapt to the epidemiological situation and NPIs by decreasing their contact numbers during epidemic waves and increasing again during interim periods. Conversely, individuals with mid-low and low education levels exhibit a lower and relatively stable number of contacts over time. Workplace contact dynamics suggest that only highly educated individuals (high and mid-high) were able to adapt to the epidemiological situation, while seemingly, those with lower education levels had less flexibility to adjust during different pandemic periods. Interestingly, mid-low educated individuals reported a higher number of workplace contacts, particularly during the second wave (*2nd W*) and the second interim period (*2nd IP*), attributable to their predominant vocational degree status and involvement in on-site interactive work. These patterns are confirmed when examining the relative number of contacts over time, as presented in Section B.4 of the Appendix.

When we group people by their employment situation, it makes sense to compare groups in the community layer. From Fig 2.4*b* it is clear that employed people maintain more contacts even outside of their workplace as compared to not-employed individuals, which is a clear sign of behavioural differences between these two groups. Meanwhile, employed individuals follow somewhat similar contact dynamics as highly educated peo-

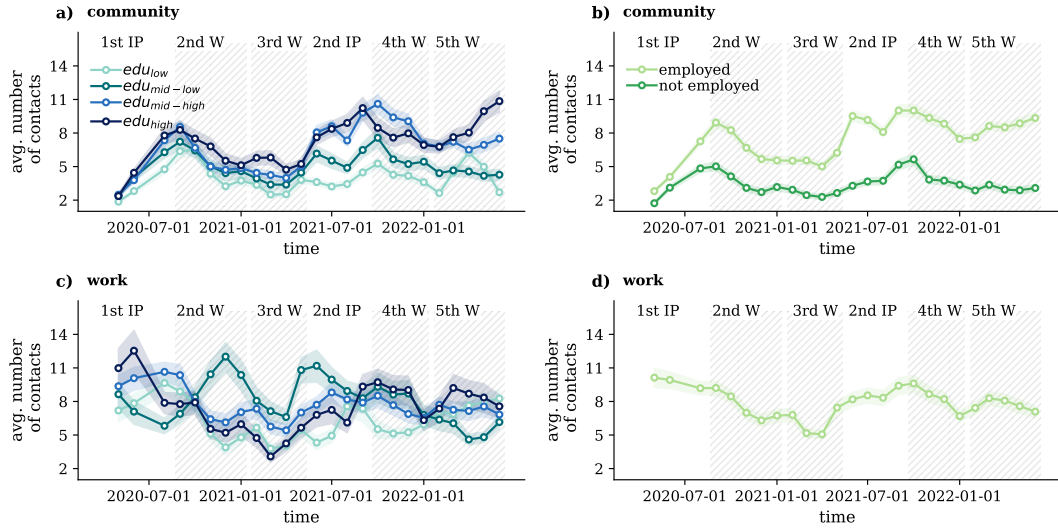


Figure 2.4: Contact dynamics by population subgroups. (a-b) Average number of contacts in the community layer (a) by different education levels and (b) employment situation. (c-d) Average number of contacts at workplace (c) by different education levels and (d) for employed people. All curves have been smoothed over the observation periods for better visualisation.

ple (see Fig. 2.4d), signalling some correlation between these two groups.

2.3 Beyond age-stratification: *the extended SEIR model*

We demonstrate that social inequalities significantly influence human contact patterns, thereby shaping the network of proxy social interactions. This is critically important for the propagation of diseases as it determines the transmission chain of an infection spreading among a susceptible population. Consequently, incorporating the contact pattern differences among individuals of different socio-economic backgrounds into epidemiological models is crucial. This could help to understand the unequal spread and uneven burden that an epidemic could impose on the different socio-demographic groups of a society. To this end, we propose a simple mathematical framework based on the extension of a conventional age-structured SEIR compartmental model (see Section 1.2.1), which we call the *extended SEIR* model. Instead of the traditional age-contact matrix C_{ij} (see Section 1.1.2, the proposed extended model incorporates $C_{\bar{d},i,j}$ age contact matrices, that are decoupled along important socio-demographic dimensions $\bar{d}$ to model epidemic spreading in different subgroups of the population.

2.3.1 Decoupled age-contact matrices

Conventionally, to compute age-contact matrices C_{ij} we divide a population into subgroups according to their age and calculate the average number of contacts that individuals in age class i have with individuals in age class j . Here, instead, we further stratify individuals from each age class i according to various dimensions, like employment status, settlement or education level.

In detail, we decouple the conventional age contact matrix C_{ij} into D number of matrices, one for each of the subgroups of the dimension that we want to take into account. More precisely, let $\bar{d}$ be the subgroup of the dimension considered and let

$\bar{d} \in 1, \dots, D$. We can write

$$C_{\bar{d}i,j} = R_{\bar{d}i,j} / N_{\bar{d}i}, \quad (2.4)$$

where $R_{\bar{d}i,j}$ is the total number of contacts that individuals of age class i and belonging to subgroup $\bar{d}$ have with individuals in age class j , regardless of the subgroup to which the contacted individuals belong; and $N_{\bar{d}i}$ is the total number of individuals in age class i and subgroup $\bar{d}$.

For example, to differentiate between *employed* and *not-employed* individuals, we compute two age contact matrices: $C_{\text{employed},i,j}$ and $C_{\text{not-employed},i,j}$.

This framework can be extended to any number of dimensions considered simultaneously, in this case, the length of the $\bar{d}$ vector will correspond to the number of combinations of the levels of the dimensions considered. Note that, the available data lack information on the subgroup membership of contacts, recording only the demographic details of survey respondents. Consequently, we opted to decouple the age contact matrices solely along the dimension of the respondent.

Symmetrization

Symmetrization of *decoupled age-contact matrices* poses a non-trivial challenge. Given a set of *decoupled* age-contact matrices along a given dimension d the reciprocity condition becomes: $\sum_d C_{d,ij} N_{di} = \sum_d C_{ji} N_{d,ij}$. Thus, the difficulty in symmetrization arises from the fact that we can not compute the transpose of a decoupled age-contact matrix. This is because the matrix is stratified only along the columns i.e. the participant's subgroup of appartenance.

To address this challenge, we implemented the following approach. For each of the decoupled age-contact matrices ($C_{d,ij}$) we compute a corresponding scaling matrices ($B_{d,ij}$). These matrices represent the proportions of contributions of each decoupled contact matrix to the aggregate one.

$$B_{d,ij} = C_{d,ij} N_{d,i} / C_{ij} N_i \quad (2.5)$$

Subsequently, akin to Equation (1.1), we computed the symmetric age-contact matrix at the aggregated level and we project back the symmetrized matrix (C_{ij}^{sym}) to the decoupled dimensions as follows:

$$C_{d,ij}^{sym} = \frac{C_{ij}^{sym} N_i \times B_{d,ij}}{N_{d,i}} \quad (2.6)$$

Derived decoupled age-contact matrices

In Fig 2.5a-f we show the age contact matrices decoupled by education level and employment situation ($C_{di,j}$) for the 4th epidemic wave for the adult population (See B.4.3 of the Appendix for the corresponding matrices including children). These matrices have been computed by considering *community*, *work* and *household* contacts together. The emerging large differences between these matrices demonstrate clearly that beyond age, the identified variables, i.e. education and employment status, induce significant differences in the contact patterns of individuals. Although these variables may not be independent of the age of people, the observed distinct patterns suggest more complex mechanisms controlling contact patterns among subgroups that can be explained by age alone.

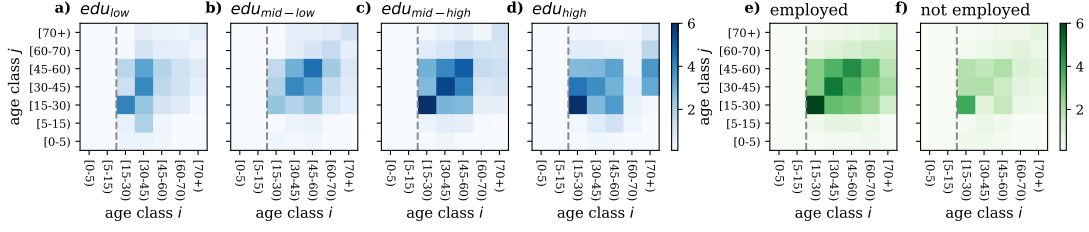


Figure 2.5: Age-contact dynamics by population subgroups. (a-d) Decoupled age contact matrices by education level for the 4th epidemic wave. (e-f) Decoupled age contact matrices by employment situation for the 4th wave. These figures depict contact numbers only for the adult population [15+), while matrices containing children are shown in Section B.4.3 of the Appendix. All the values are shown as the median and *IQR* of 1000 bootstrapped samples.

2.3.2 The *extended* SEIR model

In order to investigate the effect of the decoupled contact matrices on the dynamic of infectious disease transmission, we propose a simple mathematical framework as an extension of the conventional age-structured SEIRD compartmental model.

The model proposed here extends its definition by taking into account not only the age structure of the contacts in the population but also their differences along a set of other dimensions $\bar{d}$, such as education level, income level and employment situation. The model can be described by a set of ordinary coupled differential equations as presented in Eq. (2.7):

$$\begin{aligned}
 \partial_t S_{\bar{d},i}(t) &= -\lambda_{\bar{d},i}(t) S_{\bar{d},i}(t) \\
 \partial_t E_{\bar{d},i}(t) &= \lambda_{\bar{d},i}(t) S_{\bar{d},i}(t) - \epsilon E_{\bar{d},i}(t) \\
 \partial_t I_{\bar{d},i}(t) &= \epsilon E_{\bar{d},i}(t) - \gamma I_{\bar{d},i}(t) \\
 \partial_t R_{\bar{d},i}(t) &= \gamma I_{\bar{d},i}(t)
 \end{aligned} \tag{2.7}$$

Here i indicates the age group of the ego, j indicates the age group of the peer, $\bar{d}$ represents a vector of dimensions to which the ego belongs, β is the probability of transmission given a contact, ϵ is the rate at which individuals become infectious, γ is the recovery rate, and $C_{\bar{d}}$ is the age contact matrix corresponding to dimensions $\bar{d}$.

In this system of equations system the force of infection is defined as:

$$\lambda_{\bar{d},i}(t) = \phi \sum_j \frac{C_{i\bar{d},j}}{N_j} I_j(t), \tag{2.8}$$

In order to account for the variability of contacts in our data, for each simulation that we run we use a static decoupled contact matrix that we compute from a bootstrapped sample of our data (see Section B.3).

Results

We analyse the impact of decoupled age-contact matrices along four dimensions: employment situation, education level, settlement, and income level. Taking the decoupled contact matrices as input we simulate the spread of infectious disease among an entirely susceptible population using both the *conventional SEIR* and the *extended SEIR* models. Having fixed the epidemiological parameters such as the transmission rates and seeding strategy, other input parameters like the population distributions and contact matrices have been estimated from data as we explain in Section B.7 of Appendix in more detail.

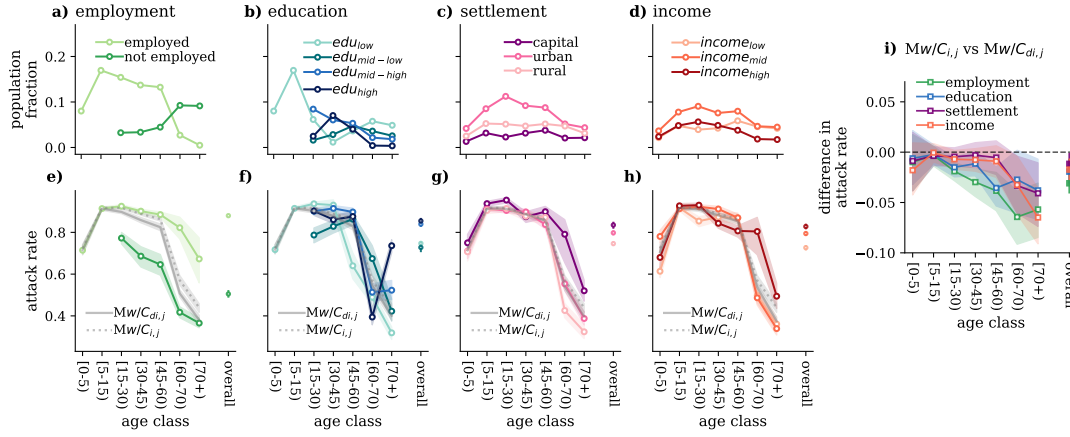


Figure 2.6: Differences in attack rates in extended SEIR models. (a-d) Survey population distribution by age and employment situation (a), education level (b), settlement (c) and income (d). (e-h) Attack rate by age and employment situation (e), education level (f), settlement (g) and income (h) as predicted by the *extended SEIR* (i.e. calculated with both age and the given socio-economic variable stratification). The grey lines represent the attack rate by age calculated only with age stratification as predicted by *classical SEIR with $C_{i,j}$* (solid lines) and *extended SEIR with $C_{di,j}$* (dotted lines). i) Difference in the attack rate by age as predicted by the *classical* ($Mw/C_{i,j}$) and the *extended* ($Mw/C_{di,j}$) model, when different dimensions are considered. Results are shown for the 4th wave. Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Simulations start with $I_0 = 5$ initial infectious seed. Results are shown as the median and IQR computed over 1000 simulations.

The proposed model allows us to investigate how differences in contact patterns along diverse social groups translate into an unequal burden of the epidemic. To quantify these differences in the epidemic outcome, we measure the attack rate defined as the population-wise normalised fraction of individuals who contracted the infection from a given group. To follow the distribution of the people along the investigated dimensions, we show the survey population fractions in the different age groups in Fig. 2.6a-d. Meanwhile, in Fig 2.6e-h we depict the attack rates calculated using the *extended SEIR* models for different age and socio-demographic groups (and as reference only for age - see grey solid lines). Results are shown for the cases when we decouple each age group along the four dimensions analyzed. As anticipated by the statistical analysis, employment and education produce the largest differences between groups in terms of attack rate by age. Interestingly, the group of employed people happened to be the most infected group in all age groups, while mid and high-educated individuals are more infected among those, who are 45-60 years old. When decoupling age contact matrices by settlement and income, although differences appear smaller between groups, high-income individuals and the ones living in the capital are more infected, particularly elderly ones with age 60+. These modelling results suggest the interesting overall conclusion that employed and wealthier group of the population, as well as those living in the capital report a higher attack rate, thus they are typically the most infected group relative to their population size.

These results also demonstrate that the *extended SEIR* model is able to capture differences introduced by the considered socio-demographic variable and, in this way, to model the epidemic impact on the different groups of the population. These differences are also visible at the population level. In Fig 2.6i we show the differences between the attack rates predicted by the *conventional* and the *extended SEIR* models, for each

age group and overall too. It is evident from these results that models using contacts only stratified by age may overestimate (negative difference) the size of the epidemic in different age groups or in the whole population. For example, our simulations based on data from the 4th wave demonstrate that the *conventional SEIR* model could predict higher attack rates for each age group with respect to the *extended SEIR*, which considers differences in contact numbers along the employment situation or the education level. (See Section B.7.1 of the Appendix for the corresponding figures for the other periods and the relative sensitive analysis). It is important to highlight that the uneven age distribution within the different subgroups sometimes reduces or annul the effect of the difference in the contact patterns when we are computing aggregate quantities at the population level. This explains why, even if there is a significant difference in contact patterns, the difference in the overall attack rates only spans a small range between the two models.

2.4 Vaccination uptake differences

Beyond the crucial role played by the network of face-to-face interactions, individual vaccination uptake may also substantially affect epidemiological outcomes by decreasing morbidity and mortality. Also in this case, different subgroups of the population may also exhibit varying vaccine attitudes, leading to lower vaccine uptake in certain groups.

2.4.1 Statistical analysis

Following the same methodology as describe above 2.2.2, we investigate the dimensions that, in interaction with age, affect the most the probability of getting vaccinated against COVID-19. In this case, we model this probability using a logistic regression model instead of a negative binomial, as the dependent variable was binary and not a count one. Namely, we model the probability of getting vaccinated for respondent i as defined in Eq. (2.9):

$$\log P_i(\text{vax} = 1) = \alpha + \beta_1 \text{age_group}_i + \beta_2 X_i + \beta_3 \text{age_group}_i * X_i + \epsilon_i \quad (2.9)$$

where age_group_i is the age class of i ; X_i is the variable of interest (e.g., education, income etc.), $\text{age_group}_i * X_i$ is the interaction term of age group and the variable of interest, and ϵ_i is the error term.

Despite the prioritization of specific groups, such as medical personnel, in the initial months of the vaccination campaign, we regard the associated impact as negligible for the purposes of this study. Interestingly, the statistical analysis in this case indicated income as the most important dimension along which individuals made different vaccination decisions (Figure 2.7).

As robustness analysis for this model, we investigate the effects of NPIs on vaccination uptake, with results presented in Section B.2.2 of the appendix. All the results of the robustness analysis aligns with those discussed earlier and lead to the same qualitative conclusions. By applying the same pipeline of statistical analysis as we explained above, we identify the dimensions along which individuals made different decisions in terms of vaccination, given their age. In this case, the interaction with age is particularly important given that the COVID-19 immunization strategy implemented in Hungary followed an age-stratified outreach by prioritizing elderly individuals [144, 145].

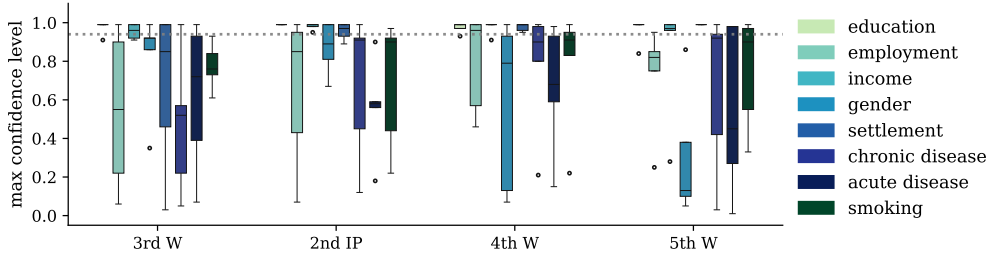


Figure 2.7: Key vaccination determinants during COVID-19 in Hungary Box-plot of the maximum confidence level at which the effect of the different categories of the variable on vaccination become significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*.

2.4.2 Investigating difference in vaccination within the *extended* SEIR model

To further investigate the difference shown by the statistical model, Fig. 2.8a-d shows the percentage of vaccinated individuals by age and the investigated dimensions, during the *4th* wave of the pandemic. Although by the *4th* epidemic wave the vaccination saturated in Hungary, the effects of the age-dependent vaccination policy are clearly visible. More strikingly, we find that higher socio-economic groups of the population were more likely to get vaccinated against the COVID-19 virus. This observation is valid for all age groups and periods considered in the analysis (See Section B.2.2 of the Appendix for the corresponding figures for the other periods).

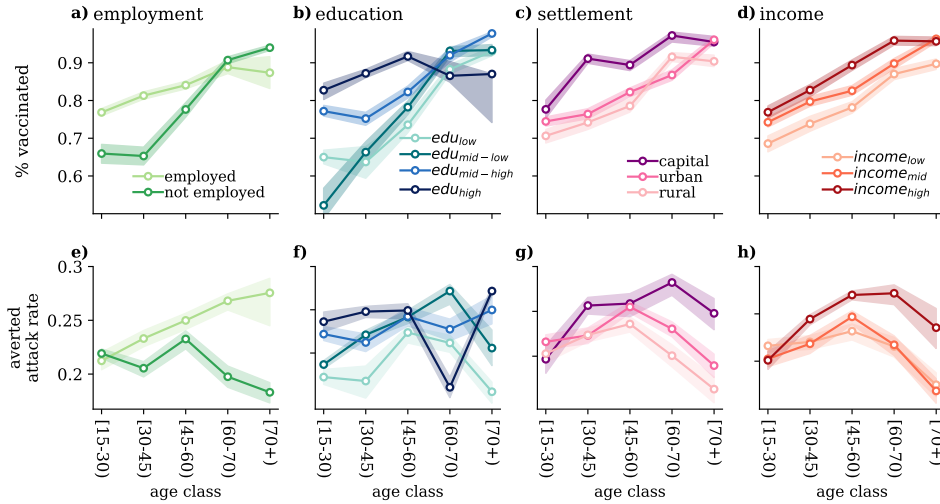


Figure 2.8: Disparities in COVID-19 vaccination and averted attack rates. (a-d) Fraction of individuals vaccinated with at least one dose against COVID-19 during the 4th wave, decoupled by age and (a) employment situation, (b) education level, (c) settlement, and (d) income. Panels (e-h) show the averted attack rates (difference between the attack rate in the non-vaccination scenario with respect to the vaccination one) by (e) age and employment situation, (f) education level, (g) settlement, and (h) income as predicted by *extended SEIR*. The model takes into account different rates of vaccination uptake by subgroups of the given variable, compared to the non-vaccination scenario. Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Simulations start with $I_0 = 5$ initial infectious seeds. Results are shown as the median and *IQR* computed over 1000 simulations.

To consider these observations, we model the vaccination uptake in the *extended SEIR* framework. More precisely, we define the probability of getting vaccinated (i.e. immune or recovered from the point of the infection) to be dependent in this case on both the age and the subgroup of the population considered. Using this *extended SEIR* model, we are able to compare the effects of vaccination uptake, while keeping fixed the structure of contacts. Fig 2.8e-h shows the *averted attack rate* due to vaccination with respect to the non-vaccination scenario. We consider the probability of getting vaccinated along the four different investigated dimensions separately. In all of these scenarios, the gain in averted infection is strongly dependent on the subgroup membership. As expected, the groups with higher vaccination uptake are the ones, which reduce their attack rate the most in the vaccination scenario. However, this pattern is not linear. For example, among individuals aged 60+, although the not-employed people report a higher vaccination uptake, they are the ones that gain less in terms of averted infections. This is because these individuals, having a low number of contacts, are already protected from exposure to the virus thus they gain less from vaccination. (See Section B.7.2 of Appendix for the corresponding figures for the other periods and the relative sensitive analysis).

2.5 Stratified modelling of the Hungarian scenario

To provide an example of how the proposed mathematical framework can be applied to a real case scenario, we model the 4th COVID-19 wave in Hungary between 09/2021 and 01/2022. As the statistical analysis showed that employment and income are the most important dimensions along which, respectively, contact patterns and vaccination uptake change the most, here we divide the population into subgroups by considering simultaneously these two additional dimensions other than age. In addition, we introduce a new compartment D to our SEIR model, that represents a dead state that infected individuals may enter with transmission rate $I \xrightarrow{\gamma IFR} D$.

To simulate the SEIRD for the 4th COVID-19 wave in Hungary we calibrate our model using the Approximate Bayesian Computation (ABC) method [146, 147] on the total number of daily deaths from 09/2021 to 01/2022 [148]. Details about the fitting method and calibrated results are summarised in Section B.8 of the Appendix.

The results of the simulated model are presented in Figure 2.9, which shows the daily fraction of newly infected (panels (a)-(c)) and new dead (panels (d)-(f)) cases for different employment, income, and age groups. As expected, these curves suggest that the group of employed people experienced the infection at a higher rate as compared to those not employed. At the same time, in terms of socio-economic status and age, more affluent and younger people got infected more during the simulated epidemic wave. On the other hand, strikingly the contrary trend is suggested in terms of mortality rates. From the simulations, we find that although not-employed, low-income and older individuals appeared with the lowest infection rates, they evolved with the highest mortality rate as compared to other groups.

Considering that the mortality rate for COVID-19 in infected individuals (IFR) depends on their age, this observation can be largely attributed to the fact that people not working and low-income are also the oldest in the population. To explicitly separate the effect of age when analyzing the burden of the epidemic, we examine the attack rates and mortality rates by age groups separately in each of the subgroups of the population stratified by income level and employment status (Fig. 2.9g-h).

As we have shown in the analysis above, these results confirm an overall decreasing

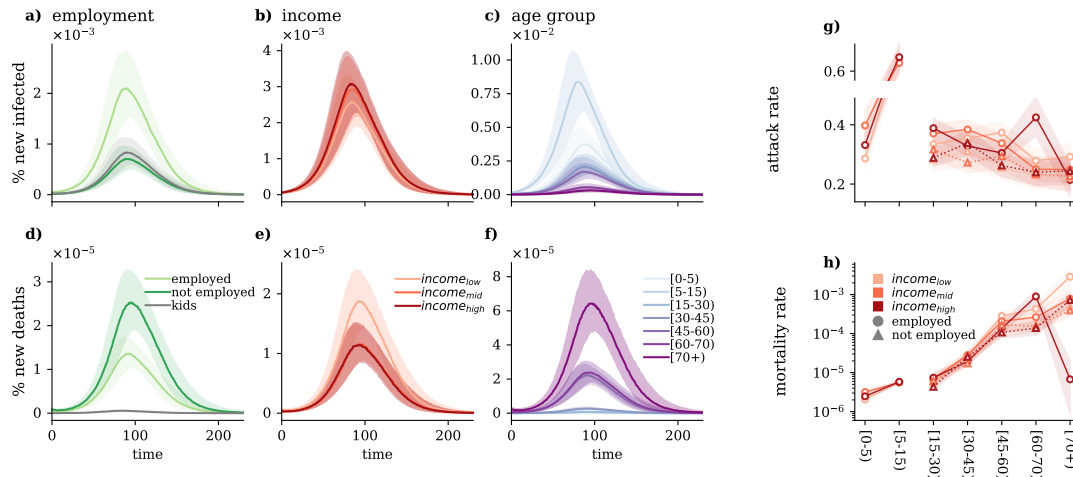


Figure 2.9: Modelled epidemic dynamics in Hungary during the 4th COVID-19 epidemic wave. Results from the simulated model for the 4th wave in Hungary. The figure shows the median and the IQR over 1000 runs. (a-c) Fraction of newly infected by (a) employment situation, (b) income level, and (c) age group. (d-f) Fraction of new deaths by (d) employment situation, (f) income level, and (g) age groups. (g) Attack rate by age, employment situation and income level. (h) Mortality rate by age, employment situation and education level. For more details about the parameters of the numerical simulations see SI.

infection rate by age and that not-employed individuals experience the lowest attack rate in each age group. The further stratification of not-employed individuals by income level reveals a clear pattern, with high-income people exhibiting a higher infection rate compared to mid-income and low-income individuals. In contrast, the infection pattern among the income levels of employed individuals is age-dependent. For young employed individuals in the age group 15-30, high and mid-income individuals register a higher share of infections, while among those older than 30, low-income individuals exhibit a higher infection rate compared to mid and high-income individuals. An exception is represented by the age group 60-70, for which the most infected group is still the high income.

In terms of mortality, as expected, we find an increasing trend by age; otherwise, we can conclude a similar pattern as for the attack rate. Particularly, the mortality rate by age group shows that employed people die at a higher rate. While, in terms of income, other than the group of age 15-30 and 60-70, the lower income people suffered more deaths according to our simulations. The extreme decrease in mortality rate for the employed high-income group is due to data sparsity in the survey data, recording only a few data points in this category (see Section B.1.2 of SI).

2.6 Discussion

The results presented in this chapter emphasizes the critical role that socio-economic factors play in influencing human behaviors relevant to the transmission of epidemics. Our findings suggest that contact patterns are shaped by social factors not only in their absolute values but also in the extent to which they vary in response to extraordinary events, such as lockdown or curfew interventions. Specifically, our statistical analysis shows that socio-economic factors such as employment situation and education level played a significant role in determining contact numbers and vaccination uptake during

the COVID-19 pandemic in Hungary. Additionally, in contrast to studies reporting a negative correlation between poverty and number of contacts [149], we find that more people with higher social-economic status tend to have a higher number of contacts and are the ones that change the most their number of interactions with respect to the epidemiological situation. Contrarily, people with lower social-economic status maintain a lower number of contacts over time with smaller oscillations. Although, we cannot establish a causal connection, as a possible, and plausible, explanation these results suggest that people from higher socio-economic groups, such as those with higher education, income, and employment status, were able to better adapt to the epidemiological situation and NPIs and were more likely to get vaccinated.

We propose a mathematical framework that extends the well-known age-stratified approach to model infectious diseases by explicitly accounting for differences in contact patterns and vaccination uptake for specific subgroups of the population. This method allows us to better understand the mechanisms underlying the emergence of inequalities in epidemiological outcomes. Results demonstrate that traditional epidemiological models, that only consider age, could overlook crucial heterogeneities along other social and demographic aspects that may impact the spreading of an epidemic. Through simulated epidemic processes, we show that significant differences in terms of attack rates arise from differences in contact patterns. By neglecting differences in vaccination uptake and the effects of vaccination campaigns among subgroups in modelling, we would miss important determinants, which significantly influence the outcome of an epidemic.

Through simulations of the COVID-19 pandemic in Hungary, we demonstrate that higher contact rates among wealthier and employed individuals lead to higher attack rates in these groups. However, their age structure and higher vaccination rates result in lower mortality rates compared to disadvantaged groups, who experience higher mortality. These findings align with empirical studies on the second and third waves of COVID-19 in Hungary, which show that individuals in deprived areas had a lower likelihood of being confirmed as COVID-19 cases but a higher risk of death, alongside lower vaccination uptake during the third wave. However, we recognize as a limitation of this work that, owing to data gaps, the initialization of population distributions within compartments and socio-demographic groups for the calibrated model is done only proportionally the population distribution. Indeed, the assumption of a homogeneous initial condition is improbable, precisely because of the different contact patterns, exposure risks, and vaccination uptake among different age and socio-economic groups (see Section B.8 of Appendix for further detail).

Additionally, we acknowledge the interplay between vaccination and contacts in mutually shaping each other, yet we have opted not to delve deeper into these mechanisms in the current work as they may fall beyond the scope of our study.

Finally, due to the limitation of the survey collection methodology, contact patterns of individuals can be differentiated only by the characteristics of participants. Indeed, the only information we know about the contacted peers is their age, while their other characteristics remain unknown. Thus, our extended *SEIR* model can only account for age-contact matrices that are decoupled along other social dimensions of the participants (*ego*). In other words, although our model incorporates additional social dimensions, given the subgroup the *ego* belongs to, it still only considers the average number of contacts stratified by the age group of the contacted (*alter*). In Chapter 3 we will tackle this limitation by introducing the *generalised contact matrix* which accounts for stratification along multiple socio-demographic dimensions of the contactee.

By shedding light on the complex interplay between social, demographic and eco-

nomic factors and disease transmission dynamics, our findings underline the need for a new mathematical framework for epidemic modelling that accounts for multidimensional inequalities. This would help us to better understand the socially stratified consequences of an epidemic and highlight non-negligible inequalities between different socio-demographic groups. Additionally, incorporating social factors into epidemiological models will provide a valuable tool to design and evaluate targeted NPIs to cope more efficiently with the spread of an infectious disease.

Chapter 3

Generalized contact matrices allow integrating socio-economic variables into epidemic models

Contact matrices have become an integral part of realistic epidemic models for respiratory infections. Interestingly, age and location are often considered jointly (See Section 1.6 for more details on contact matrices). However, despite their essential role, age and context are not the only important variables shaping contact patterns, disease dynamics, and epidemic outcomes. As described in Section 1.4 and 2.2, variables of socioeconomic status (SES), including income and education, significantly shape contact structures and impact the spread of infectious diseases. These factors are often overlooked in epidemic models, which typically stratify social contacts by age and interaction contexts.

In Chapter 2, we propose an initial modeling approach to incorporate these dimensions into epidemiological models by extending the traditional age-contact matrix to include other dimensions exclusively from the ego point of view. Here, we address this limitation by presenting an epidemic framework that features generalized contact matrices, stratified by multiple dimensions (i.e., variables).

As a first step, we study the mathematical properties of such a framework by considering age plus m other, unspecified, dimensions. We focus on prototypical SEIR compartmental models and derive a closed-form expression for the basic reproductive number R_0 . The derivation follows the next-generation matrix approach (See Section 1.2.2) applied to a flattened representation of generalized contact matrices. We quantify how much R_0 differs from the correspondent value computed in models including only one attribute (e.g., age). Interestingly, we prove a lower-bound theorem showing how increasing the number of dimensions in contact matrices cannot decrease R_0 . Furthermore, we prove that building generalized contact matrices via a random mixing assumption for the additional dimensions (i.e., contact rates are set proportional to the respective sub-group sizes) leads to the same R_0 of lower dimensional contact representations. Conversely, we numerically observe that correlations in contact patterns, such as mixing assortativity and heterogeneous activity, increase the value of R_0 . In the second step of our analysis, we highlight the importance of integrating SES dimensions into epidemic models and showcase how the proposed framework opens new avenues in this direction. Starting from synthetic scenarios, we show how adopting generalized contact matrices allows quantifying heterogeneities in epidemic outcomes and estimating the impact of NPIs across socio-economic strata. Indeed, it enables accounting not only for differences in contact patterns but also for heterogeneities in adherence to NPIs

across sub-groups of the population. We then apply our framework to empirical data from Hungary. We quantify the possible misrepresentation induced by neglecting SES in simulated outbreaks. The results confirm significant differences between models. The use of generalized contact matrices, that stratify contacts by age and one SES dimension, in agreement with the mathematical derivations, leads to higher values of R_0 for a given disease, to lower values of attack rates for a given R_0 , and allows capturing heterogeneity in disease's burden across SES groups.

This chapter is based on the paper [150].

3.1 Generalized contact matrices

We define *generalized contact matrices* as multidimensional objects that stratify contact patterns across any number of dimensions. Since age is often considered the key variable in traditional approaches, in what follows we focus on age plus other categorical dimensions. Hence, we move from the prototypical contact matrices $\mathbf{C}$, whose elements C_{ij} describe the contact rates between individuals in age bracket i and j , to $\mathbf{G}$, whose elements $G_{\mathbf{a},\mathbf{b}}$ capture the contact rates between individuals in sub-group $\mathbf{a}$ and $\mathbf{b}$. Here, $\mathbf{a} = (i, \alpha, \beta, \dots, \gamma)$ and $\mathbf{b} = (j, \eta, \nu, \dots, \xi)$ are index vectors (i.e., tuples) representing individual's membership to each category. We use Greek letters to highlight the difference between age and other dimensions, but the distinction among variables is not, in general, necessary. To provide a concrete example of a generalized contact matrix, we can imagine stratifying the population according to age ($i \in [1, \dots, K]$), income ($\alpha \in [1, \dots, V_1]$), and educational attainment ($\beta \in [1, \dots, V_2]$). In this case, the generalized contact matrix $\mathbf{G}$ would describe the average number of contacts that an individual in age bracket i , income α , and education β has with individuals in age group j , income η , and education ν in a given time window (see Figure 3.1). From this perspective, the matrix $\mathbf{C}$ can be thought of as an aggregation of contact rates at lower levels of stratification. In other words, the standard matrices can be viewed as aggregations along all dimensions affecting the organization of contacts but age.

The total number of contacts of individuals in a given age group i with others in j can be written as $R_{ij} = C_{ij}N_i$. Considering the generalized contact matrices we have $C_{ij}N_i = \sum_{\mathbf{a}', \mathbf{b}'} G_{\mathbf{a}', \mathbf{b}'} N_{\mathbf{a}'}$, where $\mathbf{a}' = \mathbf{a} - \{i\}$ and $\mathbf{b}' = \mathbf{b} - \{j\}$ are the index vectors m capturing all dimensions but age. Since the matrices $\mathbf{G}$ has $m + 1$ dimensions, we can aggregate contacts in $m + 1$ possible ways, for example computing the average contacts that an individual in one of the SES α has with others in the same SES β obtaining

$$C_{\alpha\beta}N_\alpha = \sum_{\mathbf{a}', \mathbf{b}'} G_{\mathbf{a}', \mathbf{b}'} N_{\mathbf{a}'} \quad (3.1)$$

where now $\mathbf{a}' = \mathbf{a} - \{\alpha\}$ and $\mathbf{b}' = \mathbf{b} - \{\beta\}$. In general, $C_{ij} \neq C_{\alpha\beta}$. Indeed, the number of groups and the number of individuals in each of them might be different.

We denote with K the number of age groups, while with V_p the number of groups in each m other dimensions (i.e., $p \in [1, m]$). While the generalized matrix $\mathbf{G}$ can be naturally described as a multidimensional matrix, the use of $T = K \prod_{p=1}^m V_p$ index vectors pairs allows for flattened representation in a squared bi-dimensional matrix of size $T \times T$. The formulation can easily include different contexts where interactions take place, applying (1.2), we can write $G_{\mathbf{a}, \mathbf{b}} = \sum_l \omega_l G_{\mathbf{a}, \mathbf{b}}^{(l)}$, where ω_l captures the relative importance of each social setting in the transmission.

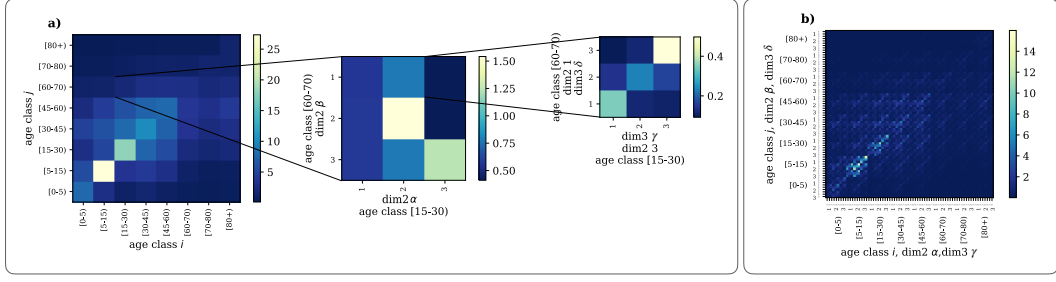


Figure 3.1: Generalized contact matrices. *a)* Schematic representation of generalized contact matrices. We can transform a $K \times K$ age-structured contact matrix $\mathbf{C}$ (first matrix on the left) into a generalized matrix $\mathbf{G}$ with $m + 1$ dimensions. Such transformation can be done according to different generalization schemes (See Sections 3.4, Appendix C.1 and C.7). In the plot, we consider $m = 2$. Hence, the second matrix describes the stratification of contacts across a second dimension i.e., $\alpha, \beta \in [1, 2, 3]$. The third matrix describes a further stratification, i.e., $\gamma, \delta \in [1, 2, 3]$. For simplicity, we show the stratification across the new dimensions only for one entry of the lower level, respectively. *b)* Example of a flattened generalized contact matrix including 3 dimensions $\mathbf{G}$ (72×72).

3.2 Epidemic model

As mentioned above, we study diseases whose natural history can be described by a SEIR compartmental model (See Section 1.2.1). Similar results can be obtained for others such as the SIS and SIR models. To include the generalized contact matrices $\mathbf{G}$, we define $X_{\mathbf{a}}$ as the number of individuals in groups $\mathbf{a}$ and compartment X where $X \in [S, E, I, R]$. The model dynamics is described by the set of differential equations:

$$\begin{aligned} d_t S_{\mathbf{a}}(t) &= -\Lambda_{\mathbf{a}}(t) S_{\mathbf{a}}(t) \\ d_t E_{\mathbf{a}}(t) &= \Lambda_{\mathbf{a}}(t) S_{\mathbf{a}}(t) - \Psi E_{\mathbf{a}}(t) \\ d_t I_{\mathbf{a}}(t) &= \Psi E_{\mathbf{a}}(t) - \Gamma I_{\mathbf{a}}(t) \\ d_t R_{\mathbf{a}}(t) &= \Gamma I_{\mathbf{a}}(t) \end{aligned} \quad (3.2)$$

where Γ is the recovery rate, and Ψ is the rate at which exposed individuals become infectious. In this framework, the force of infection can be written as follow:

$$\Lambda_{\mathbf{a}}(t) = \Phi \sum_{\mathbf{b}} G_{\mathbf{a},\mathbf{b}} \frac{I_{\mathbf{b}}(t)}{N_{\mathbf{b}}} \quad (3.3)$$

It describes the per capita rate at which susceptible individuals in categories $\mathbf{a}$ acquire the infection, and Φ is the transmissibility of the disease. To avoid confusion with the indices of the additional dimensions, the parameters regulating the disease dynamics are denoted with capital Greek letters.

3.3 Spectral properties of generalized contact matrices and the basic reproductive number

The basic reproductive number R_0 is one of the most important quantities of epidemiological relevance (See Section 1.2.2). The R_0 of a SEIR model, featuring a single-attribute contact matrix $\mathbf{C}$ defined, for example, by K age brackets, can be obtained using the next-generation matrix approach (See Section 1.2.2). The expression reads

$R_0 = \frac{\Phi}{\Gamma} \rho(\tilde{\mathbf{C}})$, where ρ denotes the spectral radius of the matrix $\tilde{\mathbf{C}} \in \mathbb{R}^{K \times K}$ whose elements are $\tilde{C}_{ij} = \frac{C_{ij} N_i}{N_j}$ (see the Section See 1.2.2 for the full derivation). As mentioned above, N_i and N_j describe the number of individuals in age-bracket i and j respectively. We can write $\tilde{\mathbf{C}} = \mathbf{R}\mathbf{N}^{-1}$ where $\mathbf{N}$ is a diagonal matrix whose elements are the number of individuals in each age group, and $\mathbf{R} = \mathbf{N}\mathbf{C}$ is a symmetric matrix describing the total number of contacts for each pair of age groups.

How does this expression change when we consider generalized contact matrices? As shown in detail in the Section C.2.2 of Appendix, the next-generation matrix approach can still be used to find the answer. The only extra step required is flattening the generalized contact matrix into a 2D matrix of size $T \times T$ (as defined above, $T = K \prod_{p=1}^m V_p$). In these settings, the expression for the basic reproductive number becomes $R_0 = \frac{\Phi}{\Gamma} \rho(\tilde{\mathbf{G}})$ where $\tilde{\mathbf{G}} \in \mathbb{R}^{T \times T}$ is the generalized contact matrix whose elements are $\tilde{G}_{\mathbf{a}, \mathbf{b}} = \frac{G_{\mathbf{a}, \mathbf{b}}}{N_{\mathbf{b}}} N_{\mathbf{a}}$. $N_{\mathbf{a}}$ and $N_{\mathbf{b}}$ describe the number of individuals in sub-groups $\mathbf{a}$ and $\mathbf{b}$ respectively. We can express this as the product of two matrices $\tilde{\mathbf{G}} = \mathbf{R}_{\mathbf{G}} \mathbf{N}_{\mathbf{G}}^{-1}$, where $\mathbf{N}_{\mathbf{G}}$ is a diagonal matrix describing the number of individuals in each sub-group and $\mathbf{R}_{\mathbf{G}} = \mathbf{N}_{\mathbf{G}} \mathbf{G}$ is a symmetric matrix describing the total number of contacts between any pairs of sub-groups.

The expressions for R_0 for single-attribute and generalized contact matrices are analogous. They both hinge on the spectral radius of a given matrix whose size and composition are generally different. Does this imply that the value of R_0 is also different? In other words, what happens to R_0 when we vary the number of variables used to stratify contact patterns? To answer this question, we can study what happens to R_0 when we increase the number of dimensions by 1. For example, let's consider the simplest case where we move from a single-attribute contact matrix to its generalization with an additional variable. In this case, but without loss of generality, the following theorem holds:

Theorem 1 (Spectral radius of generalized contact matrices). *Let $\tilde{\mathbf{C}} = \mathbf{R}\mathbf{N}^{-1} \in \mathbb{R}^{K \times K}$ be a single-attribute contact matrix where $\mathbf{R}$ is a symmetric, non-negative matrix, and $\mathbf{N}$ is a diagonal matrix, with positive diagonal elements. Let $\tilde{\mathbf{G}} = \mathbf{R}_{\mathbf{G}} \mathbf{N}_{\mathbf{G}}^{-1} \in \mathbb{R}^{KV_1 \times KV_1}$ be a two-attributes generalization of $\tilde{\mathbf{C}}$ defined such that $\sum_{\alpha, \beta=1}^{V_1} (R_{\mathbf{G}})_{i\alpha, j\beta} = R_{ij}$ and $\sum_{\alpha, \beta=1}^{V_1} (N_{\mathbf{G}})_{i\alpha, j\beta} = N_{ij}$ with $\mathbf{R}_{\mathbf{G}}$ being a symmetric, non-negative matrix, and $\mathbf{N}_{\mathbf{G}}$ being a diagonal matrix with positive diagonal elements. By denoting with $\rho(\mathbf{M})$ the spectral radius of a generic matrix $\mathbf{M}$ it follows $\rho(\tilde{\mathbf{G}}) \geq \rho(\tilde{\mathbf{C}})$.*

While the theorem explicitly treats a generalization obtained by shifting from one to two dimensions, the result holds for shifts to higher dimensions. In other words, adding any further attribute in the definition of the generalized contact matrix cannot decrease its spectral radius. The theorem does not make any assumption on the sizes K, V_1 and it can be easily adapted to generalized contact matrices with multiple dimensions. We discuss this extension along with the theorem's proof in the Supplementary Information of the paper [150]. These results are not further explored here as they were derived by the second author of the work, but I included them for the completeness of the narrative.

The theorem implies that a generalized contact matrix stratified across $m + 1$ dimensions has a spectral radius (hence basic reproductive number R_0) equal or higher than its s -dimensional counterpart ($1 \leq s \leq m$). In general however, as shown in the next sections, the variables used to capture the stratification of contacts and the way interactions are aggregated (i.e., by age or other dimensions) might affect the description of epidemic processes, especially in the presence of nontrivial correlations between

explicit and implicit (i.e., aggregated) variables. Hence, the estimation of the spreading potential of a disease via a model is not only a function of the total number of contacts but also of how these contacts are arranged across groups. This observation is one of the key insights from network-based epidemiology: the epidemic threshold (i.e., R_0) in two populations with the same number of individuals and number of contacts is drastically influenced by the way interactions are organized (i.e., the topology of the network) [65].

3.4 Numerical simulations

3.4.1 Synthetic generalized contact matrices

First, we test the analytical results obtained in the previous section by considering synthetic generalized contact matrices. To ground the results with empirical observations, we build on prototypical contact matrices, where interactions are stratified only according to age. Here, we rely on empirical data from Hungary using a pre-COVID-19 contact matrix stratified by age (See Section 1.3). We report analogous findings for a different country in Appendix C.5.

We developed a model to build synthetic generalized contact matrices $G_{\mathbf{a},\mathbf{b}}$. For simplicity and clarity of exposition let us consider only two dimensions: age and an additional variable, for example one SES indicator. Let us denote with K the number of age brackets and V_1 the number of sub-groups in the second dimension. In these settings $\mathbf{a} = (i, \alpha)$ and $\mathbf{b} = (j, \beta)$. The indices i and j refer to the age group while α and β to the SES of the ego and the alter respectively. The model takes as input an empirical contact matrix $\mathbf{C}$ describing the contact rates between age-brackets. As mentioned above, we can write $\mathbf{C} = \mathbf{N}^{-1}\mathbf{R}$ where $\mathbf{N}$ is a diagonal matrix (with positive diagonal elements) describing the number of individuals in each age bracket, and $\mathbf{R}$ is a symmetric matrix describing the total number of contacts between age groups in a given period. The generalized contact matrices $\mathbf{G}$ can be expressed as $\mathbf{G} = \mathbf{N}_{\mathbf{G}}^{-1}\mathbf{R}_{\mathbf{G}}$ where analogously $\mathbf{N}_{\mathbf{G}}$ is a diagonal matrix (with positive diagonal elements) describing the number of individual in each $\mathbf{a}$ and $\mathbf{R}_{\mathbf{G}}$ is a symmetric matrix capturing the number of total contacts between any two $\mathbf{a}$ and $\mathbf{b}$ in a given period. To build generalized contact matrices we need to define these two matrices. The first can be obtained by assuming a given population distribution for each $\mathbf{a}$. The second requires more attention. Indeed, the construction needs to respect two properties. The first is symmetry: $(R_G)_{i\alpha,j\beta} = (R_G)_{j\beta,i\alpha}$. In words, the number of contacts that individuals in age group i and SES α have with individuals in age group j and SES β , must be equal to the number of contacts that individuals in age group j and SES β have with individuals in age group i and SES α . This property implies that, for all $i = j$, $(R_G)_{i\alpha,i\beta} = (R_G)_{i\beta,i\alpha}$. In general this is not the case for $i \neq j$. The second property is that $R_{ij} = \sum_{\alpha,\beta}^{V_1} (R_G)_{i\alpha,j\beta}$. In words, by integrating across the second dimension we must obtained the initial matrix $\mathbf{R}$. For a given pair i and j $(R_G)_{i\alpha,j\beta}$ is of size $V_1 \times V_1$. Hence, for any $i \neq j$, we need a model to set a number $W = V_1^2 - 1$ of elements in the matrix. Note how the minus one stems from the constrain introduced by the second property. For all $i = j$ instead, the symmetry of the matrix is such that we need to set only $Y = V_1 + \frac{V_1(V_1-1)}{2} - 1$ elements. The first V_1 comes from the diagonal (i.e., $\alpha = \beta$), the $\frac{V_1(V_1-1)}{2}$ are instead the off-diagonal values, the minus one as before is due to the second property of such matrices. In order to reduce the number of parameters, we assume an independence between age and other dimensions. In this case, since $W > Y$ for all $V_1 \geq 2$, we can define the generalized contact matrix by setting W values. These can be obtained via a simple model where, for any i and j pair, i) each

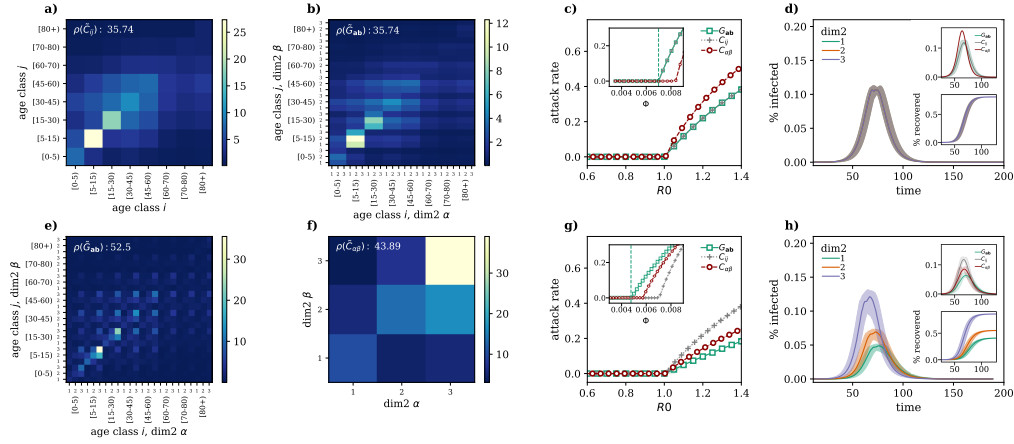


Figure 3.2: Impact of generalized contact matrices on epidemic spreading. (a) Example of a classic age contact matrix $\mathbf{C}$ (8×8); (b) and (e) generalized contact matrices with an additional dimension with three groups $\mathbf{G}$ (24×24) in the case of random mixing (b) and (e) assortative mixing in the second dimension; (f) depicts the case when we integrate the generalized contact matrix over all age classes. The ρ values indicate the spectral radius corresponding to each matrix. In panel (c), we show the attack rate as a function of R_0 (main figure) and disease transmissibility (inset). In panel (d) we show, in the case of random mixing, the prevalence (main figure and top inset) and the fraction of recovered as function of time (bottom inset); (g) and (h) are as the previous two plots but in case of assortative mixing with different levels of activity. Results refer to the median of 500 runs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, and $R_0 = 2.7$. Simulations start with a number $I_0 = 100$ of initial infectious seeds.

SES α is responsible for a fraction P_α of R_{ij} connections, ii) a fraction q_α of these are on the diagonal (i.e., in-group connections) and $1 - q_\alpha$ are instead off-diagonal. This parameter controls the assortativity of connections within each group. The fractions P_α and q_α are input parameters. By setting $V_1 - 1$ values P_α (the minus one stems from the constrain $\sum_\alpha P_\alpha = 1$) and V_1 values q_α , we obtain a model with $2V_1 - 1$ parameters. In case this number is equal to W , the constraints imposed by our assumptions allow us to define all the entries of the generalized contact matrix. If instead $W > 2V_1 - 1$ (guaranteed for any $V_1 \geq 3$), other $W - 2V_1 + 1$ free parameters (besides the values of P_α and q_α) are required for each i and j pair. We refer the reader to the Appendix C.1 for more details about synthetic generalized contact matrices.

Results

In Fig. 3.2-a we display the empirical contact matrix $\mathbf{C}$ for Hungary, which stratifies interactions in $K = 8$ age brackets. While until the 45-60 age group, the highest values of contact rates are within the same age bracket (i.e., diagonal elements), we observe strong off-diagonal values for age groups that range between 15 and 60 years. In Fig. 3.2-b we show the flattened 2D representation of a generalized contact matrix where we have a second dimension, besides age. We imagine a simple case where the second dimension contains three groups, i.e., $V_1 = 3$. For simplicity, we assume that 35%, 45%, and 20% of the population belong to these three categories across all age groups. The matrix is formed by $K \times V_1 = 24$ distinct groups.

Random mixing scenario First, we assume contacts- when it comes to the second dimension – to be proportional to the product of the population size in each group.

Hence, they are the expected values of a random mixing process. As mentioned above, this assumption leads to a generalized contact matrix with the same spectral radius as the original matrix. The effects of this property on an epidemic model, used to study the spread of a virus in the population, are shown in Fig. 3.2-c. We plot the attack rate (i.e., epidemic size) as a function of R_0 for i) a model fed only with the contact matrix C_{ij} (grey crosses), ii) a model fed with the generalized contact matrix $G_{\mathbf{a},\mathbf{b}}$ described above (green squares), iii) a model fed with the matrix $C_{\alpha\beta}$ where contacts are stratified only according to the second dimension (red circles). As a result, we observe, in perfect agreement with the analytical formulation, that $R_0 = 1$ divides the phase space into two regions. For values below this threshold, the disease does not spread in any of the three models. Only for values larger than one, the disease infects a finite fraction of the population. We observe that the threshold and the attack rates for different values of R_0 are equal in the first two models. As expected, when assuming random mixing along the second contact dimension, the results of a model that either neglects or considers such additional dimension are equal. Hence, the fraction of infected individuals in the groups of the additional category follows the same trajectory (see Fig. 3.2-d). We note, however, clear differences concerning the third model (red circles) that features a 3×3 contact matrix capturing the stratification of contacts only for the additional category. A description of the epidemic in these settings leads to higher attack rates for any given $R_0 > 1$. This observation confirms how within the same population the chosen stratification of contacts might affect the description of an epidemic unfolding in the system.

In the inset of Fig. 3.2-c we show the attack rate as a function of the transmissibility Φ . The vertical dashed line denotes the analytical critical value of Φ obtained by setting the expression for R_0 equal to one for the model featuring the generalized contact matrix. The plot shows the equivalence between the first model and the second as well as the difference with respect to the third. More precisely, the critical value of Φ in the model featuring the matrix $C_{\alpha\beta}$ is larger than the other two models. This implies a smaller value of R_0 , for a given value of Φ . The result is in line with the lower-bound theorem (Theorem 1). Indeed, $C_{\alpha,\beta}$ is a lower dimensional representation of $\mathbf{G}$. As such, its spectral radius is either equal or smaller. It is important to stress, however, that $C_{\alpha\beta}$ is obtained from $\mathbf{G}$ by aggregating across all age-dependent correlations. Hence, the corollary of the lower bound theorem, which is valid when the stratification along the additional dimension is obtained by random mixing, does not hold in general.

Assortative scenario We now consider the second scenario: assortative mixing. In Fig. 3.2-e we display the 2D flattened generalized contact matrix obtained by assuming that 60%, 50%, and 65% of the contacts in the first, second, and third group of the additional category take place within each group, this way inducing assortativity. Furthermore, we assume some levels of heterogeneity also in the activity of the different groups setting as 20%, 40%, and 40% the share of contacts of the three groups, respectively. Even a qualitative visual inspection of the generalized contact matrix shows how the introduction of assortativity and activity changes the contact rates (see panel Fig. 3.2-e) with respect to the previous case of random mixing (panel Fig. 3.2-b). This is confirmed by the value of the spectral radius which is about 46% larger, increasing from $\rho(\tilde{G}_{\mathbf{ab}}) = 35.74$ to 52.5. In Fig. 3.2-f we show the matrix $C_{\alpha\beta}$ obtained after integrating the generalized contact matrix over all age classes. As imposed by construction, the third group is characterized by a higher assortativity than the others. This group represents the smallest fraction of the population and is the most active (together with

the second group). These characteristics explain the high diagonal value in the matrix describing the contact rate between individuals in that group. As a result, the spectral radius is 24% higher ($\rho(\tilde{C}_{\alpha\beta})=43.89$) than the one of the contact matrix stratified by age (with $\rho(\tilde{C}_{ij})=35.74$). The impact of the different contact matrices on the simulated attack rate is shown in Fig. 3.2-g where we plot it as a function of R_0 . The figure shows in each case that the critical point falls at $R_0 = 1$, thus the analytical solutions match the numerical simulations very well.

Contrary to the previous case of random mixing, the attack rates of the model that is informed with the contact matrix C_{ij} and the one with the generalized matrix $G_{\mathbf{a},\mathbf{b}}$ are different for $R_0 > 1$. Interestingly, for a given R_0 the strong assortativity of contacts constrains the spreading of the disease, resulting in a smaller infected population fraction (see the green squares). Furthermore, the most active group is also the smallest one in this setting. A similar behavior is observed in contagion processes unfolding on explicit contact networks organized in clusters, where the local topological correlations might slow down the spreading of contagion processes that in turn do not evolve into an endemic state (i.e., SIR-like dynamics) [151]. If the contacts are stratified according to the second dimension only (red circles), the attack rates are closer to those emerging from the generalized contact matrices (green squares). In the inset of the figure, we show the attack rate as a function of the transmissibility Φ . The vertical dashed line denotes the theoretical prediction of its critical value when considering the generalized contact matrices. Due to the differences in the spectral radius of the various matrices, the critical values of Φ do not coincide. In the setting considered here, neglecting the assortativity and activity across the second dimension in favor of simpler representations focused only on age or the additional variable leads to a possibly large underestimation of the critical value of Φ . Assortativity and activity also introduce variations in the disease burden across the secondary dimension. Indeed, Fig. 3.2-h shows how, in these settings, individuals in the third class, are affected earlier and more intensely than the others. This is due to their high activity and assortativity. The result is confirmed by looking at the evolution of the fraction of recovered as a function of time (bottom inset of Fig. 3.2-h) where we see how individuals in the third class are significantly more affected by the disease. Finally, in the top inset of Fig. 3.2-h we show the prevalence for three models for a fixed value of $R_0 = 2.7$. Contact matrices stratified only by age feature a higher peak with respect to the other two. Even though these results are driven by the assumptions made when building the generalized contact matrix, they show how modeling outcomes, based on models that include or neglect further stratification of contacts besides age, might be extremely different. Interestingly, such differences cannot be observed *a posteriori*. For example, estimating the prevalence in a sub-group by simply considering its share of the population cannot account for differences in peak timing or heterogeneities in epidemic outcomes. In the absence of empirical contact data, it might be tempting to use sub-group sizes and build generalized contact matrices via a random mixing assumption. However, our results formally demonstrate that this would inevitably lead to the same epidemic dynamics of lower dimensional representations. In the Appendix C.6.1, we report an extensive exploration of the parameter space, confirming our analytical solutions' validity. We also include scenarios that consider additional dimensions (Appendix C.6.2). The results confirm that the analytical expression for R_0 provides a good description of the epidemic dynamics even in generalized contact matrices with three dimensions.

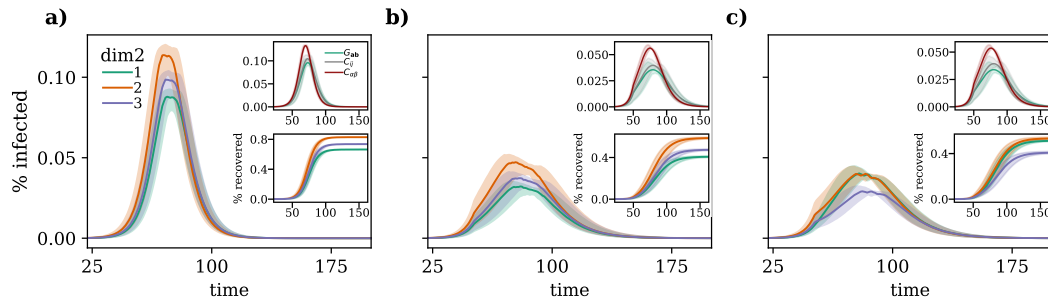


Figure 3.3: Modeling the adoption of non-pharmaceutical interventions (NPIs). Disease prevalence in (a) baseline scenario (b) and (c) with NPIs introduced at $t^* = 50$ that induce a reduction of 35% in total number of contacts with respect to the baseline scenario. In the case of (b) the three groups reduce the number of contacts equally, while in (c) the NPIs induce a redistribution of the number of contacts. Curves in the main plot indicate the prevalence for the three groups of the additional dimension. The top inset shows the prevalence for models featuring generalized contact matrices (green line), standard age-stratified contact matrices (gray line), and contact matrices stratified by the additional dimension (red line). The bottom inset shows the fraction of recovered as a function of time for the three groups of the additional dimension. Results refer to the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, and $R_0 = 2.7$. Simulations start with $I_0 = 100$ initial infectious seeds.

3.4.2 Modeling the adoption of non-pharmaceutical interventions

As mentioned above, epidemic models featuring generalized contact matrices are more expressive than traditional approaches. They allow capturing possible heterogeneities in behaviors that might span from NPIs adoption to vaccination uptake across diverse population sub-groups. Indeed, the COVID-19 pandemic has given a stark reminder that the ability to comply with NPIs or vaccine uptake is far from homogeneous. They correlate and are influenced by many socio-economic variables (See section 1.4.1). Standard models can only partially describe such heterogeneities, considering, for instance, contact reductions as a function of age and location (See Section 1.2.3).

To showcase the potential of generalized contact matrices in this context, we explore hypothetical scenarios where a given population, featuring an additional dimension besides age, changes behavior following the introduction of NPIs. We assume the ability to reduce contacts to be associated with the membership to a particular population group. We start with a pre-epidemic baseline where 33%, 50%, and 60% of the contacts in the first, second, and third groups of the additional category take place within each group. In the baseline 25%, 45%, and 30% are the shares of contacts of the three groups. For simplicity, we assume an equal population distribution across the three groups. In Fig. 3.3-a, we show the prevalence of a disease that spreads unmitigated by any NPI in our baseline population. Individuals in the second group are the most affected due to their higher activity. This is confirmed in the bottom inset, where we plot the fraction of recovered as a function of time. The second group is more affected than the other two. The top inset shows the estimation of the overall prevalence considering, as before, the three epidemic models featuring generalized contact matrices (green line), age-stratified matrices (grey line), and contact matrices stratified only for the second dimension (red line). In these settings, the last model features a higher peak, while the other two models are very similar.

Then, we imagine two scenarios where at time $t^* = 50$, due to NPIs, the population changes behaviors. (See Section C.4 for the details about the implementation) In Fig. 3.3-

b we show what would happen to the epidemic in case of a homogeneous reduction scenario when all groups would be able to reduce their contacts by 35%. The impact of the NPIs is quite strong and induces a clear reduction of the peak across the three groups of the population. The top inset of Fig. 3.3-b shows how, in this scenario, the overall prevalence, estimated with a model featuring age-stratified contact matrices (gray line), peaks a bit higher than the model fed with generalized contact matrices (green line), but they are overall very similar. The third model (red line), instead, is rather different as the peak is not only higher but also a bit earlier. This is another reminder of how the representation of contacts, the choice of which variables are considered or integrated, might affect the estimation of the epidemic dynamics. Moreover, as noted above, such differences in peak time cannot be observed by considering the additional dimensions only a posteriori. In the bottom inset of Fig. 3.3-b, we plot the fraction of recovered population as a function of time. We observe how the implementation of NPIs reduces the burden of the disease with respect to the baseline, but as before, individuals in the second sub-group of the additional dimension are the most affected.

In Fig. 3.3-c, we explore a scenario where the first group (e.g., representing for example the lowest SES class) is not able to protect itself as the other two manage to afford. We model this by reducing the overall number of contacts by 35% but assuming that the NPIs introduce changes in both assortativity and activity. We imagine that due to the NPIs 50%, 60% and 70% of the contacts in the first, second, and third group of the additional category take place within each group. Hence, across the groups, assortativity increases, but more significantly for the second and third group. Furthermore, we imagine that the NPIs shift activities to 37%, 37%, and 26% across the three groups. Hence, while activity decreases for the second and third groups, it relatively increases for the first. Observing the prevalence curves, we see how the NPIs bring a general reduction, but the changes in the contact patterns affect the first group more negatively. Indeed, we observe a switch from a baseline where this group was the least affected to a scenario where it becomes the most affected (together with the third group). This is confirmed in the bottom inset of the figure where we plot the fraction of recovered individuals as a function of time for the three sub-groups in the additional dimension. The top inset of the figure shows the overall prevalence for the three models which is very close to the previous.

3.4.3 Empirical generalized contact matrices

We built generalized contact matrices stratified in two dimensions, age and income, by using real data from the *MASZK* study (see Section 1.3).

The data provided us with a range of information about the anonymous participants including their perceived wealth with respect to the average, which is one SES variable we rely on. Information on social interactions have been collected in two different ways *i*) in an aggregate form, such that each participant reported the number of contacts they had with individuals in each of the eight age bracket considered, *ii*) in a diary in which each participant listed one by one the interactions they had on a given day by reporting some meta information of the *contactee* such as their age and SES. In particular, the average number of contacts of an individual in age class i , and SES α with an individual in age class j , and SES β is $G_{\mathbf{a},\mathbf{b}}$ where $\mathbf{a} = (i, \alpha)$ and $\mathbf{b} = (j, \beta)$. However, the *MASZK* data provided us with diary information only for the adult population (individuals older than 15 years old). For the children, their average number of contacts is available only in the aggregate form for the eight age brackets considered. Thus, we assigned the average number of contactees to the different SES as follow: $G_{i\alpha,j\beta} = G_{i\alpha,j}u_{\alpha\beta}$, where $G_{i\alpha,j}$ is

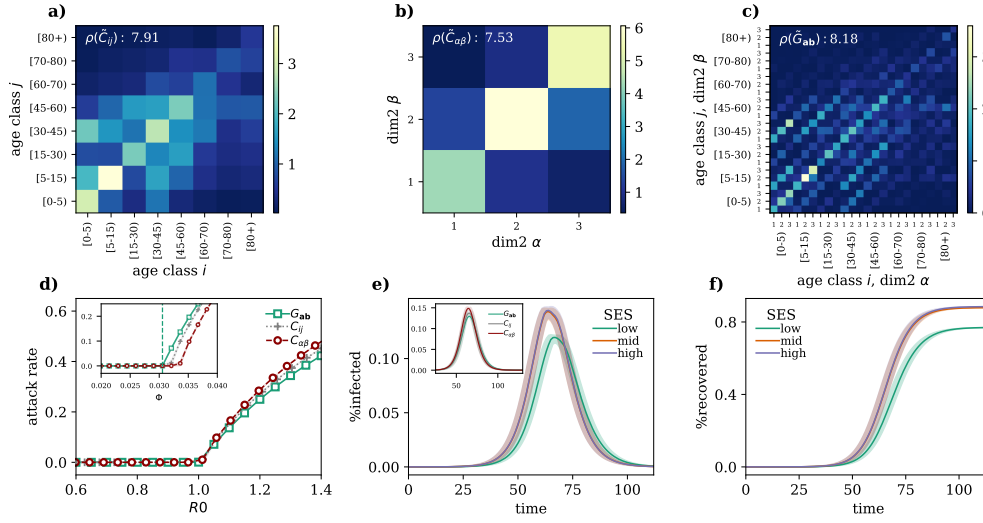


Figure 3.4: Empirical generalized contact matrices. (a) Age contact matrix $\mathbf{C}$ (8×8); (b) socio-economic status (SES) contact matrix $\mathbf{C}_{\alpha\beta}$ (3×3) and (c) generalized contact matrix with age and socio-economic status $\mathbf{G}$ (24×24); (d) attack rate as a function of R_0 (main figure) and transmissibility (inset); (e) disease prevalence in different SES groups (main figure) and different representation of the contact patterns (inset); (f) fraction of recovered individuals in different SES groups. In panels (a-c) ρ indicates the corresponding spectral radius. Results in panels (d) and (e) refer to the median of 500 runs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, and $R_0 = 2.7$. Simulations start with $I_0 = 100$ initial infectious seeds. The matrices have been computed using the contact diaries coming from the MASZK data collected in Hungary during June 2022.

the average number of contacts that individual of age group i and SES α has with all the individuals of age group j , and $u_{\alpha\beta}$ is a parameter that controls how these contacts are distributed among individuals of different SES.

Results

In Fig. 3.4-a we show the traditional contact matrix that focuses only on age. Though similar, the structure of contact patterns differs with respect to the equivalent matrix for the same country shown above in Fig. 3.2-a. This discrepancy is due to the different periods these matrices represent. While the matrix in Fig. 3.2-a has been collected before the COVID-19 pandemic, data in Fig. 3.4-a records typical contact patterns during the pandemic, specifically in June 2022. Although at that time the vaccination campaign was in full swing and the number of confirmed cases and deaths was relatively low, the previous wave had peaked only a few months before and some level of social distancing was still in place (for more details about the construction and normalization of this matrix see Section C.4 of the Appendix). The matrix confirms that children and young adults are the most active and that their interactions are largely assortative. However, the matrix features a rather smaller spectral radius with respect to the pre-pandemic contact matrix. This highlights the reduction in contacts caused by the COVID-19 emergency. In Fig. 3.4-b, we show the stratification of contacts considering only the SES variable, which divides the population into three SES groups with Low, Medium, and High wealth. Similar to other studies (See section 1.4.1), we observe a strong diagonal component, confirming high levels of assortativity within SES groups. The population experiencing mid-low SES features lower assortativity and activity. Furthermore, the off-

diagonal elements display higher levels of segregation for people experiencing low SES, as individuals experiencing mid and high SES tend to interact more with each other. In Fig. 3.4-c we show the flattened representation of the generalized contact matrix. $\rho(\tilde{G}_{\mathbf{a},\mathbf{b}})$ is higher than the other two, but the difference is clearly more limited with respect to the synthetic case we discussed in the previous section. In Fig. 3.4-d we show the attack rate as a function of R_0 for a disease spreading in a susceptible population described with each of the three contact matrices discussed. It is important to stress that these simulations do not aim to reproduce the evolution of the COVID-19 pandemic in Hungary. Instead, they describe a hypothetical disease spreading on different representations of the contact patterns measured in June 2022. Our goal is to showcase the possible variations in the description of an epidemic. The theoretical critical value for the basic reproductive number (being at $R_0 = 1$) well matches the numerical simulations. Furthermore, for a given value of $R_0 > 1$ the attack rate estimated with a model featuring the generalized contact matrix (green triangles) is always smaller than any of the two other models. Also in this case, the more realistic description of contacts across multiple dimensions constrains the unfolding of the disease with respect to scenarios that neglect one of the two dimensions. However, the model featuring the contact matrix stratified by SES only (red stars) leads to higher attack rates than the model featuring an age-stratified contact matrix. The inset confirms the validity of the mathematical formulation and highlights how a higher spectral radius results in a lower critical value for Φ . In Fig. 3.4-e, we plot the prevalence of the disease for the model with generalised contact matrix, splitting the three SES groups. Interestingly, we observe that individuals experiencing high SES are the most affected in terms of contracted infections. This is due to their higher contact activity. Those experiencing mid-SES follow very closely, partially due to the strong interaction activities with the first group. Individuals experiencing low SES, instead, are affected later and are subject to a lower prevalence. In this scenario, where the population is not subject to further NPIs nor spontaneously changing behavior during the epidemic, the higher level of segregation experienced by the low SES group has the silver lining effect of shielding that group from the epidemic, in accordance with empirical observations in the country [123]. These results are confirmed in Fig. 3.4-f where we plot the fraction of recovered as function of time for individuals in different SES groups. In the inset of Fig. 3.4-e, we show the overall evolution of the disease obtained by modeling the epidemic with each of the three contact matrices. The plot confirms how the chosen representation of contacts influences the description of the disease. Considering different numbers or types of dimensions, in this case, induces differences in the estimation of the peak, which is a key variable used to inform public health measures.

In the Appendix C.7, we repeat the same analyses considering data collected at different time points of the COVID-19 pandemic, in April and November 2021. While the overall results confirm the picture that emerged here, models fed with standard and generalized contact matrices lead to estimations of the attack rates, for a given R_0 , which are closer to each other. Nevertheless, the model featuring generalized contact matrices allows capturing heterogeneities in the prevalence across SES groups, which are invisible to standard approaches and to analyses conducted integrating SES dimensions only at posteriori.

3.5 Discussion

In this chapter, we presented a mathematical epidemic modeling framework that features generalized contact matrices, that is, contact matrices that are stratified according to

multiple dimensions, such as variables linked to socioeconomic status (SES). The goal of our work is to highlight the importance of moving beyond the traditional representation of mixing patterns, based on age-specific and/or context-specific contact rates, allowing for the development of structured epidemic models that incorporate multiple groups of the population at once.

First, we focused on the mathematical characterization of the modeling framework. We found that the basic reproductive number, R_0 , of a model featuring generalized contact matrices can still be obtained analytically via the next-generation matrix approach developed for traditional compartmental models. We then introduced a lower-bound theorem proving how the generalization to multiple dimensions leads to a value of R_0 , which is necessarily equal to or larger than for models featuring fewer grouping dimensions. Interestingly, we showed, numerically, how non-trivial correlations of contacts across and within groups, such as assortativity and heterogeneous activity, increase the value of R_0 . We then explored how the modeling framework can be used to integrate and consider explicitly SES dimensions. Through numerical simulations, we showed how models featuring generalized contact matrices better capture heterogeneous behaviors across population sub-groups, such as varying rates of adherence to NPIs across socio-economic strata, as observed in reality (See Section 1.4.1). By using both synthetic and empirical data from Hungary, we demonstrated how neglecting SES dimensions in favor of simpler representations of social contacts may result in large misrepresentations of attack rates, epidemic thresholds, and disease burden across different population sub-groups. Moreover, our results highlight how such misrepresentations cannot be simply corrected by integrating SES dimensions *a posteriori*.

As for any structured epidemic modeling approach, our model requires data to parameterize social contact rates across subgroups. Social contact surveys have been and will continue to be an important asset for the study of mixing behaviors both in normal conditions and during epidemic outbreaks (See Section 1.3.1). Future contact surveys could include additional demographic and socio-economic dimensions beyond the usual age/gender components. The work we presented here shows how such dimensions can be easily integrated by generalizing traditional epidemic models. In some cases, large-scale contact surveys including several population attributes can be unfeasible, especially in resource-poor settings. Alternative approaches could infer contact patterns from the analysis of demographic data that are routinely collected by census surveys (See Section 1.3.2). Other proxies of contact rates, derived from mobility data, could be similarly used to infer contact patterns among different socio-economic groups, as demonstrated by recent studies focusing on experienced segregation in large US metropolitan areas [152, 153].

It is important to mention the limitations of our work and the directions for future developments of our study. Admittedly, the model we developed to generate synthetic generalized contact matrices does not attempt to reproduce empirical observations from real data, but rather to provide a general test bed for investigation. This choice was guided by the lack of available data about the stratification of contacts across multiple dimensions in different countries. As described, we had access to empirical data only for Hungary. Arguably, one country is not enough to develop a general model and more work is needed to design and test other approaches. Similarly, the methodology we used to simulate the effects of NPIs was driven by simplicity rather than realism. We intended to showcase the potential of our model to capture possible heterogeneities in behaviors rather than reality. Data from Hungary contained limited information about the social contacts of children. Hence, we had to introduce a few assumptions about

their structure. Another important assumption we made is about the independence of age and other socio-economic dimensions. While this assumption was necessary to showcase our methodology in simple terms, future work is needed to investigate the impact of such correlations on the description of epidemic processes and to explore the applicability/extension of methods for dimensionality reduction.

Chapter 4

Social inequalities in vaccine coverage and their effects on epidemic spreading

Vaccination is one of the most effective public health interventions for controlling infectious diseases and protecting individuals and communities. While vaccines hold the promise of accessibility regardless of social status, numerous studies demonstrate that inequalities manifest themselves in the uptake of vaccinations. Socioeconomic disparities play a significant role in determining who gets vaccinated and who does not, thus influencing the overall effectiveness of vaccination campaigns. The mechanisms behind these disparities are multifaceted and rooted both in structural and behavioural factors (See Chapter 1.4.3).

Furthermore, for any given uptake and coverage level the impact of a vaccination campaign is closely tied to its timing and speed relative to the progression of the epidemic. A campaign that begins only after a large portion of the population has already been infected will inevitably have a much smaller effect compared to one initiated earlier. During the COVID-19 Pandemic, vaccinations started in December 2020 (in the Global North) [21, 154]. At that time many countries were facing a rise in infections induced by the winter seasonality and by the spread of a more transmissible variant (i.e., Alpha). Furthermore, doses were a scarce resource and countries had to deal with tremendous logistical challenges for their distribution. As a result, the initial vaccination rates were far from optimal. Hence, the beginning of COVID-19 vaccinations is a clear example of a campaign that starts amid an epidemic and finds a large majority of the population susceptible. Non-pharmaceutical interventions (NPIs) have been critical to support these complex phases. However, NPIs have also been associated with socioeconomic inequities. Indeed, the ability to self-isolate, work from home, and avoid crowded places varied significantly across different socioeconomic groups (See Chapter 1.4.1). As a result, disadvantaged groups of the population suffered from compounding inequities. On one side, they had limited access to vaccines. On the other, they could not afford to reduce the risk of infection by adopting NPIs. These disparities have significant implications for public health, leading to sub-optimal vaccination coverage and perpetuating the spread of vaccine-preventable diseases.

In this context, incorporating socioeconomic factors in epidemic models beyond the traditional age-wise stratification is essential [12, 13, 15, 41, 155, 156]. However, current approaches fail to do so potentially resulting in incomplete or biased predictions about epidemic burden and vaccination campaign effectiveness. To address these complexities

and limitations, here we adopt and extend the epidemic framework proposed in Chapter 3 that features *generalized contact matrices*. The model allows accounting for the stratification of contacts across multiple dimensions such as age and SES [150]. In doing so, it provides a finer description of the population under investigation and allows to assess how disparities in vaccination coverage and NPIs adoption influence the overall epidemic outcome. To isolate the interplay between social contacts and vaccination uptake across subgroups we consider both synthetic and real generalised contact matrices that stratify contacts according to age and one SES indicator. The age stratification is data-driven in both cases, while the stratification for the SES dimension is built either by using a simple model or from data collected in Hungary (See Appendix B.1).

To account for the effect of vaccination timing, these analyses are developed in two distinct scenarios. The first simulates a situation where the vaccination campaign and the epidemic are concurrent. The second scenario instead simulates a situation in which the vaccination campaign begins before the epidemic, reaching a substantial portion of the population prior to the outbreak.

Our results emphasize the importance of models that explicitly account for socioeconomic disparities to accurately estimate the epidemic burden within specific subgroups of the population. Indeed, we demonstrate that, when disparities in vaccination uptake exist across subgroups, models that do not incorporate these dimensions fail to accurately predict the epidemic burden within these groups. Furthermore, we find that the unequal distribution of vaccines across socioeconomic groups might lead to a non-linear effect on epidemic outcomes that, in general, cannot be estimated a posteriori from models that account just for the traditional age stratification. Finally, our findings show how disparities intensify when compounded by the interaction with NPIs.

Overall, our results underscore the importance of incorporating subgroup-specific factors into epidemic models to enhance the accuracy of predictions and, in turn, the effectiveness of public health interventions and vaccination strategies.

This chapter is based on the paper [157].

4.1 Epidemiological model with generalized contact matrices and vaccination

We consider a Susceptible-Exposed-Infectious-Recovered (SEIR) compartmental model (See Chapter 1 for further details) incorporating multidimensional contact matrices with an additional dimension, with three levels (i.e., subgroups), besides (ie, generalized contact matrices presented in Chapter 3).

Each compartment is divided between non-vaccinated NV and vaccinated V , thus, the total number of individuals in compartment C is $C = C_{NV} + C_V$. Individuals who received one dose of vaccine move to the compartments denoted with the superscript V . We assume all individuals except for the infectious and deaths can receive the vaccine. The number of people that get vaccinated at each time step is defined by $\Omega(t)$, and it is distributed among S_{NV} , E_{NV} , R_{NV} , proportionally to their share in the population. The effectiveness of the vaccine is described by g_1 , which represents efficacy against infection. Additionally, we assume that infected vaccinated individuals are less likely to have further infections, this is modelled by the parameter VE in the force of infection.

In particular, the SEIR model with vaccination featuring generalized contact matrices considers the population sliced in $m + 1$ dimensions the epidemic dynamics are encoded in the following set of differential equations:

$$\begin{aligned}
d_t S_{NV\mathbf{a}}(t) &= -\Lambda_{\mathbf{a}}(t) S_{NV\mathbf{a}}(t) - \Omega(t)_S, \\
d_t S_{V\mathbf{a}}(t) &= -(1 - g_1) \Lambda_{\mathbf{a}}(t) S_{V\mathbf{a}}(t) + \Omega(t)_S, \\
d_t E_{NV\mathbf{a}}(t) &= \Lambda_{\mathbf{a}}(t) S_{NV\mathbf{a}}(t) - \Psi E_{NV\mathbf{a}}(t) - \Omega(t)_E, \\
d_t E_{V\mathbf{a}}(t) &= (1 - g_1) \Lambda_{\mathbf{a}}(t) S_{V\mathbf{a}}(t) - \Psi E_{V\mathbf{a}}(t) + \Omega(t)_E, \\
d_t I_{NV\mathbf{a}}(t) &= \Psi E_{NV\mathbf{a}}(t) - \Gamma I_{NV\mathbf{a}}(t), \\
d_t I_{V\mathbf{a}}(t) &= \Psi E_{V\mathbf{a}}(t) - \Gamma I_{V\mathbf{a}}(t), \\
d_t R_{NV\mathbf{a}}(t) &= \Gamma I_{NV\mathbf{a}}(t) - \Omega(t)_R, \\
d_t R_{V\mathbf{a}}(t) &= \Gamma I_{V\mathbf{a}}(t) + \Omega(t)_R,
\end{aligned} \tag{4.1}$$

where $\mathbf{a} = (i, \alpha, \dots, \gamma)$ is the index vector that describes the membership of individuals in the $m + 1$ groups, and the subscripts NV and V indicate respectively the non-vaccinated and vaccinated individual in each compartment. The force of infection, defined as the per-capita rate at which susceptible are exposed to the disease, can be written as:

$$\Lambda_{\mathbf{a}}(t) = \Phi \sum_{\mathbf{b}} G_{\mathbf{a},\mathbf{b}} \frac{I_{NV\mathbf{b}}(t) + (1 - VE) I_{V\mathbf{b}}(t)}{N_{\mathbf{b}}} \tag{4.2}$$

where Φ is the transmissibility of the disease, VE is the vaccine efficacy against further transmission.

4.1.1 Modeling vaccination uptake among the second dimension

The daily allocation of vaccines to age groups is taken from data in Hungary reflecting the COVID-19 vaccination campaign. However, vaccination data on the daily doses administered is available, at the lower level, by age groups (See Appendix D.1) which we distribute among subgroups (SES) according to a given distribution $P_{vax}(\alpha)$, which indicates the proportion of vaccines allocated to each subgroup. Let $\Omega_i(t)$ represent the daily administered doses at time t for age group i . The number of vaccines administered to a subgroup α is then $\Omega_{i,\alpha}(t) = \Omega_i(t) \cdot P_{vax}(\alpha)$. We assume that each subgroup can be vaccinated up to 95%. If this limit is reached, any excess vaccinations are redistributed randomly within that particular age group.

4.2 Inferring epidemic burden for subgroups

The generalized model allows us to estimate disease burden across the three socio-economic subgroups by aggregating outcomes by age and SES, summing across all age groups. For instance, the total number of infected individuals in subgroup α at time t can be easily computed as the sum across age groups: $I_{\alpha}(t) = \sum_i I_{i,\alpha}(t)$. One of the questions that we are interested in assessing in this work is whether and to what extent it is possible to approximate these subgroup-specific outcomes using models that lack explicit SES stratification. Indeed, without explicit inclusion of the SES dimension, subgroup infection numbers can only be approximated *a posteriori* by utilizing outputs from traditional models, such as attack rates for vaccinated (AR_V) and non-vaccinated (AR_{NV}) populations.

To estimate the epidemic burden for specific SES subgroups α using outputs from a model stratified solely by age, we first need to infer the epidemic burden within each

age group in the subgroup (i, α) and subsequently aggregate these age-specific results to obtain AR_α . This requires estimating the number of susceptible individuals in each subgroup who became exposed during the epidemic i.e., $M_{S_{i,\alpha} \rightarrow E_{i,\alpha}}$. This involves calculating the attack rates for vaccinated AR_{Vi} and non-vaccinated AR_{NVi} individuals, in each age group i , from the age-stratified model (See Appendix D.2 for further details on this calculation). Therefore, we can multiply AR_{NVi} and AR_{Vi} by the total number of non-vaccinated and vaccinated individuals in each group (i, α) at time t , taking into account the transitions between non-vaccinated and vaccinated states over time. The equation used for this calculation reads as follows:

$$M_{S_{i,\alpha} \rightarrow E_{i,\alpha}} = AR_{Vi}(t)M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t) + AR_{NVi}(t)(N_{i,\alpha} - M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t)), \quad (4.3)$$

where $AR_{Vi}(t)$ and $AR_{NVi}(t)$ represent respectively the attack rates for vaccinated and non-vaccinated individuals in age group i and $M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t)$ denotes the number of susceptible individuals in groups (i, α) who were vaccinated at time t . The attack rate for subgroups (i, α) can be then calculated by dividing equation (4.3) for the population size as follows:

$$AR_{i,\alpha} = \frac{AR_{Vi}(t)M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t) + AR_{NVi}(t)(N_{i,\alpha} - M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t))}{N_{i,\alpha}} \quad (4.4)$$

We can then aggregate by age groups equation (4.4) and divide by the total number of individuals in subgroup α as follows:

$$AR_\alpha = \frac{\sum_i AR_{Vi}(t)M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t) + AR_{NVi}(t)(N_{i,\alpha} - M_{S_{NVi,\alpha} \rightarrow S_{Vi,\alpha}}(t))}{N_\alpha} \quad (4.5)$$

This method, however, has two primary limitations: (1) it assumes accurate knowledge of the number of vaccinated individuals within each age and SES-defined subgroup, which may not be feasible in practical settings; (2) it inherently cannot capture subgroup-specific dynamics, such as differing peak times, that the model does not explicitly represent.

In the following scenario analysis, we will evaluate the performance of this approximation technique, which we will indicate as $f(C_{ij})$ ¹, against direct predictions from the generalized model, indicated as G_{ab} .

4.3 Scenario analysis

As mentioned in the introduction, the interaction between the timing of a vaccination campaign and the progression of the epidemic plays a critical role in shaping the epidemic burden. To investigate their effect on the epidemic outcome, we focus on two simulation scenarios. In the first (*Scenario 1*), the epidemic and vaccination campaign start simultaneously. We set the basic reproduction number R_0 to 1.6, representing a relatively slow epidemic and leading to a substantial overlap between the epidemic wave and the vaccination campaign. In the second scenario, (*Scenario 2*), the vaccination campaign is completed before the epidemic begins. Here, contrary to the previous scenario, we set

¹The f describes the a posteriori estimation of disease burden from traditional models featuring only age-structured contact matrices C_{ij}

a relatively high value of the basic reproductive number, $R_0 = 3$, to simulate more rapid epidemic dynamics. The higher R_0 allows us to observe the rapid spread of the virus in a predominantly vaccinated population.

In all the experiments, we take into account eight age classes, $[0-5)$, $[5-15)$, $[15-30)$, $[30-45)$, $[45-60)$, $[60-70)$, $[70-80)$, $[80+)$, and three SES: *low*, *middle*, and *high*. The stratification for age is obtained from data from the MASZK data (See Appendix B.1), while the stratification for SES is either synthetic, obtained from a simple model, or measured from data. In the first case, we introduce assortativity by assuming that 60%, 50%, and 65% of the contacts in the first, second, and third SES categories take place within each group. Furthermore, we assume that activity levels are distributed heterogeneously, with 20%, 40%, and 40% of the total contacts allocated to the first, second, and third groups, respectively. We set an uneven population distribution with group sizes comprising 35%, 45%, and 20% of the total population. The model used to generate the synthetic population and mixing patterns along the second dimension follows the one presented in Chapter 3 (See Appendix C.1 for further details).

We study and compare three different vaccination uptake distributions (VD for short) across the SES subgroups. In the first case ($VD1$), each subgroup receives an equal share, i.e., one-third of the vaccines. In the second case ($VD2$), the second subgroup (middle SES) receives 40% of the available vaccines each day, while the first (low SES) and third (high SES) subgroups receive each 30%. Finally, in the third case ($VD3$), the distribution is more uneven, especially for the subgroups experiencing a low SES, with the three subgroups receiving 25%, 35%, and 40% of the vaccines, respectively.

4.3.1 Scenario 1: concurrent vaccinations and epidemics

In Figure 4.1a we begin by observing the epidemic wave under different vaccination distributions (i.e., $VD1$, $VD2$, $VD3$) as predicted by the generalized model (see coloured solid lines) and an age-stratified model (see black dashed line). The first observation is that, as expected from what was discussed in Chapter 3, the generalised model's projections significantly differ from those of the age-stratified model. Indeed, the latter overestimates the total number of infected individuals. The discrepancy is due to the distribution of activity, assortativity and population among subgroups. Second, the total number of infected projected by the generalized model varies across the vaccination uptake distributions. Notably, vaccination distribution of $VD3$ (in orange) proves to be the most effective, as it reduces the number of infections the most. This effectiveness is primarily due to the allocation of doses, where 75% of the total daily available vaccines are distributed to the two most socially active groups. Specifically, the third group, which, despite being the smallest by population size, exhibits the highest assortativity and activity (alongside the second group), receives 40% of doses. The combination of being both the smallest and the most active means that, in $VD3$, the third group becomes entirely vaccinated (up to the saturation value). At the same time the second group, also highly active, receives 35% of doses. Prioritizing these two active groups maximizes the impact of the vaccination campaign. On the contrary, a different outcome is produced by $VD2$ (in green). In this case, 40% of the vaccines are allocated to the second group, which is both the most active (along with group 3) and the largest in population size. However, the vaccination share allocated to the second group is insufficient to effectively hamper infections. Simultaneously, the lower vaccination rate in the third group, combined with its high activity level, allows for increased transmission rates, which affect the overall dynamics of the epidemic. Similar results are observed for $VD1$ (in blue). Although doses are distributed evenly, the unequal population size and activity levels result in a

performance comparable to $VD2$.

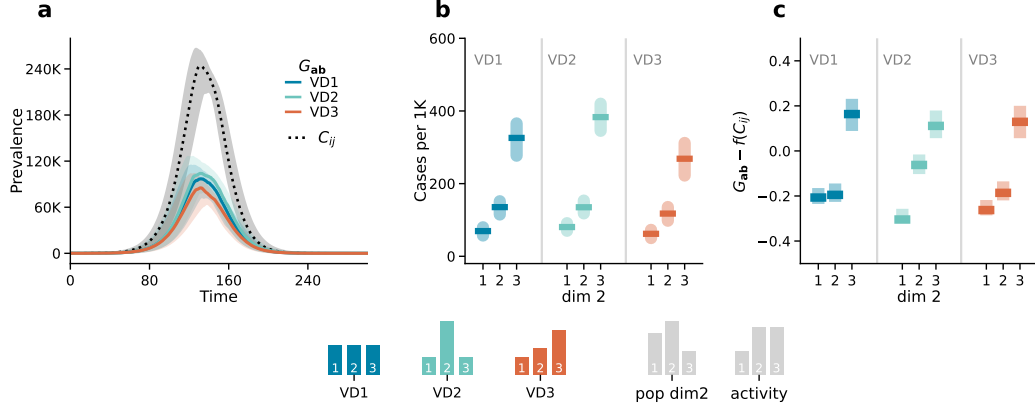


Figure 4.1: Scenario 1: epidemic outcomes. Panel *a* displays the number of infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcomes of the generalized model (G_{ab}) in the three different vaccination distribution strategies (i.e., $VD1, VD2, VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rates predicted by the generalized model G_{ab} and those estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25, \Psi = 0.4, g_1 = 0.6, R_0 = 1.6$. Simulations start with $I_0 = 100$ initial infectious seeds.

In panel *b* of Fig. 4.1, we examine the attack rates predicted by the generalised model by subgroup for each vaccination distribution. Interestingly, the third group consistently appears as the most affected, while the first group remains the least infected across all vaccination distributions. This is solely due to the differences in activity levels: as the third group is the most active, it is always the most infected relative to its population size. Notably, in $VD2$, the disparity in infections for the third group is even more pronounced, as it receives significantly fewer vaccines compared to the other groups, further amplifying its vulnerability.

In panel *c* of Fig. 4.1 we show the difference between the number of cases in each subgroup projected by the generalized model G_{ab} and estimated from an age-stratified model $f(C_{ij})$. The differences between the two are consistently away from zero. Indeed, the epidemic burden within each subgroup results from the interplay among activity levels, contact patterns, population distribution, and vaccination coverage. This cannot be captured by traditional models that do not account, explicitly, for these factors. As shown in Appendix D.4, we further explore these results by running the same analysis under different vaccination distribution settings and a homogeneous mixing case. Overall, we observe the same qualitative findings.

The interplay between vaccinations and NPIs

We now investigate how the interplay between heterogeneities in adherence to NPIs and vaccination distribution among subgroups affects epidemic outcomes. We explore hypothetical cases where a population modifies its behaviour in response to the introduction of NPIs during an ongoing vaccination campaign. We assume that, to reduce contacts, NPIs are introduced 65 days after the onset of the epidemic leading to a 20% overall reduction in contact rates, though with variations across subgroups. Indeed, we assume

that the ability to adhere to the NPIs is associated with membership in a particular population group. Specifically, we explore a case where people in the first group (i.e., experiencing the lowest SES) cannot afford to protect themselves as the other two groups. We assume that the NPIs introduce changes in both assortativity and activity. Specifically, due to the NPIs, 50%, 60%, and 70% of the contacts for the first, second, and third groups, respectively, take place within their group. Thus, assortativity increases across all groups, with the second and third groups experiencing the most significant rise. Additionally, NPIs shift activity levels to 40%, 25%, and 35% for the first, second, and third groups, respectively. This results in a decrease in activity for the second and third groups, while the first group experiences a relative increase, further exacerbating its vulnerability. Following the same logic as before, we study the effects of three different vaccination uptake distributions (i.e., $VD1, VD2, VD3$).

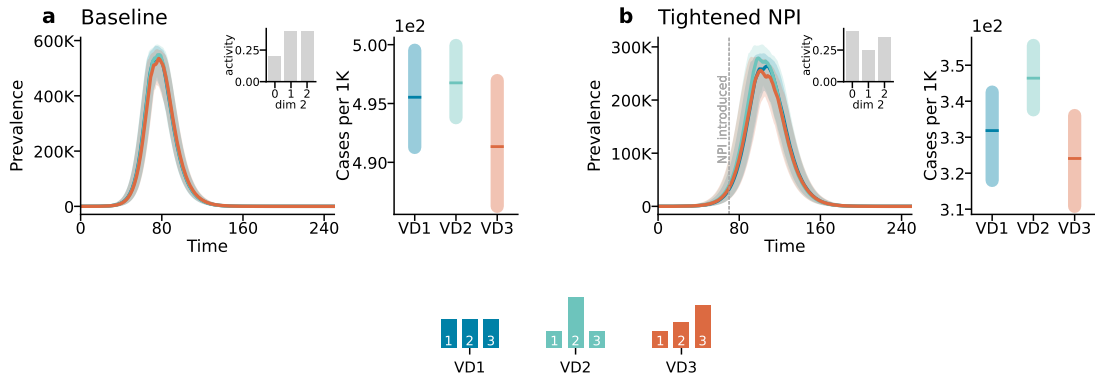


Figure 4.2: Scenario 1: The impact of NPIs under different vaccination distributions. Panel *a* refers to the baseline case, with insets depicting the activity distribution along the second dimension of the generalised contact matrix. The first column shows the prevalence, while the second column the attack rates per 1000. Panel *b* follows the same structure but refers to the case where NPIs are introduced, reducing contacts by 20%. The epidemic starts at $t_{epi} = 0$ with the onset of vaccination. NPIs are introduced 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

In Fig. 4.2*a*, we report the results in a baseline case where the disease spreads unmitigated by any NPIs. The first column of panel Fig. 4.2*a* displays the prevalence over time, while the second column shows the corresponding attack rates per 1000. The differences between all vaccination strategies are small, but $VD3$ is still the most effective, leading to a higher reduction. In contrast, $VD1$ and $VD2$ are less effective, with $VD2$ being the least effective.

However, as shown in Fig. 4.2*b*, when NPIs are introduced, the attack rate of the different SES changes (See Appendix D.5 for the corresponding plot), thus influencing the relative effectiveness of the vaccination strategies. By ensuring better coverage for the first group, which is now more vulnerable due to its higher activity, strategy $VD1$ results in a more balanced reduction of cases across all subgroups. Meanwhile, $VD2$ becomes even less effective, as the majority of vaccines are assigned to the second group, the less active, while the more active groups receive fewer vaccines. Additionally, $VD3$ loses its relative efficacy due to the reduced activity of group three, which receives the highest share of vaccines, and the increased activity in group one, which is the least vaccinated. For completeness, in the Appendix, we investigate the effect of mortality by modelling the daily number of deaths. In Appendix Section D.5, we report the

corresponding results, which show the same trend.

4.3.2 Scenario 2: vaccinations completed before epidemics

In this scenario, we explore the epidemic dynamics assuming the vaccination rollout is already completed before the onset of the epidemic. This represents situations akin to post-vaccination waves, where the majority of the population has received vaccines, but the virus continues to circulate, possibly due to new variants or waning immunity.

Following the structure of the previous scenario, Fig. 4.3a illustrates the number of newly infected individuals over time for the three vaccination distributions, as projected by the generalized model (see coloured solid lines) and the traditional age-stratified model (see black dotted line). Also in this scenario, the generalized model projects different epidemic outcomes compared to the age-stratified model. Additionally, the differences between epidemic outcomes across vaccination distributions are more pronounced. This is because the epidemic starts only after the vaccination distributions have been fully implemented, maximizing the impact of the different vaccination strategies. Notably, *VD3* remains the most effective strategy, resulting in a significantly lower number of infections. The success of *VD3* stems from the priority given to the most active groups, ensuring that the most socially active members of the population are well-protected. Conversely, *VD2*, which prioritizes the middle SES group, leads to a higher overall number of cases, as the third, more active SES group remains under-vaccinated and plays a critical role in sustaining transmission.

Fig. 4.3b breaks down the attack rates by SES, showing that even in this scenario, the most active group (SES 3) experiences the highest number of infections relative to their population size, particularly in *VD2*. This confirms that, despite overall high vaccination coverage, specific subgroups can still disproportionately contribute to the spread of the virus.

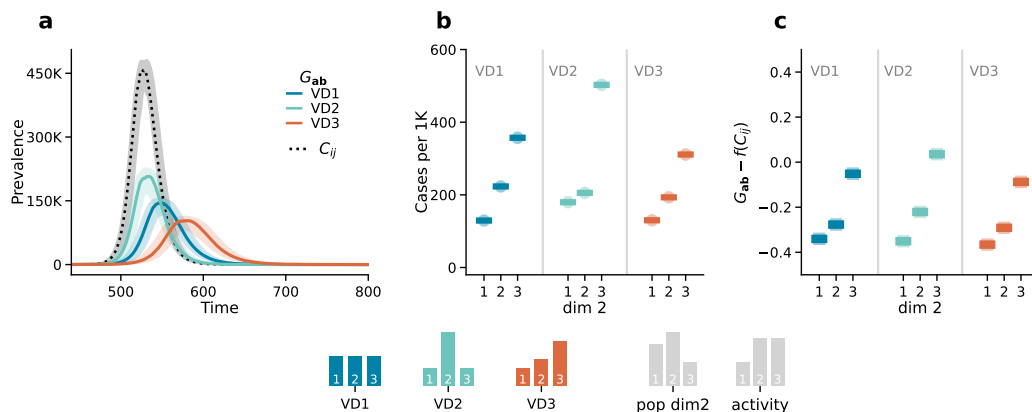


Figure 4.3: Scenario 2. Epidemic outcomes. Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcomes of the generalized model (G_{ab}) in the three different vaccination distributions strategies (i.e., *VD1*, *VD2*, *VD3*). Panel *b* shows the attack rates per 1000 by the second dimension (dim 2) projected by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rates projected by the generalized model G_{ab} and those estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, $R_0 = 3$. Simulations start with $I_0 = 100$ initial infectious seeds.

In panel 4.2c, we compare the predictions of the generalized model (G_{ab}) with those derived from the age-stratified model (C_{ij}). Once again, we observe significant discrepancies between the methods. Overall, these results highlight the importance of explicitly incorporating subgroup-specific dimensions, such as socioeconomic status, into epidemic models.

In Appendix D.4, we further explore these results by conducting a robustness analysis running the same analysis under different vaccination distribution settings. In particular, we consider a case in which age is ignored and another that assumes homogeneous mixing between groups. Notably, in the latter, we find that the difference between the projected burden of the epidemic from aggregate model output ($f(C_{ij})$) and those gathered from the generalized model (G_{ab}) is consistent 0. This implies that in case of homogeneous mixing and when the overlap between vaccination and epidemics is zero the differences across sub-groups can be obtained also a posteriori from models that neglected them.

The interplay between vaccinations and NPIs

In this scenario, the population is already fully vaccinated before the start of the epidemics. We assume that, before the start of the vaccination, some NPIs were in place. Hence we consider a situation where NPIs are then relaxed, leading to a 20% overall increase in contact rates. As before, we assume that the ability to adhere to NPIs varies across subgroups, with the variations remaining consistent with those depicted in Fig. 4.2. The NPIs are introduced for 65 days after the onset of the epidemic.

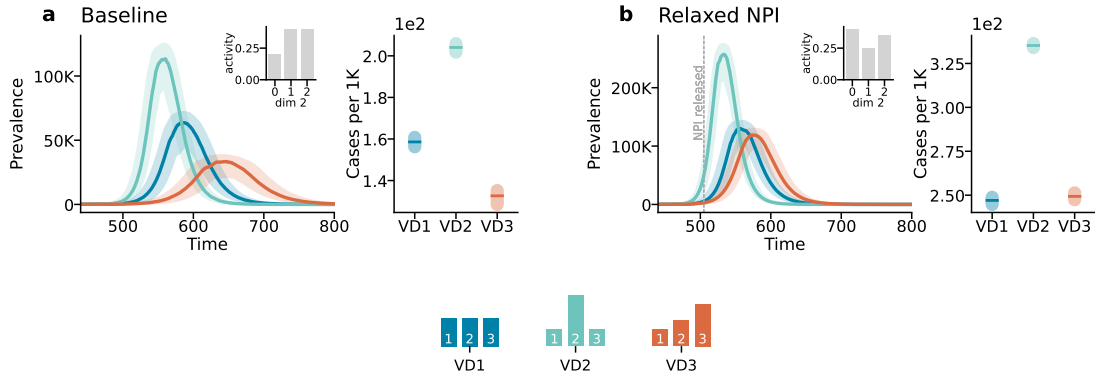


Figure 4.4: Scenario 2: The impact of NPIs under different vaccination distributions.

Panel *a* refers to the baseline case, with insets depicting the activity distribution along the second dimension. The first column shows the prevalence as a function of time, while the second column shows the attack rate scaled by 1000. Panel *b* follows the same structure but refers to the case where NPIs are relaxed, increasing contacts by 20%. NPIs are adjusted 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

In the baseline case (see Fig. 4.4a), although the different impact of vaccination distributions is more pronounced compared to the previous case, the relative effectiveness of the vaccination strategies follows the order $VD2$, $VD1$, and $VD3$, with $VD2$ being the least effective. However, when NPIs are relaxed (see Fig. 4.4b), the epidemic burden in the different SES changes according to the new activity distribution, and in turn affects the relative effectiveness of the strategies. In this case, strategy $VD1$ becomes more effective than $VD3$, while $VD2$ remains the least effective vaccination strategy.

These simulations highlight how the interplay between NPIs and vaccination strategies influences epidemic outcomes, with effectiveness depending on the timing and intensity of interventions. Again, in Appendix D.5, we investigate the effect of mortality which shows the same trends.

4.4 Hungarian contact data

We applied the model to empirical data describing social contacts stratified by age and one SES variable (i.e., self-perceived wealth with respect to the average) in Hungary during the COVID-19 Pandemic using data from the MASZK study (See Appendix B.1). To build these matrices we used two contact diaries collected in April and November 2021, respectively just before the third and the fourth wave of the COVID-19 Pandemic in Hungary.

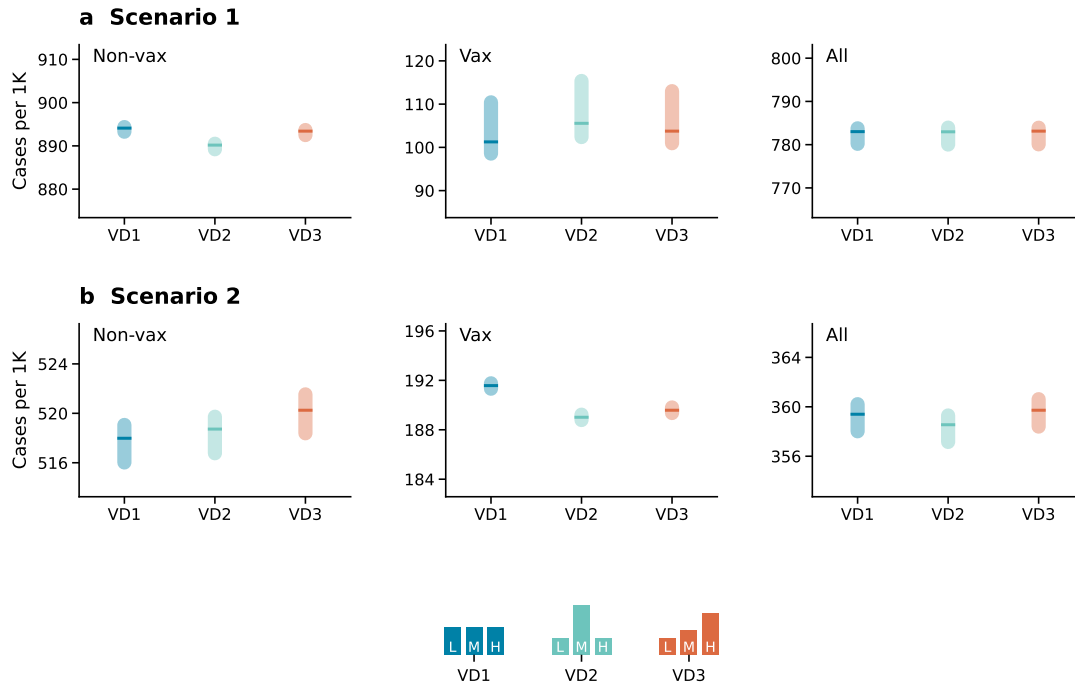


Figure 4.5: Epidemic outcomes under different vaccination distributions and real contact data. Panels in (a) refer to Scenario 1 and display the number of cases per 1000 stratified by vaccination status (non-vaccinated, vaccinated, and overall) and under different vaccination distributions ($VD1, VD2, VD3$). Panels b show the corresponding numbers for Scenario 2. Results refer to the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, $g_2 = 0.8$ and $R_0 = 3$. Simulations start with $I_0 = 100$ initial infectious seeds.

In Fig. 4.5, we present the outcomes for the Hungarian case, respectively for Scenario 1 (panel a) and Scenario 2 (panel b). In each panel, we show the number of cases per 1000 individuals stratified by the non-vaccinated ($Non-vax$) and vaccinated (Vax) subpopulations and overall (All). To facilitate straightforward visual comparison between outcomes, the y -axis of the three subplots in each panel is set to span the same range. In other words, the difference between the maximum and minimum values on the y -axis is fixed across all subplots within the same panel.

In Scenario 1, the population was initialized with 20% of individuals being immune,

representing those who had residual immunity. Vaccination was assumed to have started simultaneously with the onset of the epidemic, setting the day of the epidemic as time $t_{\text{epi}} = 0$. For Scenario 2, the population was initialized with 40% of individuals being immune, reflecting a higher level of residual immunity. In this case, the epidemic was set to begin at $t_{\text{epi}} = 220$ when 55% of the population is vaccinated, which implies 220 days had passed since the start of the vaccination campaign before the epidemic started. This adjustment mirrors the same time gap observed between the start of the COVID-19 vaccination campaign and the onset of the fourth epidemic wave in Hungary.

The scenarios, modelled under different vaccination distributions (i.e., $VD1$, $VD2$, $VD3$), reveal distinct outcomes. In Scenario 1, the differences across vaccination distributions are small, as shown in Fig. 4.5a. This result can be explained by the fact that vaccination overlapped with a fairly fast epidemic. As a result, the effects of varying vaccination uptake are less visible, leading to similar outcomes.

In contrast, Scenario 2 (see Fig. 4.5b) shows more pronounced differences among the vaccination strategies. Here, $VD2$ appears to be the most effective in reducing the epidemic burden. Interestingly, as shown in Appendix D.6, we find significant differences also in the mortality of the non-vaccinated population, indicating that even non-vaccinated individuals benefit indirectly from effective vaccination strategies.

4.5 Discussion

The COVID-19 Pandemic has provided a stark example of how disparities in vaccine accessibility and non-pharmaceutical intervention compliance can shape the trajectory of an epidemic. Building on the mathematical framework of generalized contact matrices (Chapter 3), here we analyzed the impact of these inequities. As the timing of vaccinations relative to the epidemic is a factor, we explored two scenarios. In the first, the epidemic and vaccinations begin simultaneously. In the second, the epidemic starts in an almost fully vaccinated population. Particularly, we analyze the impact of different vaccination strategies, where subgroups of the population exhibit varying levels of vaccination uptake.

Our results highlight the importance of using generalized models, that explicitly account for socioeconomic inequities, to accurately estimate the epidemic burden within different subgroups of the population. Indeed, we compared the attack rates predicted by generalized models in each subgroup of the population (SES) with those inferred a posteriori by using the outputs of age-stratified models. Interestingly, we found consistent differences between the two. Hence, disparities in vaccination uptake and contact patterns among subgroups make it challenging to predict outcomes in subgroups of the population from models that do not consider these stratifications explicitly. To support these findings further, we conducted a sensitivity analysis by varying values of R_0 and the timing of the epidemic's onset (t_{epi}). Adjusting these two parameters modulates the overlap between the epidemic wave and the vaccination campaign, allowing us to explore situations in between the two scenarios. The analysis showed that, as we fix R_0 and increase t_{epi} —thus reducing the overlap between the epidemic wave and the vaccination efforts—the difference between the two methods converges to a stable value, typically different than zero. A similar trend is observed when we set t_{epi} to zero and increase R_0 . In this configuration, a more intense simulated epidemic results in less overlap between the epidemic wave and the pandemic, as the epidemic reaches extinction more rapidly during the early stages of the vaccination campaign. (For further details, see D.4.2).

Additionally, we showed how NPIs, introduced during an ongoing epidemic, can

shift the relative effectiveness of vaccination strategies, with certain distribution of doses becoming more or less effective depending on how contact patterns change. Finally, using real generalized contact matrices from Hungary, we examined, in more realistic settings, how different vaccination distributions could influence the trajectory of an epidemic. Also the results obtained considering real contact patterns underscore the importance of incorporating socioeconomic factors into epidemic models.

It is important to acknowledge the limitations of our work. Since the methods in this chapter build extensively on those employed in Chapter 3, the limitations discussed previously also apply here. Specifically, these limitations pertain to the simplified model used to generate synthetic contact matrices and the basic approach to simulating non-pharmaceutical interventions (NPIs). Moreover, our analysis of real-world data relies solely on data from Hungary, which may not fully represent the diversity of contact patterns and vaccination behaviours present in other populations. Nevertheless, our objective was not to predict precise outcomes but to illustrate the dynamic interactions among vaccination uptake, contact patterns, and NPIs and their influence on epidemic progression.

Overall, our study shows that incorporating socioeconomic heterogeneities into epidemic models allows for more accurate and nuanced predictions of disease burden across different population subgroups. By accounting for disparities in both contact patterns and vaccination uptake, our model provides a nuanced view of how these factors interact to influence epidemic dynamics. Our framework offers a straightforward, easy-to-implement, approach that enables fast and precise predictions of epidemic burden and vaccination coverage across different subgroups. This method could assist public health officials in tailoring vaccination campaigns and other interventions more effectively, ultimately reducing the unequal impact of epidemics on vulnerable populations.

Chapter 5

Conclusion

Several factors may determine how an infection would turn out for a given person. Some of them are coded genetically, or determined by physiological conditions, but many of them are environmental and correlate with one's socio-demographic characters. In any society, people show uneven patterns along numerous social, demographic, and economic characteristics, like age, income or employment status. These characteristics not only induce medical disparities between people (as in immunity, overall health conditions, or chronic diseases) but naturally translate to differences in adaption capacities and other behavioural patterns, allowing certain groups to be more exposed to infection. The simultaneous actions of all these factors lead to observable inequalities in terms of epidemic burden between different groups at the population level.

The COVID-19 pandemic has tragically reminded us that health emergencies do not affect populations equally and that social, economic, and cultural forces fundamentally shape the outcome of disease outbreaks, reflecting the socio-economic inequalities of the society at large [158–161]. These observations are in stark contrast with the traditional structure of epidemic models, which often neglect all variables but age as the key demographic feature defining disease transmission in a population [64].

Recent studies have highlighted the urgent need for novel modelling approaches that can account for the multiple social dimensions that define the risk of infection and infection outcomes [12, 13, 15]. In the wake of the COVID-19 pandemic, research studies have introduced social contact matrices stratified by alternative demographic groups, such as racial and ethnic subpopulations [162], but usually without considering more than one dimension at once. As an alternative, scholars have developed large-scale individual-based models that reconstruct synthetic populations of millions of agents with extreme realism, including many key socio-economic characteristics [163–166]. Such models, however, require the availability of highly resolved mobility data derived e.g., from mobile phone logs, and large computational infrastructures for data processing and simulations. Previous studies have compared different data representation methods, ranging from fully homogeneous mixing to temporal networks, to identify the relevant trade-offs between compactness and realism [167–170].

The works presented here aim to make a significant contribution to this scientific landscape. First, by identifying the key dimensions that play a crucial role in shaping epidemic-relevant behaviors (Chapter 2). Second, we attempt to strike a balance between traditional approaches that are too simplistic and the complexity of large-scale synthetic populations, providing a general and flexible scheme to define epidemic models with multiple interacting population sub-groups (Chapters 3 and 4).

In detail, in Chapter 1 this work begins with a review of the key findings from exist-

ing literature in both epidemiology and social sciences to provide a foundation for the subsequent research objectives. Chapter 2 focuses on the importance of social inequalities, identifying the socio-demographic and economic factors that shape contact patterns and vaccination uptake. Using data from the *MASZK* study [92, 171], we developed a statistical model that demonstrates how contact patterns and vaccination uptake are significantly influenced by socio-economic factors, such as employment status and education level, during the COVID-19 pandemic in Hungary. Additionally, this chapter introduces the possibility of extending traditional age-structured models by decoupling the contact matrices along other socio-economic dimensions of the *participant* (*ego*). Chapter 3 expands the modelling approach further, introducing a framework of *generalized age contact matrices* that stratifies contacts across multiple social dimensions, of both the *participant* (*ego*) and the *contacted* (*alter*), providing a more comprehensive understanding of how these interactions drive disease spread. Finally, Chapter 4 integrates vaccination strategies into the framework of generalized age contact matrices, offering a more nuanced model for understanding the interplay between vaccination uptake, contact patterns and epidemic outcomes across various socio-economic groups.

While the specific limitations of each study are addressed in the discussion sections of their respective chapters, there are broader limitations that apply to this work as a whole. First, much of the research focuses on Hungary. Hungary’s unique demographic and economic structure, as well as its healthcare system, may limit the generalizability of these findings to other countries with different socio-economic dynamics and public health infrastructures. Although the models and frameworks developed here offer valuable insights, further research is necessary to validate these results in diverse global contexts.

Second, much of the analysis relies on simulated, synthetic scenarios (Chapters 3 and 4), which are not intended to replicate real-world data exactly. Instead, the simulations serve as a general test bed to explore theoretical concepts and investigate potential epidemic outcomes based on socio-economic characteristics. This choice was largely due to the absence of detailed empirical data on socio-economic factors in contact patterns and vaccination uptake. While such simulated environments are valuable for understanding broad trends, they cannot fully capture the complexity of real-world epidemics. Future research should focus on leveraging empirical data to enhance the applicability of generalized models in studying the spread of infectious diseases across different population subgroups.

In this regard, the thesis underscores the urgent need for more comprehensive contact surveys. Current data collection efforts tend to focus primarily on demographic factors like age, but must be expanded to include other crucial socio-economic dimensions—such as employment status, education, and income—of both the participant (*ego*) and their contact (*alter*).

However, it is important to acknowledge important ethical and privacy concerns linked with data collection efforts that inspect individuals about several variables. Indeed, by increasing the number of dimensions the size of each surveyed group rapidly decreases. Risks of re-identification are real and of particular concern, especially for minorities and underrepresented groups [172]. Progresses in the direction we have proposed here are necessarily linked to progresses in data collection and dissemination that reconcile the development of better public health tools on the one side with privacy rights on the other [173]. Arguably, it is this unresolved tension that, among other reasons, has so far limited the collection and sharing of contact survey data including more dimensions besides age. Nevertheless, the data we use demonstrates that such data collection is

possible in an anonymous and privacy-protected way.

Despite these limitations, the insights gained from this research offer substantial implications for public health and policy. By demonstrating how socio-economic inequalities shape epidemic outcomes, this work underscores the need for targeted interventions that address these disparities. Policies that prioritize vaccination for lower-income or less-educated individuals, implement tailored non-pharmaceutical interventions (NPIs) for high-risk communities, and ensure equitable access to healthcare can help mitigate the unequal burden of disease. This research thus not only advances the academic field of epidemic modeling but also provides actionable recommendations for reducing health inequities during public health crises.

In this regard, our work introduces critical tools, specifically through the *extended SEIR model* and the *generalized model*, which incorporate socio-economic dimensions such as income, education, and employment into the framework of epidemic modeling. The flexibility of these models allows for the inclusion of various factors that differ across subgroups of the population. While our focus was on contact patterns, compliance with NPIs, and vaccination uptake, the model can be easily extended to account for differences in susceptibility to infection, or comorbidities that increase the probability of fatality for certain groups.

These models enable public health authorities to predict not only the overall spread of an epidemic but also its differential impact across various social strata. By providing a more precise understanding of how inequalities drive transmission dynamics, interventions can be more strategically designed to protect vulnerable groups, potentially reducing both morbidity and mortality rates in disadvantaged populations.

Moreover, this approach opens avenues for future research to incorporate other variables that may be particularly relevant to specific epidemics or illnesses, such as race, preexisting chronic health conditions, or gender. Such expansions will enhance the applicability of epidemic models to diverse populations, making them more effective in guiding policy decisions.

Overall, this work contributes to the literature by placing socio-economic and other dimensions at the forefront of epidemic modeling. Addressing these factors is crucial for developing more precise descriptions of epidemics and designing more effective strategies to contain them.

Appendix A

Appendix chapter 1

A.1 The next-generation matrix approach

A commonly used method to find R_0 is the *next-generation matrix approach* (NGM) [68, 69]. This method is centered on considering the dynamics near the disease-free equilibrium (DFE). In the following we provide a summary of its derivation in the case of (i) SEIR model with homogeneous mixing hypothesis, and (ii) SEIR model with age structure population. To keep equations more readable, we henceforth drop the explicit time dependence notation.

A.1.1 SEIR with homogeneous mixing

The next-generation matrix approach requires focusing on the compartments that describe any stages of the infection: in our case, E and I . It is convenient to rewrite the differential equations for these compartments as:

$$d_t \mathbf{x} = f(\mathbf{x}) - w(\mathbf{x}) \quad (\text{A.1})$$

where $\mathbf{x} \in [E, I]$, $f(\mathbf{x})$ encodes all terms that lead to new infections and $w(\mathbf{x})$ all other transitions in and out of the compartments. More explicitly, we can write the following.

$$d_t \mathbf{x} = \begin{bmatrix} d_t E \\ d_t I \end{bmatrix} = f(\mathbf{x}) - w(\mathbf{x}) = \begin{bmatrix} \lambda S \\ 0 \end{bmatrix} - \begin{bmatrix} \epsilon E \\ \gamma I - \epsilon E \end{bmatrix} \quad (\text{A.2})$$

The expression of R_0 is linked to the early epidemic dynamics which can be linearized by calculating the Jacobian at the disease free equilibrium (DFE) $(S^*, E^*, I^*, R^*) = (N, 0, 0, 0)$. Using the Jacobian matrix $\mathbf{J}$ (which has size 2×2) at the DFE we can write

$$d_t \mathbf{x} = \mathbf{J} \mathbf{x} \quad (\text{A.3})$$

This expression can be conveniently factorized as $d_t \mathbf{x} = (\mathbf{F} - \mathbf{W}) \mathbf{x}$ where the matrix $\mathbf{F}$ is Jacobian applied to $f(\mathbf{x})$ and similarly the matrix $\mathbf{W}$ is the Jacobian applied to $w(\mathbf{x})$. In particular, we have:

$$\mathbf{F} = \begin{bmatrix} \partial_E(\lambda N) & \partial_I(\lambda N) \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} \quad (\text{A.4})$$

$$\mathbf{W} = \begin{bmatrix} \partial_E(\epsilon E) & \partial_I(\epsilon E) \\ \partial_E(\gamma I - \epsilon E) & \partial_I(\gamma I - \epsilon E) \end{bmatrix} = \begin{bmatrix} \epsilon & 0 \\ -\epsilon & \gamma \end{bmatrix} \quad (\text{A.5})$$

The expression for R_0 is linked to the two matrices as $R_0 = \rho(\mathbf{F}\mathbf{W}^{-1})$, where ρ denotes the spectral radius. It is easy to show how

$$\mathbf{F}\mathbf{W}^{-1} = \frac{1}{\epsilon\gamma} \begin{bmatrix} \phi\epsilon & \phi\epsilon \\ 0 & 0 \end{bmatrix} \quad (\text{A.6})$$

hence we can finally write:

$$R_0 = \frac{\phi\epsilon}{\epsilon\gamma} = \frac{\phi}{\gamma} \quad (\text{A.7})$$

A.1.2 SEIR with structured population

As before, the first step is to focus our attention only on the compartments that describe any stages of the infection: E_i and I_i $i \in [1, K]$ in our case. The differential equations for these compartments can be rewritten according to (A.1) as follows:

$$d_t \mathbf{x} = \begin{bmatrix} d_t E_1 \\ \vdots \\ d_t E_K \\ d_t I_1 \\ \vdots \\ d_t I_K \end{bmatrix} = f(\mathbf{x}) - w(\mathbf{x}) = \begin{bmatrix} \lambda_1 S_1 \\ \vdots \\ \lambda_K S_K \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \epsilon E_1 \\ \vdots \\ \epsilon E_K \\ \gamma I_1 - \epsilon E_1 \\ \vdots \\ \gamma I_K - \epsilon E_K \end{bmatrix} \quad (\text{A.8})$$

As above, we linearize the early epidemic dynamics by calculating the Jacobian at the disease free equilibrium (DFE) $(S_i^*, E_i^*, I_i^*, R_i^*) = (N_i, 0, 0, 0)$ this time, for all age groups. Using the Jacobian matrix $\mathbf{J}$ (which has size $2K \times 2K$) at the DFE we can write the expression for the $\mathbf{F}$ matrix

$$\begin{aligned} \mathbf{F} &= \begin{bmatrix} \partial_{E_1}(\lambda_1 N_1) & \dots & \partial_{E_K}(\lambda_1 N_1) & \partial_{I_1}(\lambda_1 N_1) & \dots & \partial_{I_K}(\lambda_1 N_1) \\ \partial_{E_1}(\lambda_2 N_2) & \dots & \partial_{E_K}(\lambda_2 N_2) & \partial_{I_1}(\lambda_2 N_2) & \dots & \partial_{I_K}(\lambda_2 N_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \partial_{E_1}(\lambda_K N_K) & \dots & \partial_{E_K}(\lambda_K N_K) & \partial_{I_1}(\lambda_K N_K) & \dots & \partial_{I_K}(\lambda_K N_K) \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \dots & 0 & \partial_{I_1}(\lambda_1 N_1) & \dots & \partial_{I_K}(\lambda_1 N_1) \\ 0 & \dots & 0 & \partial_{I_1}(\lambda_2 N_2) & \dots & \partial_{I_K}(\lambda_2 N_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \partial_{I_1}(\lambda_K N_K) & \dots & \partial_{I_K}(\lambda_K N_K) \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \end{aligned} \quad (\text{A.9})$$

It is important to recall that the components of each gradient, are computed at the DFE, i.e., $S_i \rightarrow N_i$. By looking at Eq. 1.5 we note how each partial derivative in p

selects the p -th element of the sum hence:

$$\mathbf{F} = \phi \begin{bmatrix} 0 & \dots & 0 & \frac{C_{11}}{N_1} N_1 & \dots & \frac{C_{1K}}{N_K} N_1 \\ 0 & \dots & 0 & \frac{C_{21}}{N_1} N_2 & \dots & \frac{C_{2K}}{N_K} N_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \frac{C_{K1}}{N_1} N_K & \dots & \frac{C_{KK}}{N_K} N_K \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} = \phi \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (\text{A.10})$$

where we denoted with $\mathbf{0}$ $K \times K$ blocks of zeros and the generic entry of $\tilde{\mathbf{C}}$ is $\tilde{C}_{ij} = \frac{C_{ij}}{N_j} N_i$. Using the same method we can easily write the expression of the matrix $\mathbf{W}$ as:

$$\mathbf{W} = \begin{bmatrix} \epsilon & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \epsilon & 0 & \dots & 0 \\ -\epsilon & \dots & 0 & \gamma & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & -\epsilon & 0 & \dots & \gamma \end{bmatrix} = \begin{bmatrix} \epsilon \mathbf{I} & \mathbf{0} \\ -\epsilon \mathbf{I} & \gamma \mathbf{I} \end{bmatrix} \quad (\text{A.11})$$

where $\mathbf{I}$ are identity matrices of size $K \times K$. The expression for R_0 is linked to the two matrices as $R_0 = \rho(\mathbf{F}\mathbf{W}^{-1})$, where ρ denotes the spectral radius. It is easy to show how

$$\mathbf{F}\mathbf{W}^{-1} = \phi \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \frac{1}{\epsilon} \mathbf{I} & \mathbf{0} \\ \frac{1}{\gamma} \mathbf{I} & \frac{1}{\gamma} \mathbf{I} \end{bmatrix} = \frac{\phi}{\gamma} \begin{bmatrix} \tilde{\mathbf{C}} & \tilde{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (\text{A.12})$$

hence, we finally can write

$$R_0 = \frac{\phi}{\gamma} \rho(\tilde{\mathbf{C}}) \quad (\text{A.13})$$

In the case in which contact matrices are stratified for the context (1.2) as the ω_l are assumed to be homogeneous across age-brackets the inclusion of the different contexts does not change the expression of R_0 .

Appendix B

Appendix chapter 2

B.1 MASZK dataset

B.1.1 Socio-demographic dimensions

The MASZK dataset provides us with an extensive set of information on *social-demographic characteristics* of the participants. In this section, we provide a detailed explanation of all the variables used in this work.

- *Education level* which can have four possible levels: low, mid-low, mid-high and high. We considered low-educated individuals with any primary school degrees, mid-low-educated those holding certificates from professional and vocational schools, mid-high-educated those with secondary school, and highly educated with a university education or above (eg. BSc, MSc, PhD).
- *Employment situation* which can be either employed or not-employed. In the not-employed category, we include students and retired individuals.
- *Income level* can have three possible levels: low, mid and high. In particular, individuals were asked to report their *perceived income* with respect to the average using a scale from 1 to 10. We consider low-income individuals who answered from 1 to 4, mid-income those who answered 5 or 6 and high-income those who answered 7 or above. Although this is actually *perceived income*, for conciseness we refer to it in the text as income. Please note that the results concerning this dimension are missing for the 2nd W. This is due to the fact that this variable has been collected only from the 9th data collection onwards.
- *Settlement* refers to the area where individuals live and can be either capital that is Budapest and its districts, rural which includes larger and smaller sizes villages, or urban which includes cities and towns except the capital.
- *Gender* refers to the biological gender and can be either female or male.
- *Chronic disease* is a Boolean dimension indicating if an individual is affected by any chronic disease or not.
- *Acute disease* is a Boolean dimension indicating if an individual is affected by any acute disease or not.

- *Smoking* is a Boolean dimension indicating if an individual is a smoker or not. We consider as *non-smoker* individuals who declared themselves as not-smokers or who stopped, while we consider *smoker* individuals who smoke frequently or occasionally.

B.1.2 Sample sizes

We report here the sample size by age and the variable of interest. That small variation in the total numbers is due to some missing answers from the participants. In addition, when aggregating by income level we can only show the results starting from the *3rdW*. This is due to the fact that this information has been collected only from the *9th* data collection. Finally, the overall amounts of individuals in the table by smoking behaviour amount to the adult population.

Education Level

Age Group	Education Level	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	low	637	710	376	520	503	604
[5 – 15)	low	1368	1497	849	1086	1113	1314
[15 – 30)	low	483	490	317	422	422	472
	mid-low	122	140	84	121	107	109
	mid-high	547	802	548	514	519	656
	high	358	199	129	90	95	163
[30 – 45)	low	181	74	49	66	43	64
	mid-low	381	261	135	153	147	150
	mid-high	504	508	403	377	358	435
	high	485	670	412	383	471	558
[45 – 60)	low	476	288	200	161	138	235
	mid-low	359	440	250	295	262	405
	mid-high	424	461	331	303	338	416
	high	267	332	258	283	270	281
[60 – 70)	low	413	689	353	297	314	355
	mid-low	302	333	244	196	211	280
	mid-high	316	221	109	126	90	94
	high	34	37	18	48	12	27
[70+)	low	341	269	277	330	346	485
	mid-low	216	189	153	142	184	235
	mid-high	240	148	89	100	114	105
	high	40	27	25	38	5	13
Total		8494	8785	5609	6051	6062	7456

Table B.1: Number of survey participants in the different aggregated periods by age group and the education level

Employment situation

Age Group	Employment situation	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	employed	637	710	376	520	503	604
[5 – 15)	employed	1368	1497	849	1086	1113	1314
[15 – 30)	employed	1206	1399	878	947	961	1156
	not employed	304	232	200	200	182	244
[30 – 45)	employed	1127	1241	823	797	863	991
	not employed	424	272	176	182	156	216
[45 – 60)	employed	993	1135	792	808	803	1068
	not employed	533	386	247	234	205	269
[60 – 70)	employed	225	264	176	174	144	178
	not employed	840	1016	548	493	483	578
[70+)	employed	33	30	37	31	28	50
	not employed	804	603	507	579	621	788
Total		8494	8785	5609	6051	6062	7456

Table B.2: Number of survey participants in the different aggregated periods by age group and the education level

Income Level

Age Group	Income Level	3rd W	2nd IP	4th W	5th W
[0 – 5)	high	87	148	173	190
	mid	176	218	219	269
	low	113	154	111	145
[5 – 15)	high	245	275	326	367
	mid	375	515	488	579
	low	229	296	299	368
[15 – 30)	high	291	380	346	432
	mid	562	527	558	674
	low	225	240	239	294
[30 – 45)	high	297	254	308	372
	mid	453	438	462	551
	low	249	287	249	284
[45 – 60)	high	203	224	238	293
	mid	475	498	442	600
	low	361	320	328	444
[60 – 70)	high	82	111	126	152
	mid	302	275	277	325
	low	340	281	224	279
[70+)	high	69	93	132	156
	mid	240	279	273	338
	low	235	238	244	344
Total		5609	6051	6062	7456

Table B.3: Number of survey participants in the different aggregated periods by age group and the income level.

Settlement							
Age Group	Settlement	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	capital	53	107	57	63	89	109
	rural	226	219	110	194	172	178
	urban	357	384	209	263	242	317
[5 – 15)	capital	137	292	172	144	187	258
	rural	489	474	249	363	359	408
	urban	742	731	428	579	567	647
[15 – 30)	capital	190	172	117	114	116	190
	rural	433	435	302	327	330	381
	urban	887	1023	658	706	697	829
[30 – 45)	capital	178	271	238	149	213	291
	rural	481	406	247	291	261	315
	urban	891	836	513	539	545	600
[45 – 60)	capital	292	359	247	231	192	291
	rural	461	439	285	339	281	372
	urban	773	723	506	472	535	674
[60 – 70)	capital	240	213	85	164	114	83
	rural	377	540	286	224	247	335
	urban	448	527	352	279	266	338
[70+)	capital	241	157	103	135	153	125
	rural	218	175	186	181	243	332
	urban	378	301	253	294	253	381
Total		8492	8784	5603	6051	6062	7454

Table B.4: Number of survey participants in the different aggregated periods by age group and the settlement.

Gender							
Age Group	Gender	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	Female	282	354	191	257	266	307
	Male	295	356	185	263	237	297
[5 – 15)	Female	626	741	403	561	552	657
	Male	602	756	446	525	561	657
[15 – 30)	Female	683	934	619	705	672	788
	Male	765	697	459	442	471	612
[30 – 45)	Female	720	629	449	429	453	558
	Male	831	884	550	550	566	649
[45 – 60)	Female	882	749	517	458	482	609
	Male	644	772	522	584	526	728
[60 – 70)	Female	600	811	429	367	353	434
	Male	465	469	295	300	274	322
[70+)	Female	490	347	307	388	401	524
	Male	347	286	237	222	248	314
Total		8232	8785	5609	6051	6062	7456

Table B.5: Number of survey participants in the different aggregated periods by age group and gender.

Chronic disease							
Age Group	Chronic disease	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	no	599	687	369	505	484	574
	yes	23	23	7	15	19	30
[5 – 15)	no	1228	1415	808	1025	1036	1213
	yes	90	82	41	61	77	101
[15 – 30)	no	1327	1449	975	1029	1022	1246
	yes	164	182	103	118	121	154
[30 – 45)	no	1254	1209	788	761	785	914
	yes	297	304	211	218	234	293
[45 – 60)	no	854	873	578	607	592	769
	yes	672	648	461	435	416	568
[60 – 70)	no	362	372	220	203	168	211
	yes	703	908	504	464	459	545
[70+)	no	239	170	145	121	150	181
	yes	598	463	399	489	499	657
Total		8410	8785	5609	6051	6062	7456

Table B.6: Number of survey participants in the different aggregated periods by age group and chronic disease.

Acute disease

Age Group	Acute disease	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[0 – 5)	no	618	696	371	511	475	586
	yes	4	14	5	8	28	18
[5 – 15)	no	1303	1472	836	1069	1076	1283
	yes	15	25	13	16	37	31
[15 – 30)	no	1458	1585	1051	1097	1083	1348
	yes	33	46	27	50	60	52
[30 – 45)	no	1503	1456	957	923	948	1152
	yes	48	57	42	56	71	55
[45 – 60)	no	1424	1405	969	972	931	1223
	yes	102	116	70	70	77	114
[60 – 70)	no	980	1135	652	611	571	672
	yes	85	145	72	56	56	84
[70+)	no	767	569	501	551	593	756
	yes	70	64	43	59	56	82
Total		8410	8785	5609	6049	6062	7456

Table B.7: Number of survey participants in the different aggregated periods by age group and acute disease.

Smoking

Age Group	Smoking	1st IP	2nd W	3rd W	2nd IP	4th W	5th W
[15 – 30)	no	569	549	370	353	357	419
	yes	452	505	324	349	340	444
[30 – 45)	no	965	964	628	600	620	768
	yes	586	549	371	379	399	439
[45 – 60)	no	879	913	674	646	633	864
	yes	647	608	365	396	375	473
[60 – 70)	no	711	839	458	431	404	500
	yes	354	441	266	236	223	256
[70+)	no	724	527	453	520	531	688
	yes	113	106	91	90	118	150
Total		6000	6001	4000	4000	4000	5001

Table B.8: Number of survey participants in the different aggregated periods by age group and smoking behaviour.

B.1.3 Data pre-processing

All the analyses on the number of contacts have been performed after having deleted the outliers at the 98% percentile with respect to the period of interest.

All the results presented in this work have been computed by accounting for each participant according to its weight. The weight has been provided by the survey company. In the raking procedure for the creation of the weights, in each iteration weights were calculated for one dimension to align the sample total with the population total by the given dimension (weighting variable): first for the first dimension, then for the second dimension, and so on. These iterations continued across all dimensions until the final weights made the distributions of all weighting variables the same as the population distribution (within an acceptable margin of error). [174] Weighting dimensions included gender, age, education and domicile. As a result of the procedure, each respondent was assigned with a weight, indicating whether they should have higher, or lower relative influence in the analysis. For the calculation of the weights, SPSS's SPSS_RAKE function was used [174].

B.2 Robustness checks on the statistical model

In this section, we provide all the results of the additional statistical analysis that we performed in order to support the findings presented in chapter 2.

B.2.1 Model on contacts

Accounting for the high number of zeros

All contacts Due to the implementation of NPIs (lockdowns and curfews) throughout the pandemic, people were forced, when possible, to reduce or completely reset their number of contacts. Thus, when analysing the distribution of the number of contacts we found a high presence of zeros.

To test the robustness of the results of the negative binomial regression model (*nb*) to this zero-inflated mechanism we implemented two additional models:

1. After having excluded the observations where the number of contacts was zero we re-run the *negative binomial regression model* on the non-zero number of contacts ($nb_{\text{contacts}>0}$).

$$\mu_i = \alpha + \beta_1 \text{age_group}_i + \beta_2 X_i + \beta_3 \text{age_group}_i * X_i + \epsilon_i \quad (\text{B.1})$$

2. We modelled the probability to have at least one contact using a *logistic regression model (logit)*.

$$\log P_i(\text{anycontact} = 1) = \alpha + \beta_1 \text{age_group}_i + \beta_2 X_i + \beta_3 \text{age_group}_i * X_i + \epsilon_i \quad (\text{B.2})$$

where age_group_i is the age class of i ; X_i is the variable of interest (e.g., education, income etc.), $\text{age_group}_i * X_i$ is the interaction term of the age group and the variable of interest, and ϵ_i is the error term. In model (B.1) given μ_i , we define $\lambda_i = \exp(\mu_i)$ to be the expected number of contacts for respondent i . While in model (B.2) $P_i(\text{anycontact} = 1)$ indicate the probability for respondent i to have at least one contact.

By applying the same methodology as explained in the Method section we computed the *max confidence level* by age for each of the variables and period considered in the analysis for model (B.1) and (B.2). Fig B.1 shows the results of the three models that we implemented. Interestingly, we can see that the same qualitative patterns result from both the negative binomial models, that is if we include the observations where the number of contacts is 0 (nb), or we discard it ($nb_{ncontacts>0}$) (Fig. B.1a, b). In particular, although the *maximum confidence level* computed with these models differ in terms of variability, education, employment and income seem to remain the most significant dimensions in terms of explaining differences in contact numbers among subgroups of the population. The logistic regression model (*logit*) (Fig. B.1c), shows similar results regarding the education and employment situation while it indicates that the other variables analyzed are not determining if individuals have contacts or not.

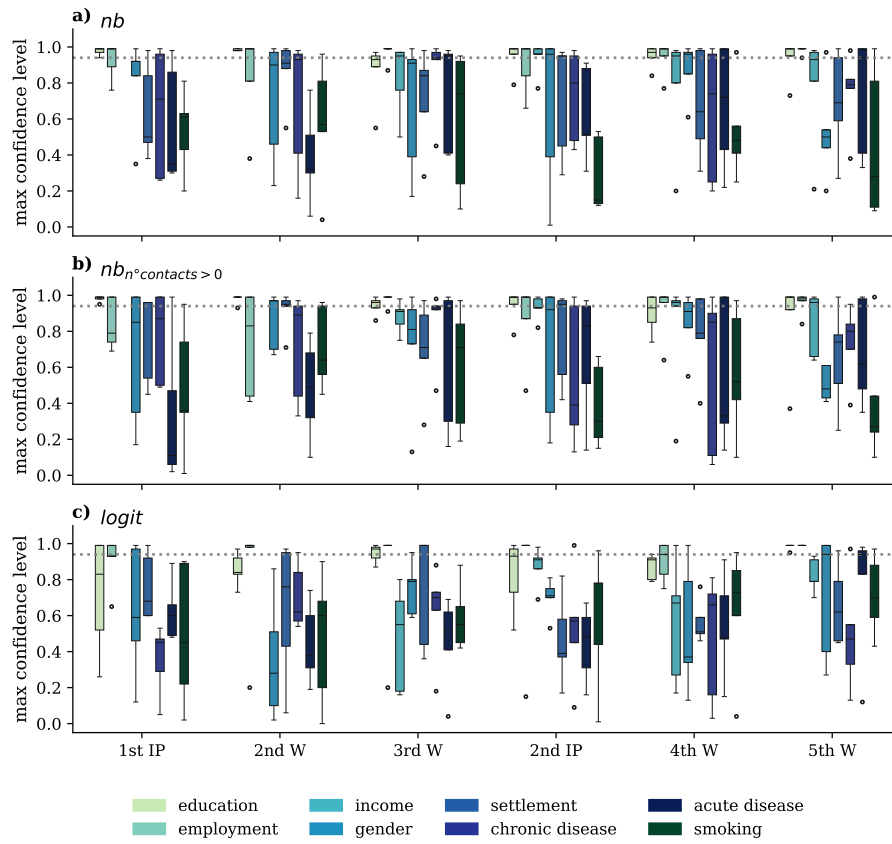


Figure B.1: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented *nb* (negative binomial) (a), $nb_{ncontacts>0}$ (negative binomial on non-zeros number of contact) (b), and *logit* (logistic regression) (c).

Community contacts Furthermore, we also investigated the contacts happening exclusively in the community layer, excluding the ones happening at work. Here we present the results of the statistical analysis on the number of contacts in the community layer. Particularly, in Fig. B.2 we show the results of the three models we are implementing: 1. negative binomial regression (*nb*), 2. negative binomial regression only on positive observations of the number of contacts ($nb_{\text{contacts}>0}$), and 3. logistic regression to model the probability of having at least one contact (*logit*). In all these models the dependent variable (y_i) refers to the community-level contacts of the individuals. Results are similar to the ones obtained considering all the types of contacts together. Indeed, also in this case results indicate that education, employment and income are the most significant dimensions in terms of explaining differences in contact numbers in the community layer among subgroups of the population. Also the logistic regression model (*logit*) (Fig. B.2c), shows similar results indicating that only education and employment are significant in determining if individuals have contacts or not in the community layer.

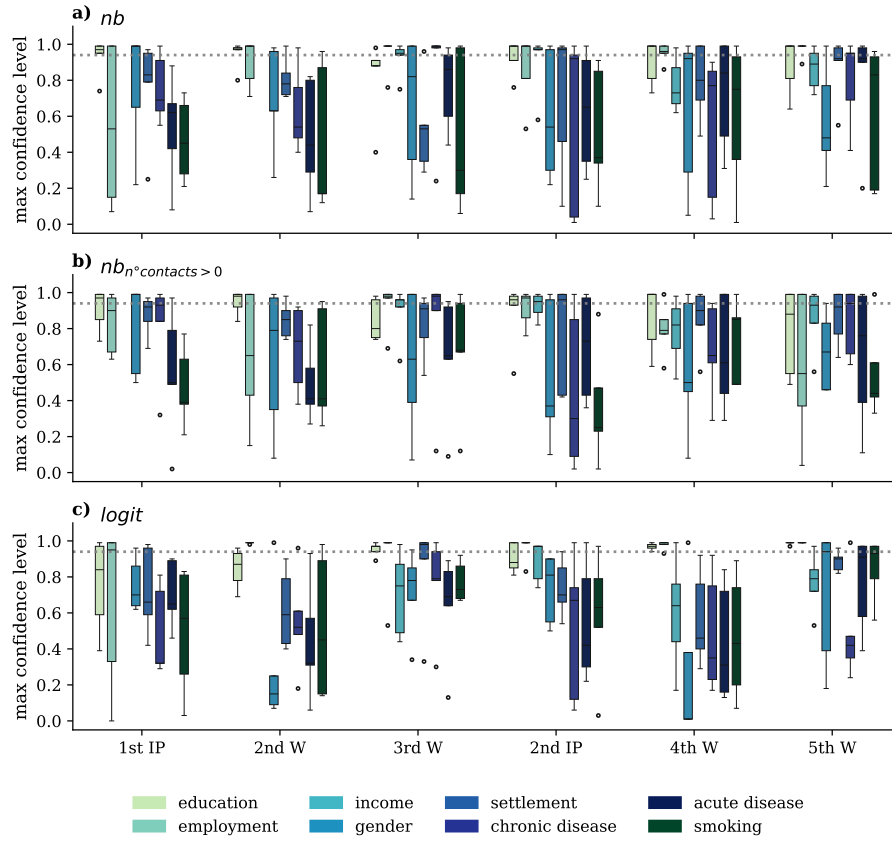


Figure B.2: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented *nb* (negative binomial) (a), $nb_{\text{contacts}>0}$ (negative binomial on non-zeros number of contact) (b), and *logit* (logistic regression) (c).

Accounting for the effect of control variables

NPIs and vaccination behaviour Although, the data were aggregated to enhance homogeneity in implemented non-pharmaceutical interventions (NPIs), yet minor variations in restrictions persist within certain periods (Fig. B.3a). In addition, we account for the interplay of vaccination behaviour and contacts which can be quite important in a dynamic situation when vaccination strategies (Fig. B.3b) are implemented over an epidemic crisis such as the one of COVID-19 [143] Thus, we have incorporated two additional controls variable to the statistical model:

- the *Oxford Stringency Index (OSI)*. This index is a composite measure that aims to quantify the strictness of the government's response to the COVID-19 pandemic. It is computed as a composite measure of nine of nine metrics: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. [175]
- a dummy variable that indicates if an individual is vaccinated (vax_i), this variable is available only from the 3rd wave on.

The results are shown in Fig. B.4 and Fig. B.5 respectively for the total number of contacts and the community ones. As we can observe the results are very similar to the one shown above, thus they bring to the same qualitative conclusions.

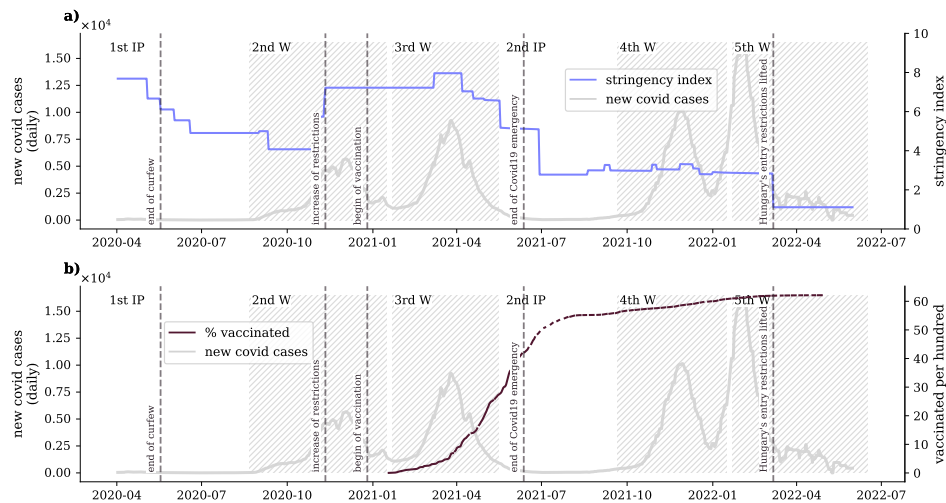


Figure B.3: **a)** *Left axis:* number of new daily COVID-19 cases in Hungary from 2020/04 to 2022/07. *Right axis:* Stringency Index. **b)** *Left axis:* number of new daily COVID-19 cases in Hungary from 2020/04 to 2022/07. *Right axis:* Percent of vaccinated people in the Hungarian population. Source: our-word-in-data. In both panels, the white and grey areas delimit the periods that have been aggregated in the analysis: two interim periods (IPs) (white areas) and four epidemic waves (W) (grey dashed areas).

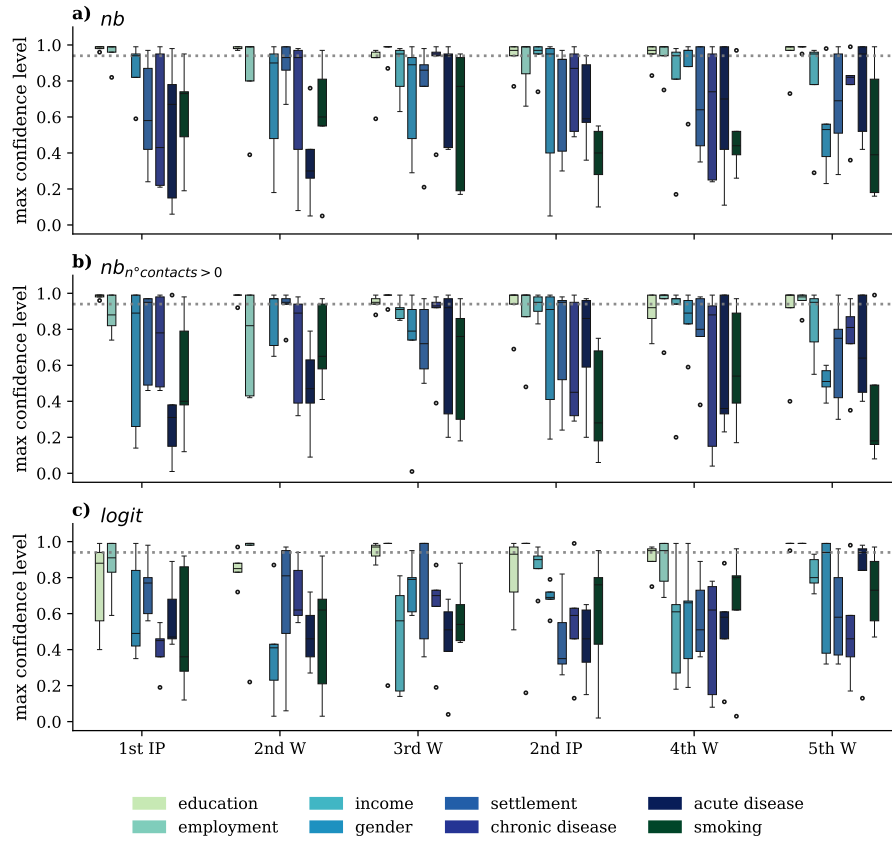


Figure B.4: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented *nb* (negative binomial) (a), *nb_{n^{*}contacts > 0}* (negative binomial on non-zeros number of contact) (b), and *logit* (logistic regression) (c).

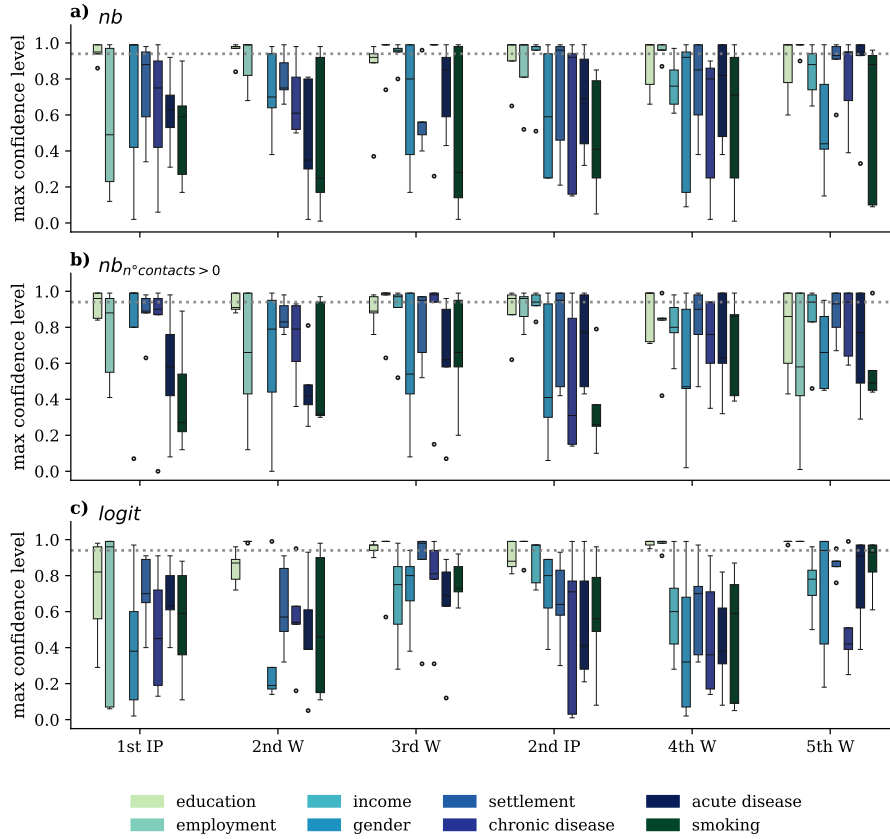


Figure B.5: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented *nb* (negative binomial) (a), *nb_{n^{*}contacts>0}* (negative binomial on non-zeros number of contact) (b), and *logit* (logistic regression) (c).

Other control variables To enhance the statistical robustness of our findings, additional analyses were conducted by incorporating supplementary control variables into the model. Specifically, for each variable of interest in the model, all the other variables were included as controls, in addition to the one presented above (OSI and vax_i). The outcomes are shown in Fig. B.6 for total contacts and Fig. B.7 for community contacts. Notably, the results closely align with those presented earlier, reinforcing the robustness of our qualitative conclusions.

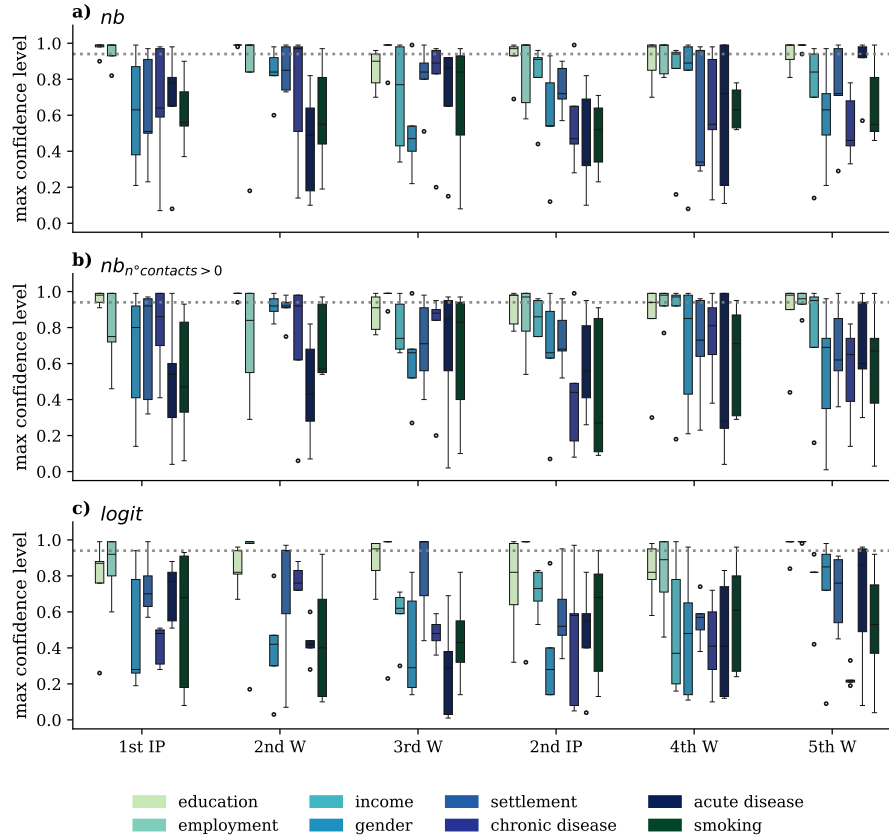


Figure B.6: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented nb (negative binomial) (a), $nb_{n^{*}contacts > 0}$ (negative binomial on non-zeros number of contact) (b), and $logit$ (logistic regression) (c).

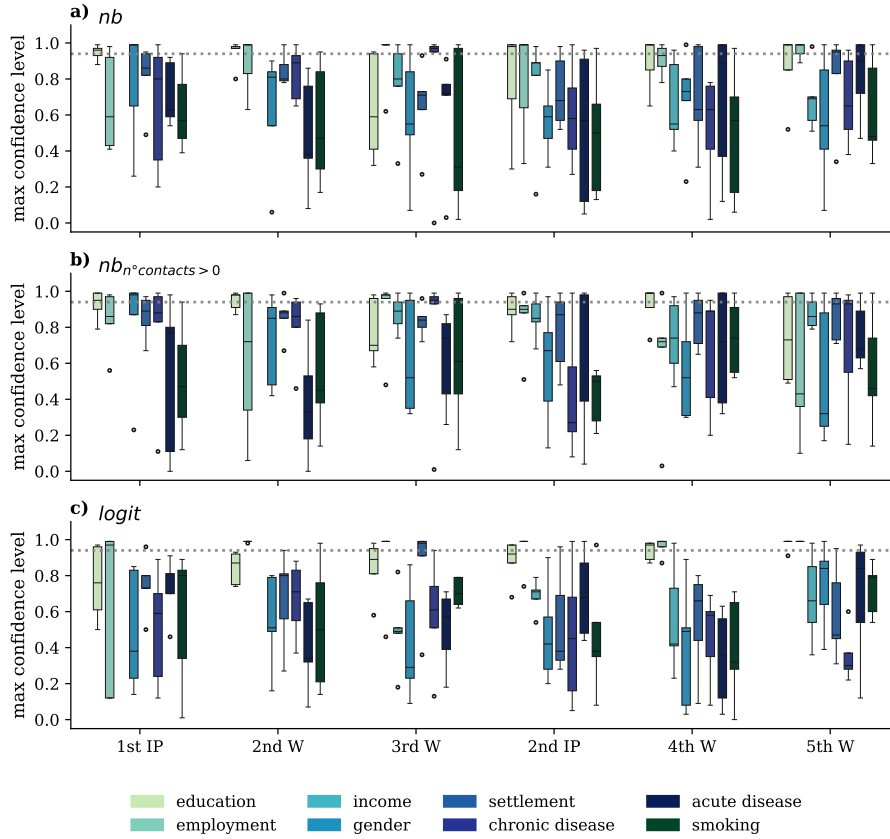


Figure B.7: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on the dependent variable of contacts becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level*, *employment situation*, *income level*, *gender*, *settlement*, *chronic disease*, *acute disease*, and *smoking behaviour*. Results for the three models implemented *nb* (negative binomial) (a), *nb_{n^{*}contacts > 0}* (negative binomial on non-zeros number of contact) (b), and *logit* (logistic regression) (c).

B.2.2 Model on vaccination

Accounting for the effect of NPIs

As we did for the model of contacts, to incorporate the impact of non-pharmaceutical interventions (NPIs) on vaccination uptake, we have incorporated an additional control variable to the statistical model: the *Oxford Stringency Index (OSI)* [175]. The results are shown in Fig. B.8. As we can observe the results are very similar to the one shown above, thus they bring to the same qualitative conclusions.

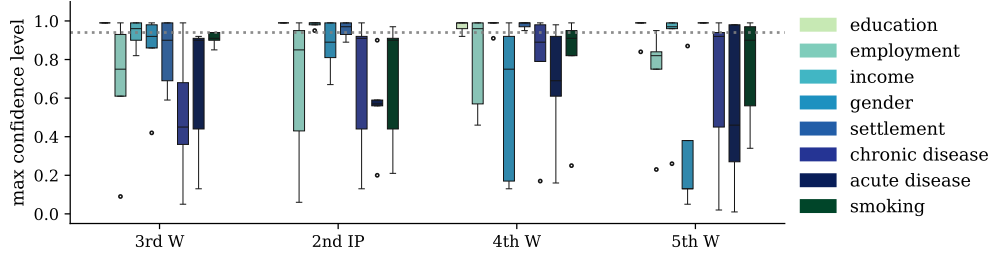


Figure B.8: Box-plot of the maximum confidence level at which the effect of the different categories of the variable on vaccination becomes significantly different. The dispersion of the box plot refers to the variation of this value over different age groups. Results are shown for *education level, employment situation, income level, gender, settlement, chronic disease, acute disease, and smoking behaviour.*

B.3 Bootstrapping

In this study, we employ bootstrapping sampling technique to assess the uncertainty of the estimation of contacts. We used bootstrapping for the estimation of average contact numbers by the different levels of the social-demographic variables, and also for the estimation of the decoupled age contact matrices. In particular, the bootstrapping method involves resampling with replacement from the original dataset, specifically focusing on a subset of the data stratified simultaneously by the given variable of interest (*employment, education level, income level and settlement*) and in case of matrices by age as well. By iteratively drawing multiple samples from each subset, we create a distribution of the average number of contacts by the levels of the variable of interest (and in the case of matrices by age as well), allowing us to quantify the uncertainty in the estimate and derive confidence intervals. All the results are conveyed as the median of 1000 bootstrapped samples, accompanied by their interquartile range (*IQR*)

B.4 Contacts

B.4.1 Average number of contacts in the community and in the work layer by sub-groups

In the main text, we show the evolution of contacts over time decoupled by education level and employment. For completeness, here we report the same figures for income level Fig B.9-a,c and settlement Fig B.9-b,d. Looking at the community contacts we can observe that there is a clear rank among the income levels in their number of contacts, with high-income individuals having the highest number of contacts and low-income individuals having the lowest. The same is arguable for the individuals living in the capital, which appear to be the most active in the community layer, while, individuals living in rural areas are the less active. To what concern the contacts at work, we can clearly observe that high-income individuals were the ones who would better adapt to the epidemiological situation, while mid and low income maintained a fairly stable number of contacts over time at the workplace. Particularly, they reported a higher number of contacts at the workplace during the COVID-19 waves. A similar conclusion can be drawn when we decoupled individuals according to their settlement. In this case, individuals living in rural areas appear to be the most active in the workplace while the ones living in the capital tend to have a lower number of interactions in the workplace.

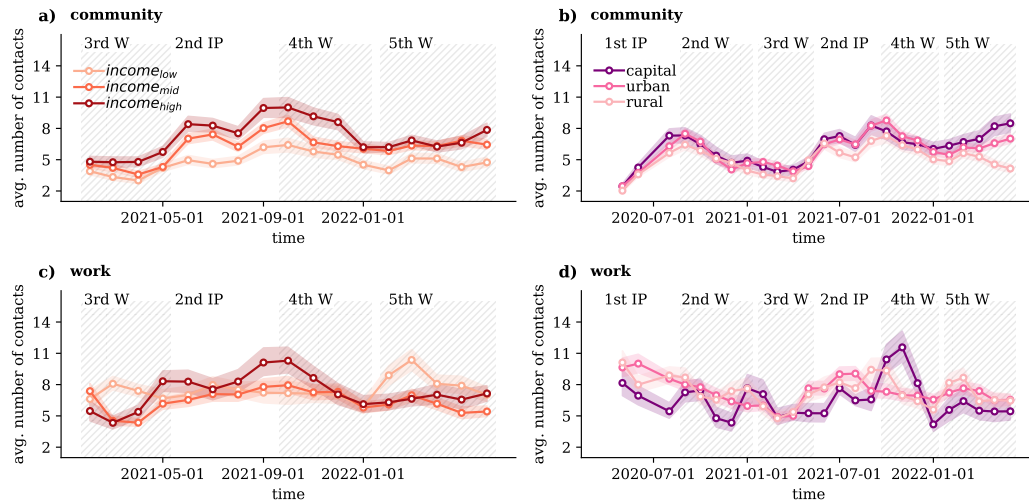


Figure B.9: (a-b) Smoothed average number of contacts in the community layer by different income levels a) and settlement types b). (c-d) Smoothed average number of contacts at the workplace by different income levels c) and settlement types d). All the values are shown as the median and the IQR of 1000 bootstrapped samples.

Relative contact variation

Here we present the relative number of contacts categorized by employment situation and education level (Fig. B.10), as well as by income level and settlement (Fig. B.11). Namely, in each month, the relative number of contacts of each subgroup is computed as the ratio between the average number of contacts for each subgroup and the average of the total number of contacts in the population.

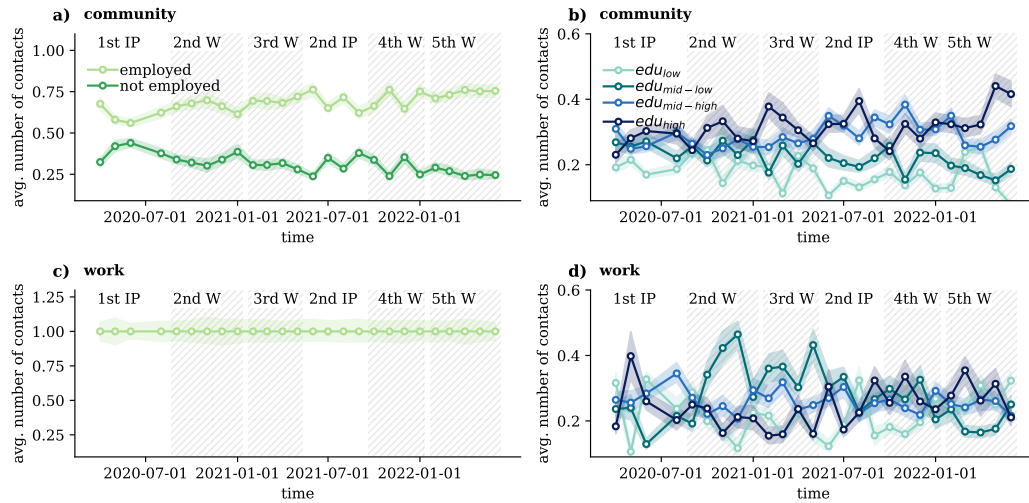


Figure B.10: (a-b) Relative smoothed average number of contacts in the community layer by different employment situation *a*) and education levels *b*). (c-d) Relative smoothed average number of contacts at the workplace by different employment situation *c*) and education level *d*). All the values are shown as the median and the IQ of 1.000 bootstrapped samples.

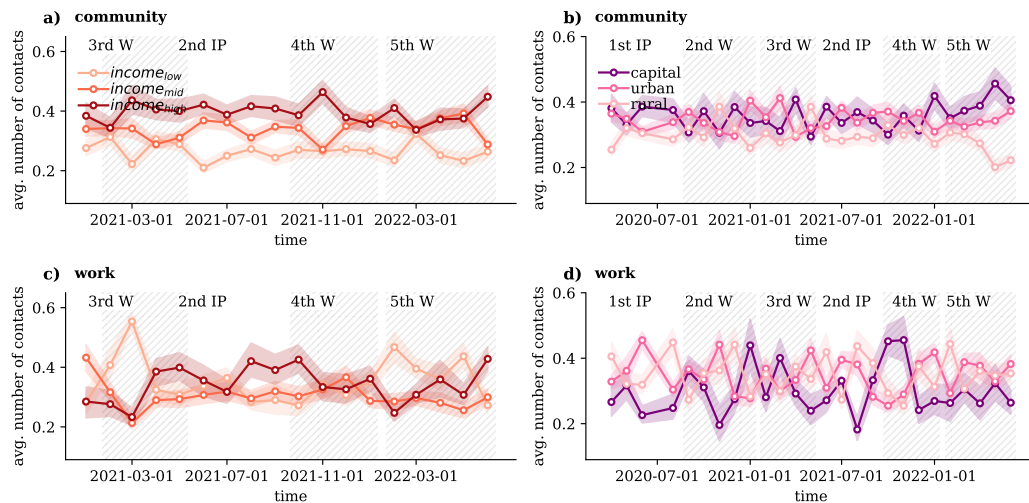


Figure B.11: (a-b) Relative smoothed average number of contacts in the community layer by different income levels *a*) and settlement types *b*). (c-d) Relative smoothed average number of contacts at the workplace by different income levels *c*) and settlement types *d*). All the values are shown as the median and the IQR of 1000 bootstrapped samples.

B.4.2 Average number of community contacts by sub-groups and age groups

To show the robustness of our findings over different age groups in Fig. B.12 we report the evolution of community contacts over time decoupled by education level, employment, income level and settlement. While the correlation with age influences the magnitude of differences among the examined sub-groups, the conclusion discussed in the main text appears to be still valid.

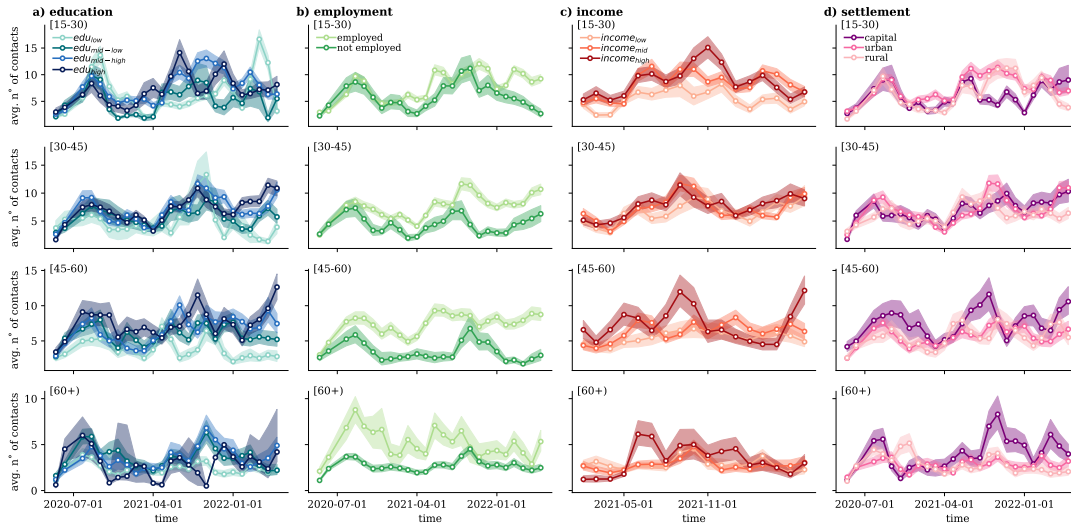


Figure B.12: Average number of contacts in the community layer by different age groups (*rows*) and sub-groups (*columns*: education levels *a*), employment situation *b*, income levels *c*) and settlement *d*). All the values are shown as the median and the IQR of 1000 bootstrapped samples.

B.4.3 Contact Matrices

For each of the periods considered in this study, we computed the age contact matrix C_{ij} considering the whole population as shown in Fig. B.13. These are the matrices that have been fed to the *conventional* SEIR model. Instead for the *extended* SEIR model, we computed the decoupled contact matrices ($C_{di,j}$) considering different dimensions as well: (i) employment situation (Fig. B.14), (ii) education level (Fig B.15), (iii) settlement (Fig. B.16), and (iv) income level (Fig. B.17). For all the matrices we report the median of 1000 bootstrapped samples. All the matrices have been computed considering the contact at work, in the community and in the household.

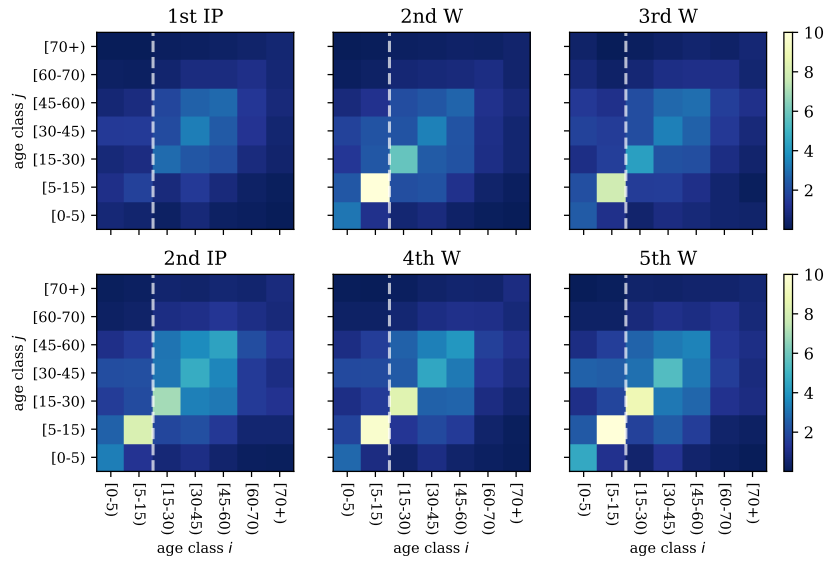


Figure B.13: Age contact matrices (C_{ij}) for different periods.

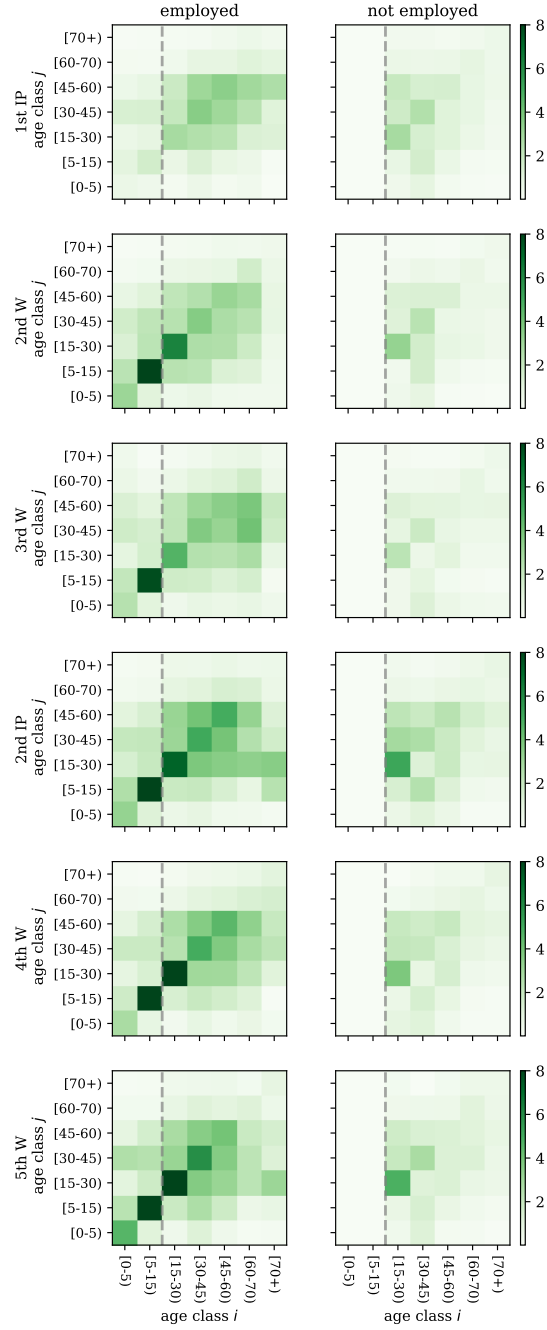


Figure B.14: Age contact matrices decoupled by employment situation ($C_{\text{employment},i,j}$) for different periods.

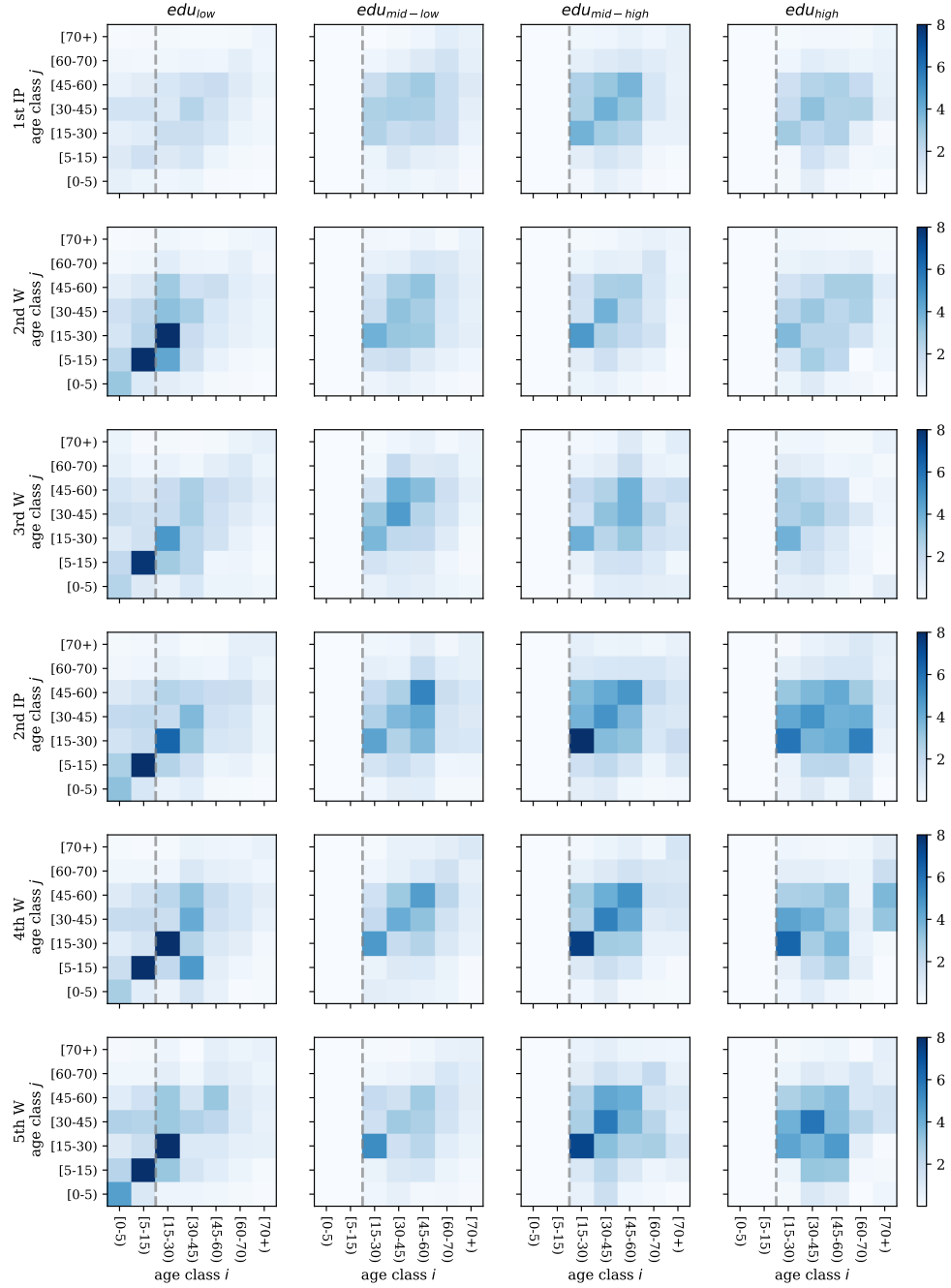


Figure B.15: Age contact matrices decoupled by education level ($C_{\text{education},i,j}$) for different periods.

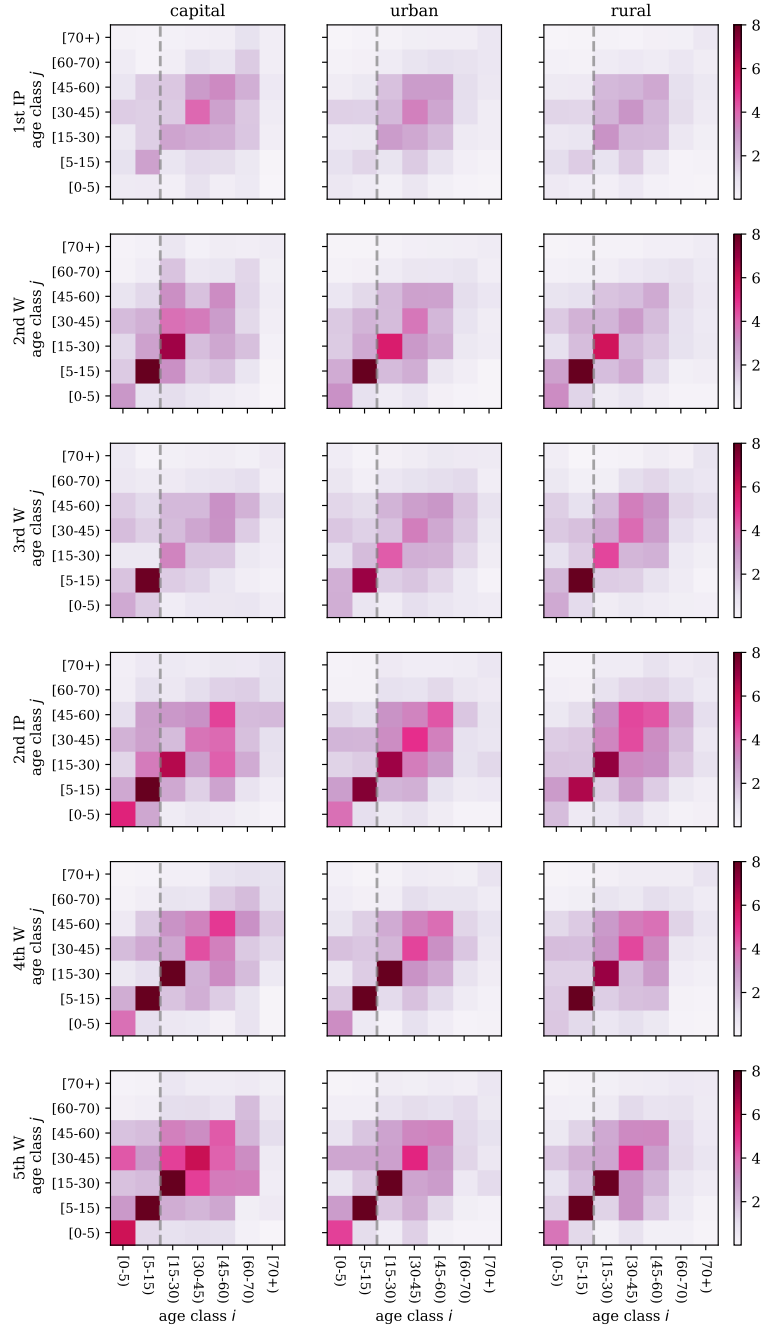


Figure B.16: Age contact matrices decoupled by settlement type ($C_{\text{settlement } i, j}$) for different periods.

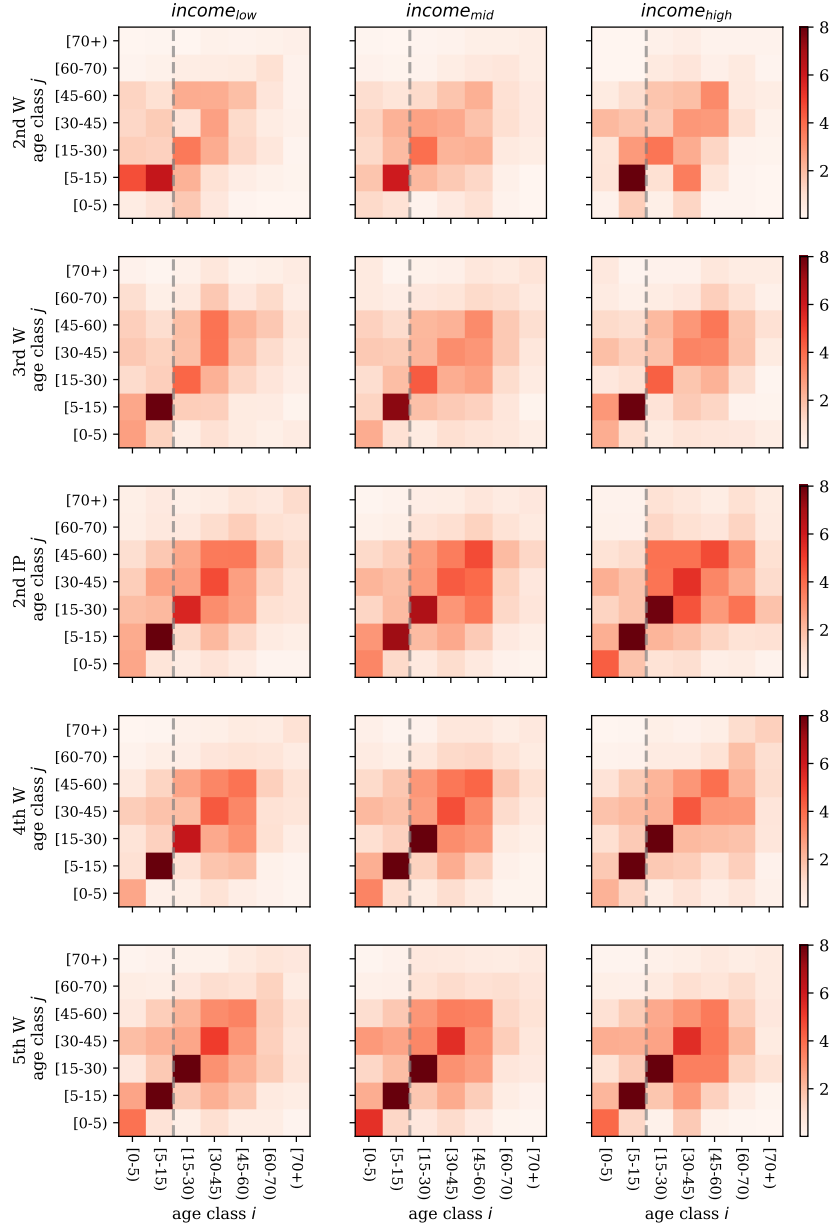


Figure B.17: Age contact matrices decoupled by income level ($C_{income_{i,j}}$) for different periods.

B.5 Vaccination uptake

Here we show the probability of getting vaccinated against COVID-19 given age and another dimension of interest. Namely, we consider (i) employment situation, (ii) education level, (iii) settlement, and (iv) income level. From Fig. B.18 we can observe that privileged groups of the population tend to have higher vaccination uptake across all age groups. This finding is consistent over the four different periods considered in the analysis.

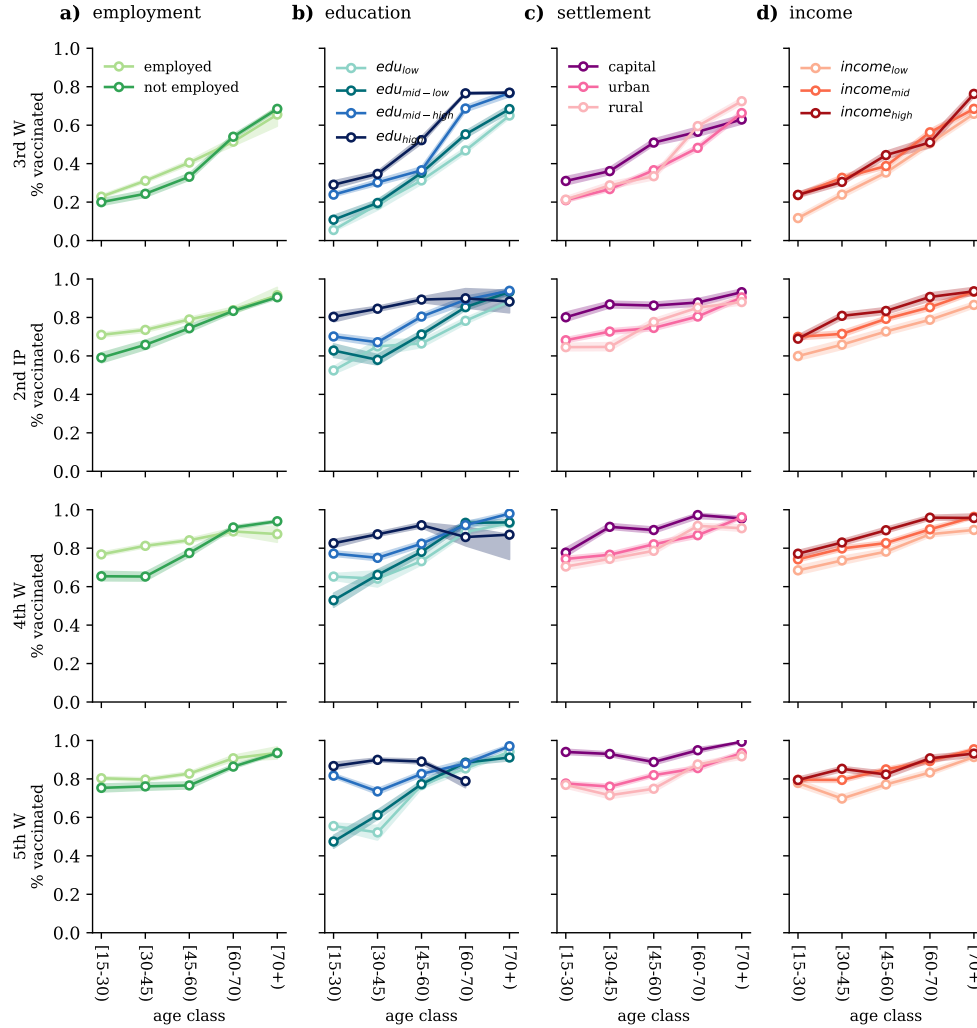


Figure B.18: Probability of getting vaccinated given age and (i) employment situation, (ii) education level, (iii) settlement, and (iv) income level. All the values are shown as the median and the IQR of 1000 bootstrapped samples.

B.6 Extended SEIRD with vaccination

We extend the traditional *SEIRD* model to account for different vaccination uptake of different groups of the population. Namely, each of the compartments is now considered separately for vaccinated and unvaccinated individuals. We define as g_1 and g_2 the efficiency of the vaccination respectively against infection and against death. In addition,

we consider the delay in the official registrations of deaths by adding a new compartment Da and a delay of Δ^{-1} days. The equations of the model are presented in equation (B.3).

$$\begin{aligned}
\dot{S}_{\bar{d},i} &= -\lambda_{\bar{d},i}S_{\bar{d},i} \\
\dot{S}v_{\bar{d},i} &= -(1-g_1)\lambda_{\bar{d},i}Sv_{\bar{d},i} \\
\dot{E}_{\bar{d},i} &= \lambda_{\bar{d},i}S_{\bar{d},i} - \epsilon E_{\bar{d},i} \\
\dot{E}v_{\bar{d},i} &= (1-g_1)\lambda_{\bar{d},i}Sv_{\bar{d},i} - \epsilon E v_{\bar{d},i} \\
\dot{I}_{\bar{d},i} &= \epsilon E_{\bar{d},i} - \gamma I_{\bar{d},i} \\
\dot{I}v_{\bar{d},i} &= \epsilon E v_{\bar{d},i} - \gamma I v_{\bar{d},i} \\
\dot{R}_{\bar{d},i} &= \gamma(1 - \text{IFR}_i)I_{\bar{d},i} \\
\dot{R}v_{\bar{d},i} &= \gamma(1 - (1-g_2)\text{IFR}_i)I v_{\bar{d},i} \\
\dot{D}_{\bar{d},i} &= \gamma \text{IFR}_i I_{\bar{d},i} \\
\dot{D}v_{\bar{d},i} &= \gamma(1-g_2)\text{IFR}_i I v_{\bar{d},i} \\
\dot{D}_{a\bar{d},i} &= \Delta^{-1}\dot{D}_{\bar{d},i} \\
\dot{D}v_{a\bar{d},i} &= \Delta^{-1}\dot{D}v_{\bar{d},i}
\end{aligned} \tag{B.3}$$

The force of infection is defined as in eq. (B.4)

$$\lambda_{\bar{d},i}(t) = \phi \sum_j \frac{C_{\bar{d},ij}}{N_j} [I_j + I v_j] \tag{B.4}$$

B.7 Epidemic Simulations

We developed stochastic, discrete-time, compartmental models using chain binomial processes to simulate the transitions among compartments. Specifically, at each time step t , the model samples the number of individuals in group $(\bar{d}, i)$ and compartment X transitioning to compartment Y from $PrBin(X_{\bar{d},i}(t), p_{X_{\bar{d},i} \rightarrow Y_{\bar{d},i}}(t))$. Here, $p_{X_{\bar{d},i} \rightarrow Y_{\bar{d},i}}(t)$ represents the transition probability.

To illustrate this, let's consider the number of individuals in the group $(\bar{d}, i)$ and compartment S that at time t become exposed transiting to compartment E . Thus, the number of individuals in $S_{\bar{d},i}(t)$ getting exposed are extracted from a $PrBin(S_{\bar{d},i}(t), \lambda_{\bar{d},i}(t))$ where $\lambda_{\bar{d},i}(t)$ is the *force of infection*.

Furthermore, in order to account for the variability of contacts in our data, for each simulation that we run we use a static decoupled contact matrix that we compute from a bootstrapped sample of our data as described in Section 3. The decoupled contact matrices have been computed considering the contacts happening at work, in the community and with family members.

The model has been initialized by computing the population distributions from the MASZK data, while we set the Hungarian population size to 9.750.000.

While the epidemiological parameters are set to realistic values to closely simulate the characteristics of Covid-19. These values are retrieved from the literature. In particular, ϵ is set to 0.25; γ is set to 0.4 [41, 176, 177]. The transmission rate β is computed in each of the periods using the Next Generation Matrix approach on the aggregate age-contact matrices corresponding to the periods analysed. We fixed $R_0 = 2.5$ and we derived β using Eq. (A.13).

In the simulations in which we introduce vaccination, we respectively set the $g_1 = 0.6$ and $g_2 = 0.8$ [178, 179]. The initial size of the epidemic is set to 5. All the results in the main text refer to the median over 1000 simulations of the model.

B.7.1 Impact of different contact patterns

In this section, we show the results of the *extended SEIR* model when differences in age contact matrices are considered for different sub-groups of the population. Particularly, we model age contact matrices differentiating individuals along their employment situation, education level, settlement and income level. For each of the dimensions considered, we run the *extended SEIR* model. We look at (i) how the prediction of this model differs from the *conventional SEIR* and (ii) how the attack rate differs for different subgroups as a result of their differences in contact patterns. Specifically, in Fig. B.19 we show the difference between attack rate by age group as predicted by the *extended SEIR* model and the *conventional SEIR* model. The results are shown for each of the periods considered in the analysis. Again, as demonstrated in the main text, the *conventional SEIR* model tends to overestimate the attack rate by age group with respect to *extended SEIR* model, particularly when employment situation and education are taken into account.

In Fig. B.20 we show the output of each of the *extended SEIR* models in terms of attack rate by age by differentiating along the subgroups taken into account. Results are shown for each of the periods considered in the analysis.

As shown in the main text the analysis over the other periods confirms that employed and highly educated individuals happened to be the most infected groups in all age groups. When decoupling age contact matrices by settlement and income, although differences appear smaller between groups, high-income individuals and the ones living in the capital are more infected, particularly elderly ones with age 60+.

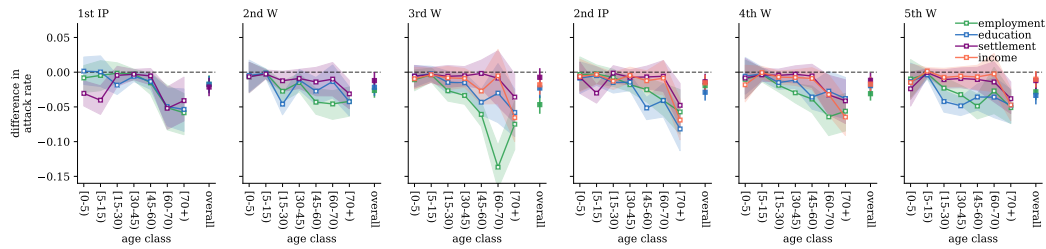


Figure B.19: Difference in the attack rate by age as predicted by the *classical* ($Mw/C_{i,j}$) and the *extended* ($Mw/C_{di,j}$) model, when different the dimensions are considered. Results are shown for each of the periods (*columns*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

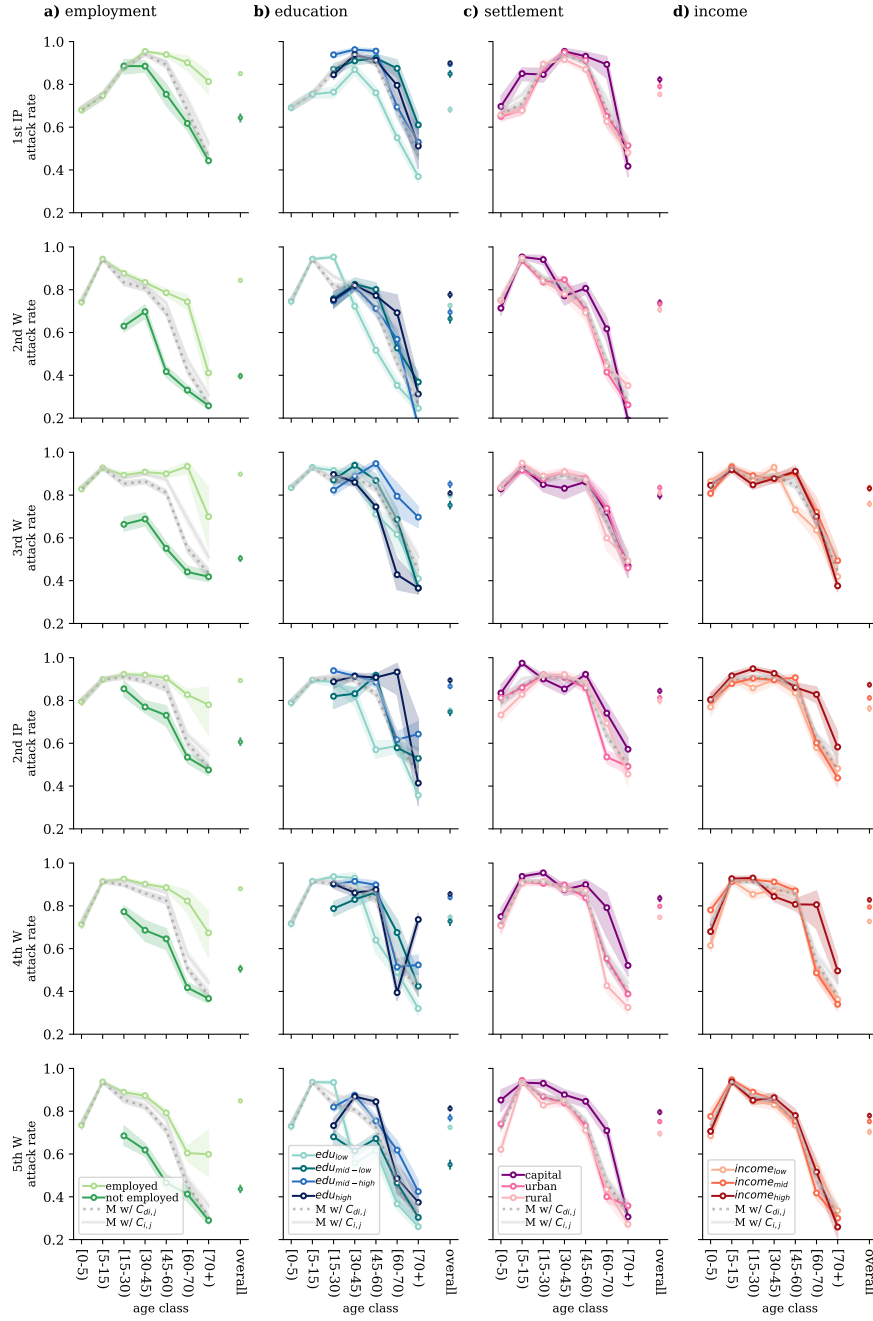


Figure B.20: Attack rate by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

Sensitivity analysis of R_0

Here we conduct a sensitivity analysis by reiterating the analysis with varying values of the basic reproductive number (R_0). Figure B.21 illustrates the variation in the difference of attack rates, across *age groups* and *overall (rows)* as predicted by the *classical* ($Mw/C_{i,j}$) and *extended* ($Mw/C_{di,j}$) models, as R_0 increases. These findings indicate that, as the epidemic intensifies with higher R_0 values, the difference in attack rates exhibits a decreasing trend with lower uncertainty for younger age groups. Conversely, most older age groups display an increasing trend, also with lower uncertainty. The overall difference in the attack rate of the population shows an increasing trend with lower uncertainty. This sensitivity analysis underscores significant discrepancies between the two models in numerous scenarios, highlighting that the *classical SEIR* model tends to overestimate the attack rate, particularly when compared to the *extended SEIR* model, which incorporates employment.

Furthermore, Figures B.22 and B.23 present the attack rates by age for different subgroups of the population respectively for $R_0 = 1.5$ and $R_0 = 3.5$. Here the results show that the increased transmissibility of the epidemic impacts equally each group of the population by leaving almost unchanged the differences among the subgroups of the population.

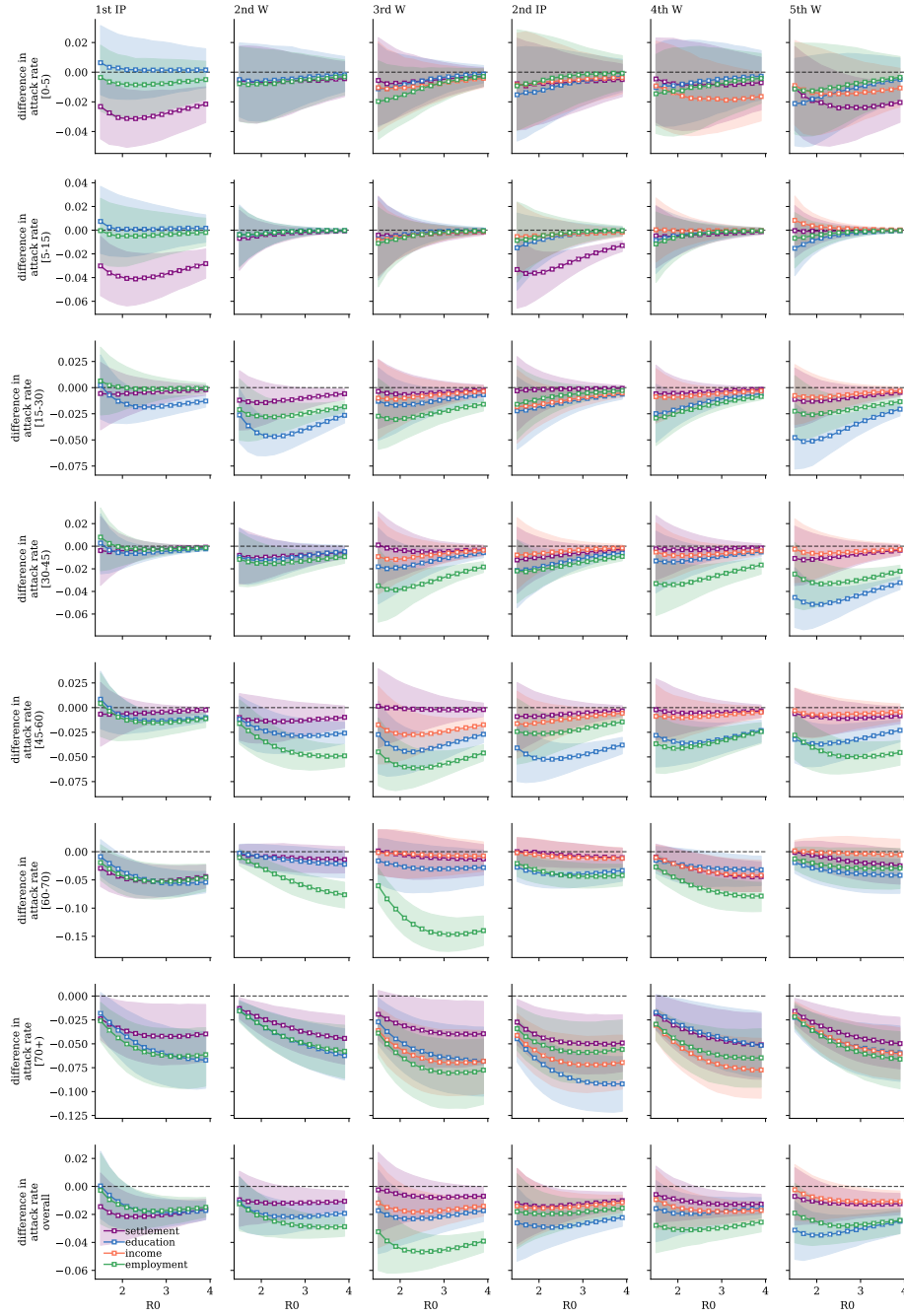


Figure B.21: Difference in the attack rate by age and overall (*rows*) as predicted by the *classical* ($Mw/C_{i,j}$) and the *extended* ($Mw/C_{di,j}$) model, for different value of R_0 . Results are shown for each of the periods (*columns*). Epidemiological parameters: $\mu = 0.4, \epsilon = 0.25$. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulation.

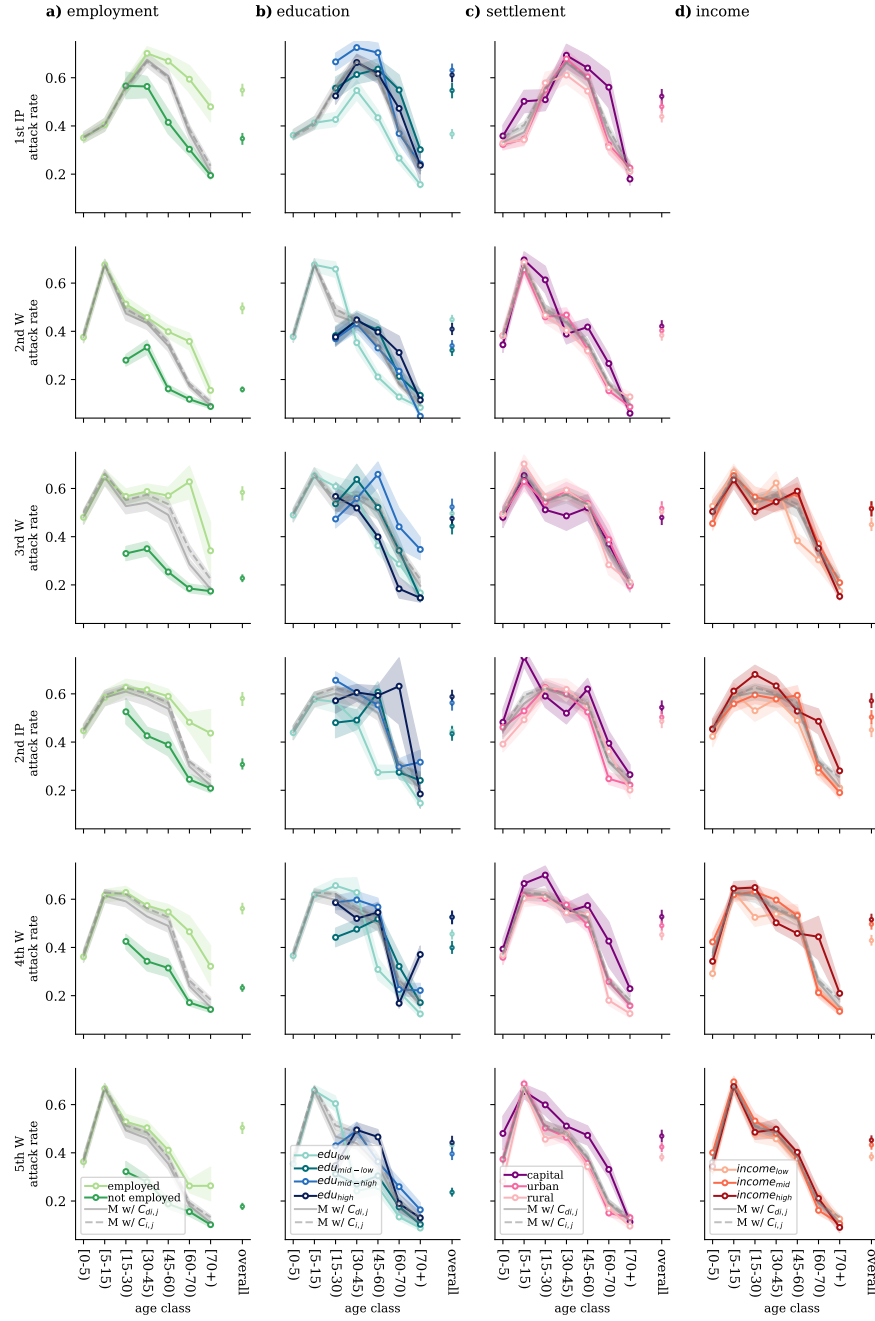


Figure B.22: Attack rate by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 1.5$. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

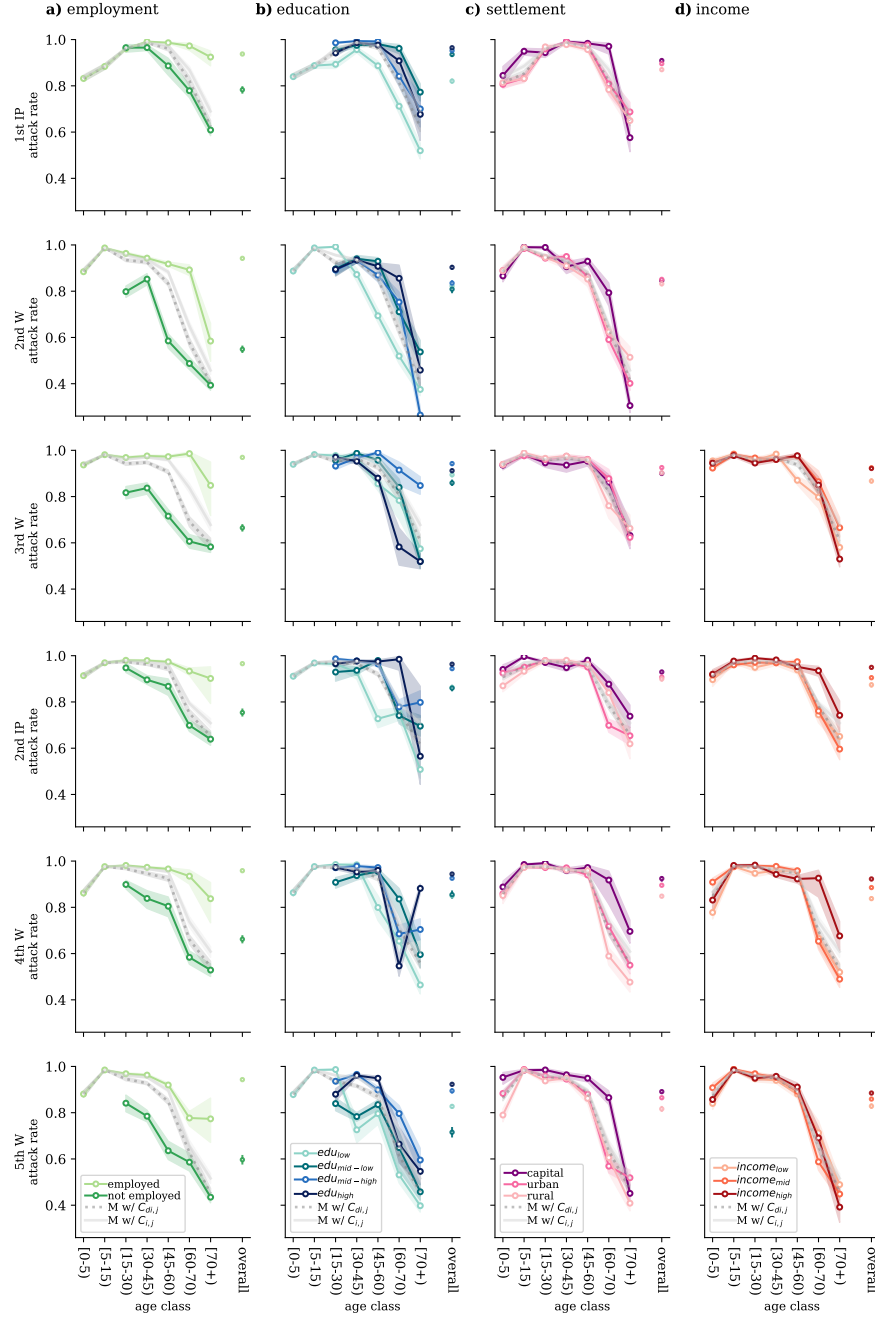


Figure B.23: Attack rate by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 3.5$. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

B.7.2 Impact of different vaccination uptake

In order to show the impact of different vaccination uptake here we show the *averted attack rate* by age due to vaccination (Fig. B.24). Specifically, we show the difference among the attack rate by age, for the different subgroups as predicted by the *extended SEIR* in the non-vaccination scenario with respect to the one in which individuals get vaccinated according to their age and subgroup- as shown in Fig. B.18.

The findings from the additional periods support the observations discussed in the main text. Indeed, Fig. B.24 clearly shows that vaccination benefits are disproportionately advantageous for more privileged population groups.

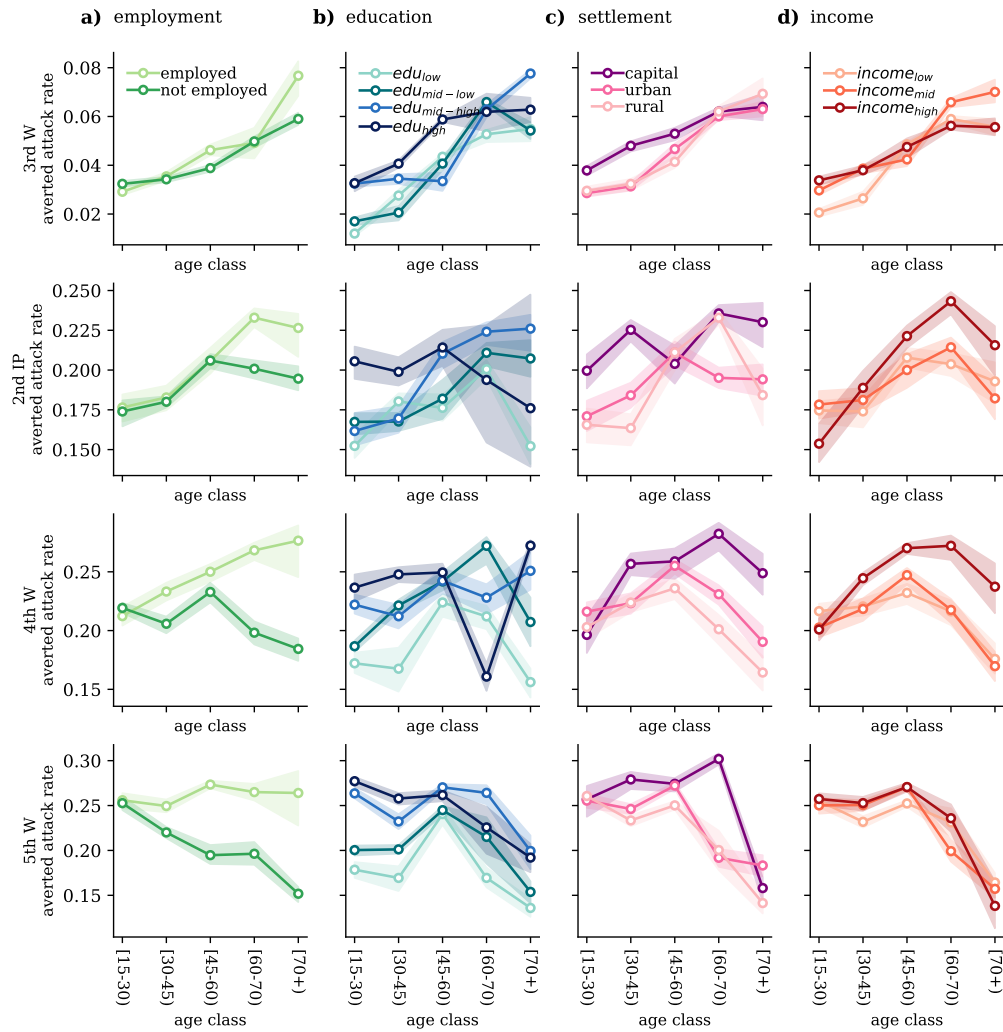


Figure B.24: Difference in attack rate due to vaccination by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Efficacy of vaccination against infection (g_1) is set to 0.6. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

Sensitivity analysis of vaccination efficacy against infection (g_1)

To investigate the robustness of the results presented above we conduct a sensitivity analysis by varying the efficacy of vaccination (g_1). Namely, Figure B.25 presents results for simulations with g_1 set to 0.2, while Figure B.26 displays outcomes for simulations with g_1 set to 0.9.

By design, $g_1 = 0$ corresponds to the non-vaccination scenario, thus the disparity in attack rate differences attributable to vaccination diminishes towards zero as g_1 decreases. Consequently, as expected, when g_1 is set to 0.2 (Fig. B.25), the overall magnitude of the difference becomes smaller. Nevertheless, discernible and meaningful differences persist among subgroups of the population. Conversely, when g_1 is set to 0.9 (Fig. B.26), both the overall magnitude of the difference and the disparities among subgroups increase, accentuating the distinctions among population groups that are undergoing higher vaccination rates.

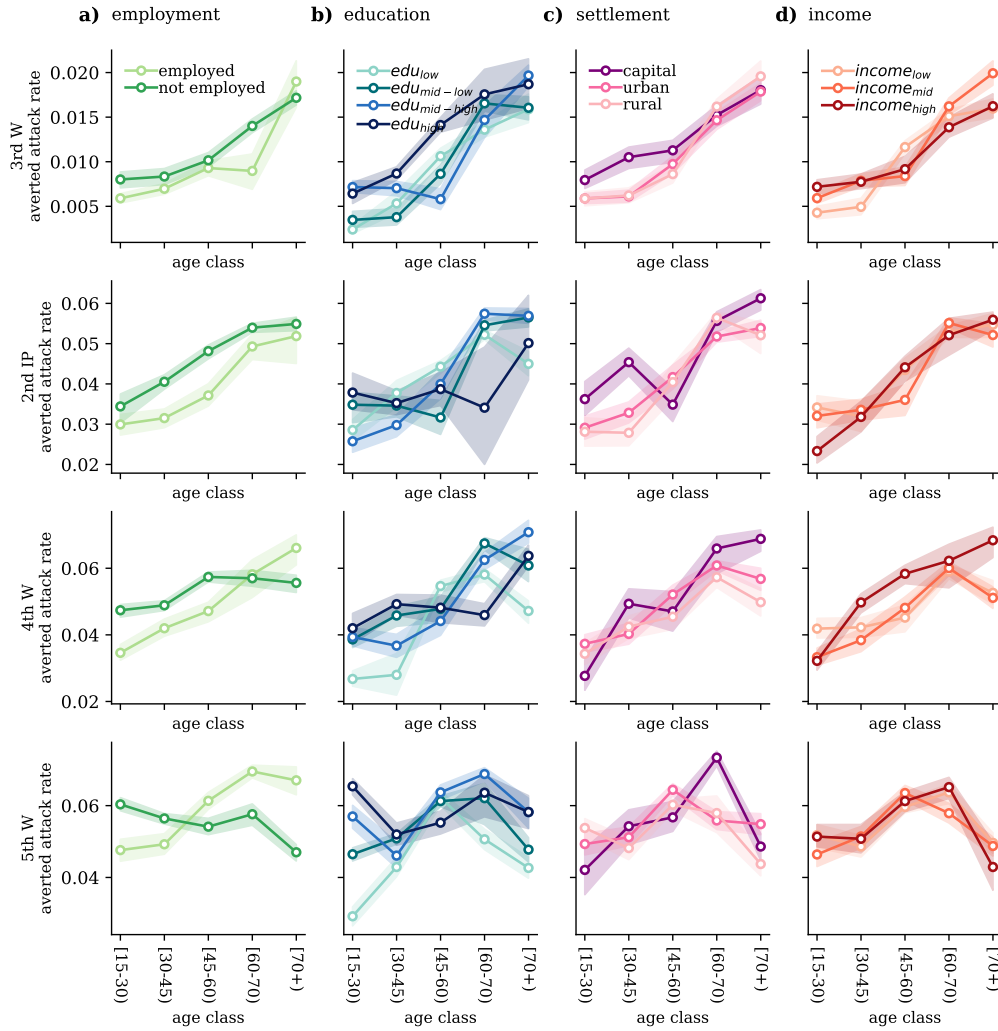


Figure B.25: Difference in attack rate due to vaccination by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Efficacy of vaccination against infection (g_1) is set to 0.2. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

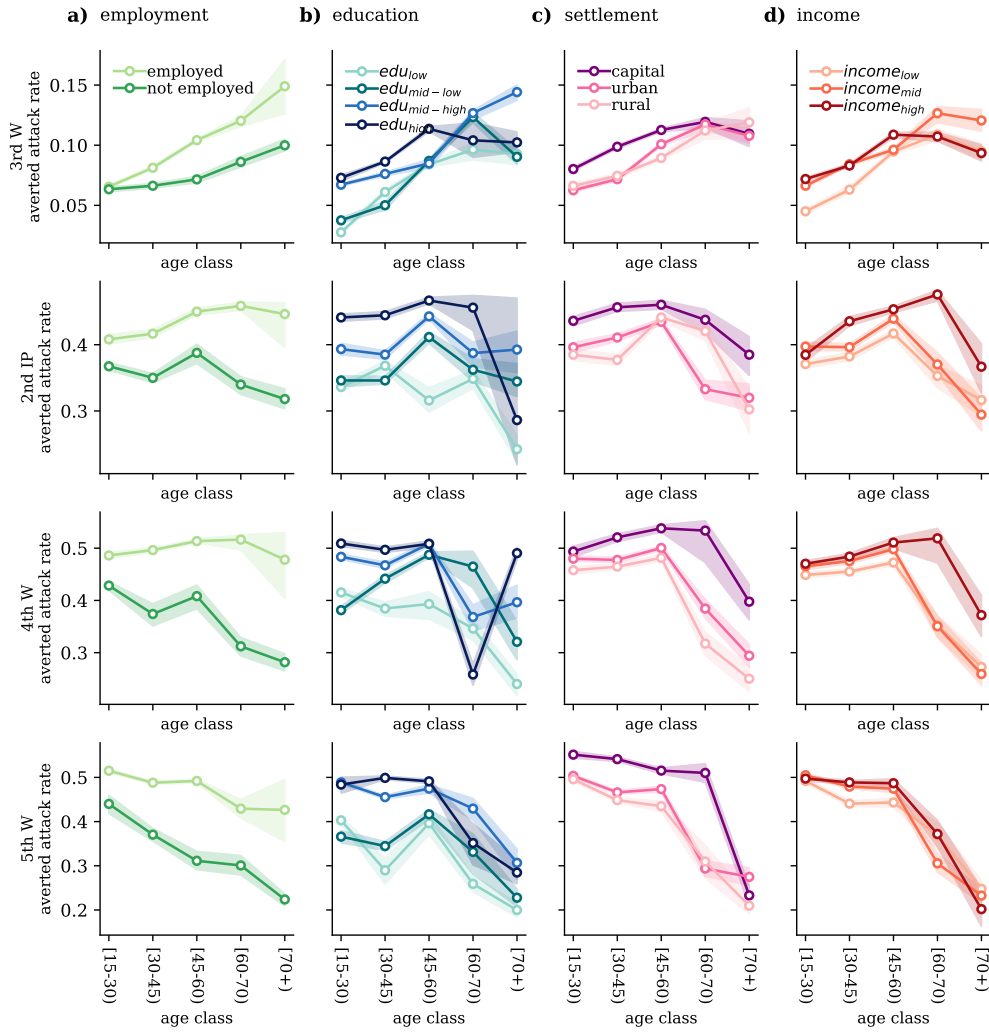


Figure B.26: Difference in attack rate due to vaccination by age and employment situation, education level, settlement and income (*columns*), in different periods (*rows*). Epidemiological parameters: $\mu = 0.4$, $\epsilon = 0.25$, and $R_0 = 2.5$. Efficacy of vaccination against infection (g_1) is set to 0.9. Simulations start with $I_0 = 5$ initial infectious seed. Results were computed over 1000 simulations.

B.8 Model Calibration

We calibrate a SEIRD model with vaccination by modelling differences in contact patterns and vaccination uptake among employed and not-employed individuals belonging to different education levels. In particular, we calibrate the free parameters of the model using an Approximate Bayesian Computation (ABC) technique [146, 147]. First, we define the prior distributions of the free parameters $P(\theta)$, a number of accepted sets N , an error metric $m(E, E')$, and a tolerance δ . We start sampling a set of parameters θ from $P(\theta)$, and generate an instance of the model using these parameters. Then, using the chosen error metric we compare an output quantity E' of the model with the corresponding real quantity E : if $m(E, E') < \delta$ then we accept the set θ , otherwise we reject it. We repeat this accept/reject step until N parameter sets are accepted. The empirical distribution of the accepted sets is an approximation of their real posterior distribution. Finally, we generate an ensemble of possible epidemic trajectories sampling parameter sets from the posterior distributions. In this work, we consider the following free parameters and prior distributions:

- the transmission rate parameter β : the prior distribution is set to $U(0.02, 0.15)$
- the delay in reporting deaths Δ : the prior distribution is set to $U(5, 20)$
- the initial recovered population R : the prior distribution is set to $U(700K, 3500K)$
- the initial exposed population E : the prior distribution is set to $U(100, 3K)$
- the initial infected population I : the prior distribution is set to $U(200, 9K)$

Given the total number of individuals Exposed, Infected and Recovered at $t = 0$, these are assigned to age group, employment situation and income level according to the population distribution.

We calibrate our model on the aggregate number of daily deaths from 09/2021 to 01/2022. For simplicity, as the percentage of those who were vaccinated was quite stable [180] in the period considered we assume that the population got vaccinated at time 0. As an error metric, we use the *Median Absolute Percentage Error (MdAPE)*. We also set the number of accepted sets $N = 3000$ and the tolerance $\delta = 0.35$. In Fig. B.27 are shown the posterior distributions of the parameters calibrated through the ABC rejection algorithm.

The fixed parameters of the model have been informed from the literature. In particular:

- the efficacy of the vaccine against infection and against death, is modelled as a normal distribution with mean respectively $g_1 = 0.7$, $g_2 = 0.8$, and standard deviation 0.05 [178, 179]. This choice has been made to account for the variability of the efficacy of the different vaccination types and against the different variants.
- the recovery rate $\mu = 1/2.5$ [41, 176, 177]
- the incubation period $\epsilon = 1/4$ [41, 176, 177]
- the infection fatality rate by age IFR_i is set as in Table B.9 [181]

In Fig. B.27 we report the number of daily real and simulated deaths (median and IQR).

Age group	IFR
[0-5)	0.001%
[5-15)	0.001%
[15-30)	0.005%
[30-45)	0.02%
[45-60)	0.2%
[60-70)	0.7%
[70+)	3.1%

Table B.9: Infection Fatality Rate by age group

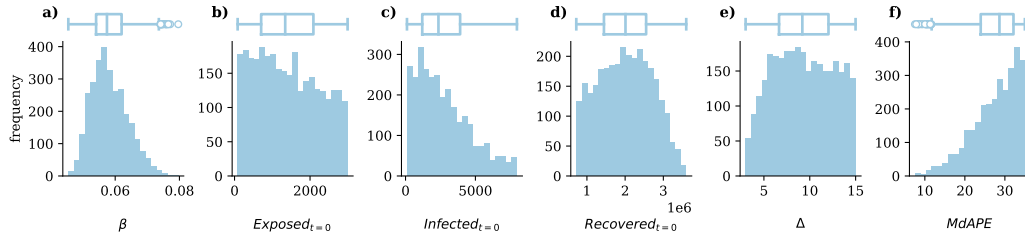


Figure B.27: (a-e) Posterior distribution of the calibrated parameters of the model (f) Distribution of the MdAPE (ϵ).

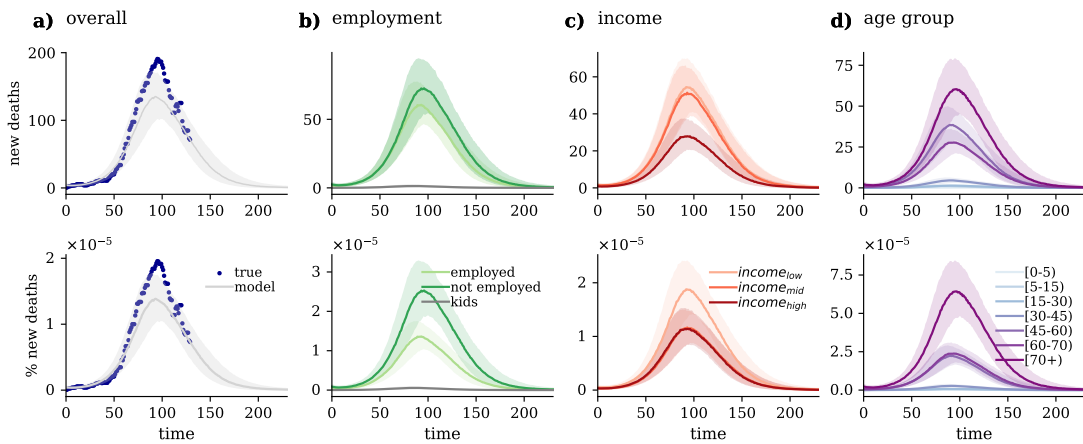


Figure B.28: (a-d) Number of daily deaths real (blue dots) and simulated (grey) overall (a), by education (b), employment (c), and age group (d). (e-h) Percentage of daily deaths real (red line) and simulated (grey) overall (e), by education (f), employment (g), and age group (h). Results refer to the median and IQR of 1000 simulations.

Appendix C

Appendix chapter 3

C.1 Synthetic generalized contact matrices

In this section, we provide details about the generation of synthetic generalized contact matrices we built to explore the possible effects of different mixing patterns among individuals. For simplicity, we first consider only two dimensions: age and a socio-economic status (SES) variable (though any other categorical variable could be used).

We developed a model to derive $G_{\mathbf{a},\mathbf{b}}$ where $\mathbf{a} = (i, \alpha)$ and $\mathbf{b} = (j, \beta)$, and i and j refers to the age group while α and β to the SES of the ego and the alter respectively. We start from an empirical contact matrix $\mathbf{C}$ describing the contact rates between age brackets i and j . As mentioned above, we can define $\mathbf{R} = \mathbf{NC}$ as the total number of contacts between the two groups in the given period. While the matrix C_{ij} is not symmetric, R_{ij} is since the number of *raw* contacts between two groups is the same, i.e., $R_{ij} = C_{ij}N_i = R_{ji} = C_{ji}N_j$. To build the generalized contact matrices $G_{\mathbf{a},\mathbf{b}}$ we first split the total contacts R_{ij} across the second dimension, and then we compute the contacts rates. In other words $R_{ij} = \sum_{\alpha,\beta} (R_G)_{i\alpha,j\beta}$. The problem is then how to get values for the elements $(R_G)_{i\alpha,j\beta}$. It is useful to reflect on the properties of these matrices. First, $(R_G)_{i\alpha,j\beta} = (R_G)_{j\beta,i\alpha}$. In other words, the number of contacts that individuals in age group i and SES α have with individuals in age group j and SES β must equal the number of contacts that individuals in age group j and SES β have with individuals in age group i and SES α . This property implies that $(R_G)_{i\alpha,j\beta} = (R_G)_{i\alpha,j\beta}^\top$. Indeed, for a given pair of indices i and j , one can think about $(R_G)_{i\alpha,j\beta}$ as a $V_1 \times V_1$ matrix (since $\alpha, \beta \in [1, \dots, V_1]$), which describes the contact patterns between those two age groups across SES. For all $i = j$ the symmetry of matrix $(R_G)_{i\alpha,i\beta}$ implies $(R_G)_{i\alpha,i\beta} = (R_G)_{i\beta,i\alpha}$. In general this is not the case for $i \neq j$. Indeed, the overall symmetry of the matrix it is not enough to guarantee that $(R_G)_{i\alpha,j\beta} = (R_G)_{i\beta,j\alpha}$ (note how just the indices of the second dimension have been swapped). This would imply a much stronger internal symmetry for each pair i and j which is not generally required. Second, as mentioned above the sum over all α and β is set by the number of contacts between age group i and j : $R_{ij} = \sum_{\alpha,\beta} (R_G)_{i\alpha,j\beta}$.

Given these features, how can we set the entries of these matrices? As mentioned, for a given pair i and j the matrix $(R_G)_{i\alpha,j\beta}$ is of size $V_1 \times V_1$ since $\alpha, \beta \in [1, \dots, V_1]$. For any $i \neq j$, we need a model to set $W = V_1^2 - 1$ elements of the matrix (the minus one is due to the constraint introduced by the second property defined above). For all $i = j$ instead, the symmetry of the matrix is such that we need to set only $Y = V_1 + \frac{V_1(V_1-1)}{2} - 1$ elements. The first V_1 comes from the diagonal (i.e., $\alpha = \beta$), the $\frac{V_1(V_1-1)}{2}$ are instead the off-diagonal values, the minus one is due to the constraint of the total number of

contacts which is set by R_{ij} . Since $W > Y$ for all $V_1 \geq 2$, we can define the matrix by defining W values. In doing so, we assume an independence between age and SES. To set the W values we consider that for any i and j pair:

1. each SES α is generally responsible for P_α fraction of R_{ij} connections, where $\sum_\alpha P_\alpha = 1$
2. we assume that a fraction q_α of these are on the diagonal (i.e., in-group connections) and $1 - q_\alpha$ are instead off-diagonal. This parameter controls the assortativity of connections within each group.

In what follows, we consider the fractions P_α and q_α as input parameters. In case W is equal to this number ($2V_1 - 1$), the constraints imposed by our assumptions allow to define all the entries of the matrix. If instead $W > 2V_1 - 1$, other $W - 2V_1 + 1$ parameters are required.

Let us consider first the case in which $V_1 = 3$, $\alpha, \beta \in [1, 2, 3]$. In these settings $W = 8$, and $Y = 5$ which is equal to $2V_1 - 1$. Hence, defining the two set of input parameters is enough to obtain the matrix for any $i = j$, while still 3 additional parameters are needed for $i \neq j$. It is important to stress how the first property described above implies that we need to define only this number for all $i < j$. The correspondent values for $i > j$ are readily obtained just by transposing the matrices. It is useful, to split the two cases:

Case $i = j$. For all $i = j$ the matrices $(R_G)_{i\alpha, i\beta}$ take the form of

$$(R_G)_{i\alpha, i\beta} = R_{ii} \begin{pmatrix} q_1 P_1 & p_{i1, i2} & p_{i1, i3} \\ p_{i1, i2} & q_2 P_2 & p_{i2, i3} \\ p_{i1, i3} & p_{i2, i3} & q_3 P_3 \end{pmatrix} \quad (\text{C.1})$$

where each $p_{i\alpha, i\beta} \in [0, 1]$ is defined such that $(R_G)_{i\alpha, i\beta} = R_{ii} p_{i\alpha, i\beta}$. As the matrix needs to be symmetric we have written, for example, $p_{i1, i2}$ instead of $p_{i2, i1}$. Due to our assumptions, the sum of the off-diagonal elements of each row should sum to $(1 - q_\alpha)P_\alpha = \Pi_\alpha$, while on the diagonal we have $q_\alpha P_\alpha$. Thus the following system of equations must be respected:

$$\begin{cases} p_{i1, i2} + p_{i1, i3} = \Pi_1 \\ p_{i1, i2} + p_{i2, i3} = \Pi_2 \\ p_{i1, i3} + p_{i2, i3} = \Pi_3 \end{cases} \quad (\text{C.2})$$

The solution, if any, is unique and can be written as

$$\hat{\mathbf{p}} = \mathbf{A}^{-1} \mathbf{\Pi} = \frac{1}{2} \begin{pmatrix} \Pi_1 + \Pi_2 - \Pi_3 \\ \Pi_1 + \Pi_3 - \Pi_2 \\ \Pi_2 + \Pi_3 - \Pi_1 \end{pmatrix} \quad (\text{C.3})$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{\Pi} = \begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \Pi_3 \end{pmatrix}, \quad \hat{\mathbf{p}} = \begin{pmatrix} p_{i1, i2} \\ p_{i1, i3} \\ p_{i2, i3} \end{pmatrix} \quad (\text{C.4})$$

Let us consider now the case in which $V_1 = 4$. In this case, $Y = 9$. As before, $2V_1 - 1$ parameters are inputs (three P_α and four q_α). In this case, $Y > 2V_1 - 1$. Hence, the system of V_1 differential equations in $V_1(V_1 - 1)/2 = 6$ (off-diagonal) elements requires $Y - 2V_1 + 1 = 2$ free parameters. Indeed, the matrix $(R_G)_{i\alpha, i\beta}$ can be written as:

$$(R_G)_{i\alpha,i\beta} = R_{ii} \begin{pmatrix} q_1 P_1 & p_{i1,i2} & p_{i1,i3} & p_{i1,i4} \\ p_{i1,i2} & q_2 P_2 & p_{i2,i3} & p_{i2,i4} \\ p_{i1,i3} & p_{i2,i3} & q_3 P_3 & p_{i3,i4} \\ p_{i1,i4} & p_{i2,i4} & p_{i3,i4} & q_4 P_4 \end{pmatrix} \quad (C.5)$$

Considering that the sum of each off-diagonal element in each row α should sum to $(1 - q_\alpha)P_\alpha$ we obtain a system of four equations:

$$\begin{cases} p_{i1,i2} + p_{i1,i3} + p_{i1,i4} = \Pi_1 \\ p_{i1,i2} + p_{i2,i3} + p_{i2,i4} = \Pi_2 \\ p_{i1,i3} + p_{i2,i3} + p_{i3,i4} = \Pi_3 \\ p_{i1,i4} + p_{i2,i4} + p_{i3,i4} = \Pi_4 \end{cases} \quad (C.6)$$

As before, we can write the system of 6 equations in matrix form as $\mathbf{A}\hat{\mathbf{p}} = \mathbf{\Pi}$ where:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}, \hat{\mathbf{p}} = \begin{pmatrix} p_{i1,i2} \\ p_{i1,i3} \\ p_{i1,i4} \\ p_{i2,i3} \\ p_{i2,i4} \\ p_{i3,i4} \end{pmatrix}, \mathbf{\Pi} = \begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \Pi_3 \\ \Pi_4 \end{pmatrix} \quad (C.7)$$

Clearly the system is under-determined as it has four equations and six variables. To find the solution we need to fix two of such variables, for example, $p_{i2,i4}$ and $p_{i3,i4}$. Doing so, we can write the system of equations as $\mathbf{A}_r \hat{\mathbf{p}}_r = \mathbf{\Pi}_r$ where

$$\mathbf{A}_r = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \hat{\mathbf{p}}_r = \begin{pmatrix} p_{i1,i2} \\ p_{i1,i3} \\ p_{i1,i4} \\ p_{i2,i3} \end{pmatrix}, \mathbf{\Pi}_r = \begin{pmatrix} \Pi_1 \\ \Pi_2 - p_{i2,i4} \\ \Pi_3 - p_{i3,i4} \\ \Pi_4 - p_{i2,i4} - p_{i3,i4} \end{pmatrix} \quad (C.8)$$

The subscript r stands for reduced, as by fixing the $Y - 2V_1 + 1$ variables we can move from a 4×6 matrix to a 4×4 . The solution, if any, is now unique and can be written as:

$$\hat{\mathbf{p}}_r = \mathbf{A}_r^{-1} \mathbf{\Pi}_r \quad (C.9)$$

For any general V_1 , the method described above allows to define the matrix $(R_G)_{i\alpha,i\beta}$ passing through an under-defined system of linear equations. By setting the potential $Y - 2V_1 + 1$ free parameters the solution, if it exists, is unique. However, we expect that some solutions might not be physical. Indeed, the matrix entries must be all equal or larger than zero. Any solution that leads to negative values is not physical and must be neglected. This condition adds constraints to the possible values of the Y parameters. It is important to stress how the free parameters are in general functions of i . For simplicity, and to avoid the need to set values across all age groups, we set them equal for all age groups.

Case $i < j$. For all $i < j$ the matrices $(R_G)_{i\alpha,j\beta}$ take the form of

$$(R_G)_{i\alpha,j\beta} = R_{ij} \begin{pmatrix} q_1 P_1 & p_{i1,j2} & p_{i1,j3} \\ p_{i2,j1} & q_2 P_2 & p_{i2,j3} \\ p_{i3,j1} & p_{i3,j2} & q_3 P_3 \end{pmatrix} \quad (\text{C.10})$$

where each $p_{i\alpha,j\beta} \in [0, 1]$ is defined such that $(R_G)_{i\alpha,j\beta} = R_{ij} p_{i\alpha,j\beta}$. The difference with respect to the previous case is that, since $i \neq j$, we have $p_{i\alpha,j\beta} \neq p_{i\beta,j\alpha}$. Hence, the matrix $\mathbf{p}$ is not symmetric for swaps of indices in the second dimension. As such, we need to find a model to set the $W = V_1^2 - 1$ values. However, as in the previous case, due to our assumptions, the sum of the off-diagonal elements of each row should sum to $(1 - q_\alpha)P_\alpha = \Pi_\alpha$, while on the diagonal we have $q_\alpha P_\alpha$. Thus the following system of equations must be respected:

$$\begin{cases} p_{i1,j2} + p_{i1,j3} = \Pi_1 \\ p_{i2,j1} + p_{i2,j3} = \Pi_2 \\ p_{i3,j1} + p_{i3,j2} = \Pi_3 \end{cases} \quad (\text{C.11})$$

The system of equations is clearly under-defined. Hence, even for the scenario with $V_1 = 3$ we have three parameters (one per equation). For example, we could set as parameters $p_{i1,j2}$, $p_{i2,j1}$, and $p_{i3,j1}$. This choice leads to the solution

$$\begin{cases} p_{i1,j3} = \Pi_1 - p_{i1,j2} \\ p_{i2,j3} = \Pi_2 - p_{i2,j1} \\ p_{i3,j2} = \Pi_3 - p_{i3,j1} \end{cases} \quad (\text{C.12})$$

Since the values of all the entries are defined between 0 and 1, each parameter is bounded by the following conditions: $0 \leq p_{i1,j2} \leq \Pi_1$, $0 \leq p_{i2,j1} \leq \Pi_2$, and $0 \leq p_{i3,j1} \leq \Pi_3$. Considering these constraints and the simple structure of Eq. C.12 for $V_1 = 3$, we can set the three free parameters as $p_{i1,j2} = a_1 \Pi_1$, $p_{i2,j1} = a_2 \Pi_2$, and $p_{i3,j1} = a_3 \Pi_3$ where a_1 , a_2 , and a_3 are unknowns defined between zero and one. The use of a_i with $i = [1, 2, 3]$, though not necessary, allows for a more physical interpretation of the free parameters as a fraction of the off-diagonal contacts. It is important to stress how once we have picked these three parameters, the matrix $(R_G)_{i\alpha,j\beta}$ is fully defined as well as the matrix $(R_G)_{j\beta,i\alpha}$ which is the transpose. Furthermore, we have three parameters for each $i < j$, since the entries are in principle functions of the age groups. However, for simplicity we set the same three free parameters across all age combinations.

Let us consider now the case in which $V_1 = 4$. In this case, $W = V_1^2 - 1 = 15$. As before, $2V_1 - 1$ parameters are inputs (three P_α and four q_α). In this case, $W > 2V_1 - 1$. Hence, the system of V_1 differential equations in $V_1(V_1 - 1)$ (off-diagonal) elements required $W - 2V_1 + 1 = 8$ free parameters. We note how in this case, and more in general for any $V_1 > 3$, we cannot simply adopt, without introducing further assumptions, a_i values as we did above for $V_1 = 3$. The most general solution can be obtained by defining directly some of the entries of the matrix (i.e., free parameters) and solving the resulting system of equations.

C.1.1 Random mixing regime

By setting $P_\alpha = q_\alpha = N_\alpha/N$ (where $N_\alpha = \sum_i N_{i\alpha}$) the model drastically simplifies and the contacts division of the total contact R_{ij} between age groups i and j across α and β equal the product of the fraction of these two groups (independently of age).

This is the random mixing regime, indeed, if contacts are random across groups the expectation value of this process leads to the aforementioned product. To provide a concrete example, let us consider a simple case where we have only one SES dimension and $V_1 = 3$. As described above, for all $i < j$ in these settings $W = 8$ and

$$(R_G)_{i\alpha,j\beta} = R_{ij} \begin{pmatrix} q_1 P_1 & p_{i1,j2} & p_{i1,j3} \\ p_{i2,j1} & q_2 P_2 & p_{i2,j3} \\ p_{i3,j1} & p_{i3,j2} & q_3 P_3 \end{pmatrix} \quad (\text{C.13})$$

Now we can set the parameters in the diagonal equal to the fraction of the population of each group:

$$(R_G)_{i\alpha,j\beta} = R_{ij} \begin{pmatrix} \frac{N_1}{N} \frac{N_1}{N} & p_{i1,j2} & p_{i1,j3} \\ p_{i2,j1} & \frac{N_2}{N} \frac{N_2}{N} & p_{i2,j3} \\ p_{i3,j1} & p_{i3,j2} & \frac{N_3}{N} \frac{N_3}{N} \end{pmatrix} \quad (\text{C.14})$$

The off-diagonal values are defined similarly

$$(R_G)_{i\alpha,j\beta} = R_{ij} \begin{pmatrix} \frac{N_1^2}{N^2} & \frac{N_1 N_2}{N^2} & \frac{N_1 N_3}{N^2} \\ \frac{N_1 N_2}{N^2} & \frac{N_2^2}{N^2} & \frac{N_2 N_3}{N^2} \\ \frac{N_1 N_3}{N^2} & \frac{N_2 N_3}{N^2} & \frac{N_3^2}{N^2} \end{pmatrix} \quad (\text{C.15})$$

The assumption that the division of the R_{ij} contacts for SES depends only on the total fraction of individuals in SES induces symmetry in these matrices even for the case $i < j$. We note how, independently of the number of SES groups, assigning contacts at random proportionally to the populations allows us to define matrices without the need to set free parameters. It is easy to show that by setting $P_\alpha = q_\alpha = N_\alpha/N$, the matrices are defined in such a way that the on-diagonal elements have a $P_\alpha q_\alpha$ fraction of contacts while the off-diagonal elements in each row sum to $P_\alpha(1 - q_\alpha) = \Pi_\alpha$.

Finally, we note how the more general way to define a homogeneous mixing scenario is done considering $(R_G)_{i\alpha,j\beta} = R_{ij} \frac{N_{i\alpha}}{N_i} \frac{N_{j\beta}}{N_j}$. In this case, we split the total contacts considering the fraction of individuals in each age and SES group. In doing so, the potential dependence of age is fully considered and the homogeneous mixing is defined for each pair of age groups. This is the general case we have considered in the corollary of the theorem described in details below. It is important to notice that this model does not, in general, lead to the symmetry of the $\mathbf{p}$ matrices for $i < j$. However, it reduces to the previous assuming that $\frac{N_{i\alpha}}{N_i} = \frac{N_\alpha}{N}$ for $\forall i$, hence when the fraction of individuals in each SES is independent of age.

C.2 Epidemic models with generalized contact matrices

C.2.1 Numerical simulations

To investigate the effect of the generalized contact matrices $G_{\mathbf{a},\mathbf{b}}$ on the transmission dynamics, we developed a stochastic, discrete-time, compartmental model where the transitions among compartments are simulated through chain binomial processes. In particular, at time step t the number of individuals in group $\mathbf{a}$ and compartment X transiting to compartment Y is sampled from $PrBin(X_{\mathbf{a}}(t), p_{X_{\mathbf{a}} \rightarrow Y_{\mathbf{a}}}(t))$, where $p_{X_{\mathbf{a}} \rightarrow Y_{\mathbf{a}}}(t)$ is the transition probability.

C.2.2 Derivation of R_0

The derivation of the basic reproductive number is done following the next generation matrix approach (See Section 1.2.2). As done for standard contact matrices, the first step is to focus only on the compartments that describe any stages of the infection: $E_{\mathbf{a}}$ and $I_{\mathbf{a}}$. It is convenient to re-write the differential equations for these compartments as

$$d_t \mathbf{x} = f(\mathbf{x}) - w(\mathbf{x}) \quad (\text{C.16})$$

where $\mathbf{x} \in [E_1, \dots, E_{\mathbf{T}}, I_1, \dots, I_{\mathbf{T}}]$, $f(\mathbf{x})$ encodes all terms that lead to new infections and $w(\mathbf{x})$ all other transitions in and out of the compartments. Furthermore, we denote with $\mathbf{a}$ ($a \in [1, T]$ and $T = K \prod_{p=1}^m V_p$) the index vectors. In other words, $\mathbf{1}$ describes the index vector $(1, 1, \dots, 1)$ of size $m+1$, $\mathbf{2}$ the index vector $(1, 1, \dots, 2)$, and $\mathbf{T}$ instead $(K, V_1, \dots, V_m)$. Using this notation, we map the generalized problem to the same structure described above for only one dimension (i.e., age). Indeed, we can write, once again dropping the time dependence for convenience,

$$d_t \mathbf{x} = \begin{bmatrix} d_t E_1 \\ \vdots \\ d_t E_{\mathbf{T}} \\ d_t I_1 \\ \vdots \\ d_t I_{\mathbf{T}} \end{bmatrix} = f(\mathbf{x}) - w(\mathbf{x}) = \begin{bmatrix} \Lambda_1 S_1 \\ \vdots \\ \Lambda_{\mathbf{T}} S_{\mathbf{T}} \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \Psi E_1 \\ \vdots \\ \Psi E_{\mathbf{T}} \\ \Gamma I_1 - \Psi E_1 \\ \vdots \\ \Gamma I_{\mathbf{T}} - \Psi E_{\mathbf{T}} \end{bmatrix} \quad (\text{C.17})$$

Which has exactly the same structure as the analogous described above, but where $i \rightarrow \mathbf{a}$ and $\mathbf{a} = (i, \alpha_1, \dots, \alpha_m)$. As shown above, the expression of R_0 is linked to the early epidemic dynamics which can be linearized by calculating the Jacobian at the disease-free equilibrium (DFE) $(S_{\mathbf{j}}^*, E_{\mathbf{j}}^*, I_{\mathbf{j}}^*, R_{\mathbf{j}}^*) = (N_{\mathbf{j}}, 0, 0, 0)$ for all age groups. Using the Jacobian matrix $\mathbf{J}$ (which has size $2K \prod_{p=1}^m V_p \times 2K \prod_{p=1}^m V_p$) at the DFE we can write

$$d_t \mathbf{x} = \mathbf{J} \mathbf{x} \quad (\text{C.18})$$

This expression can be conveniently factorized as $d_t \mathbf{x} = (\mathbf{F} - \mathbf{W}) \mathbf{x}$ where the matrix $\mathbf{F}$ is Jacobian applied to $f(\mathbf{x})$ and similarly the matrix $\mathbf{W}$ is the Jacobian applied to $w(\mathbf{x})$. In particular, we have:

$$\begin{aligned}
\mathbf{F} &= \begin{bmatrix} \partial_{E_1}(\Lambda_1 N_1) & \dots & \partial_{E_T}(\Lambda_1 N_1) & \partial_{I_1}(\Lambda_1 N_1) & \dots & \partial_{I_T}(\Lambda_1 N_1) \\ \partial_{E_1}(\Lambda_2 N_2) & \dots & \partial_{E_T}(\Lambda_2 N_2) & \partial_{I_1}(\Lambda_2 N_2) & \dots & \partial_{I_T}(\Lambda_2 N_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \partial_{E_1}(\Lambda_T N_T) & \dots & \partial_{E_T}(\Lambda_T N_T) & \partial_{I_1}(\Lambda_T N_T) & \dots & \partial_{I_T}(\Lambda_T N_T) \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \dots & 0 & \partial_{I_1}(\Lambda_1 N_1) & \dots & \partial_{I_T}(\Lambda_1 N_1) \\ 0 & \dots & 0 & \partial_{I_1}(\Lambda_2 N_2) & \dots & \partial_{I_T}(\Lambda_2 N_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \partial_{I_1}(\Lambda_T N_T) & \dots & \partial_{I_T}(\Lambda_T N_T) \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}
\end{aligned} \tag{C.19}$$

The components of each gradient are computed at the DFE, i.e., $S_j \rightarrow N_j$. By looking at Eq. 3.3 we note how each partial derivative in $\mathbf{j}$ selects the j -th index vector:

$$\mathbf{F} = \Phi \begin{bmatrix} 0 & \dots & 0 & \frac{G_{1,1}}{N_1} N_1 & \dots & \frac{G_{1,T}}{N_T} N_1 \\ 0 & \dots & 0 & \frac{G_{2,1}}{N_1} N_2 & \dots & \frac{G_{2,T}}{N_T} N_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \frac{G_{K,1}}{N_1} N_T & \dots & \frac{G_{T,T}}{N_T} N_T \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} = \Phi \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{C.20}$$

where we denoted with $\mathbf{0}$ $K \prod_{p=1}^m V_p \times K \prod_{p=1}^m V_p$ blocks of zeros and the generic entry of $\tilde{\mathbf{G}}$ is $\tilde{G}_{\mathbf{a},\mathbf{b}} = \frac{G_{\mathbf{a},\mathbf{b}}}{N_{\mathbf{b}}} N_{\mathbf{a}}$. Using the same method we can easily write the expression of the matrix $\mathbf{V}_1$ as:

$$\mathbf{W} = \begin{bmatrix} \Psi & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \Psi & 0 & \dots & 0 \\ -\Psi & \dots & 0 & \Gamma & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & -\Psi & 0 & \dots & \Gamma \end{bmatrix} = \begin{bmatrix} \Psi \mathbf{I} & \mathbf{0} \\ -\Psi \mathbf{I} & \Gamma \mathbf{I} \end{bmatrix} \tag{C.21}$$

where $\mathbf{I}$ are diagonal matrices of size $K \prod_{p=1}^m V_p \times K \prod_{p=1}^m V_p$. The expression for R_0 is linked to the two matrices as $R_0 = \rho(\mathbf{F}\mathbf{W}^{-1})$, where ρ denotes the spectral radius. It is

easy to show how

$$\mathbf{F}\mathbf{W}^{-1} = \Phi \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \frac{1}{\Psi}\mathbf{I} & \mathbf{0} \\ \frac{1}{\Gamma}\mathbf{I} & \frac{1}{\Gamma}\mathbf{I} \end{bmatrix} = \frac{\Phi}{\Gamma} \begin{bmatrix} \tilde{\mathbf{G}} & \tilde{\mathbf{G}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (\text{C.22})$$

hence, we can finally write

$$R_0 = \frac{\Phi}{\Gamma} \rho(\tilde{\mathbf{G}}) \quad (\text{C.23})$$

The expression has the same general form as the one derived for standard contact matrices stratified just for age. However, the entries of the matrix are different and those encode the multiple dimensions.

C.3 Other contacts data

Beyond the data on Hungary, we use data on Zimbabwe age-contact patterns [78]. Also, in this case, the contact numbers among age classes have been collected through surveys.

C.3.1 Age-contact matrices

Although contacts (i.e., raw contacts), theoretically, are expected to be symmetrical, due to survey limitations and reporting errors they do not achieve perfect reciprocity. Given that our model to build synthetic generalized contact matrices respects the symmetric assumption on the raw contacts, we needed to ensure that also the matrix $\mathbf{R}$ that we fed to these models respects this assumption as well. Thus, all the raw contact matrices used in this study have been corrected for reciprocity *a priori* according to equation (1.1).

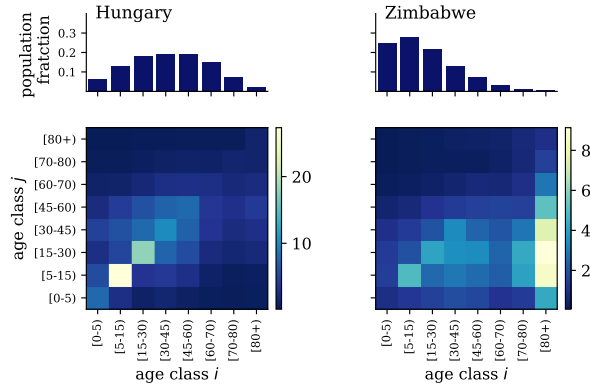


Figure C.1: (1st row) Population distribution by age (2nd row) Age-contact matrices referred to the pre-pandemic period for Hungary and Zimbabwe.

C.4 Parameters and implementation

To ensure the reproducibility of our work, in this section we include a comprehensive list of all the parameters employed to generate the synthetic and empirical generalized contact matrices used for the analysis presented in the main text. In addition, we provide all the technical details about the implementation of the analysis.

Figure 3.2

In Table C.1 we report the list of parameters used to generate the generalized contact matrices for the analysis presented in Figure 3.2 in the main text. For the random mixing scenario the only set of parameters of interest is the population distribution. Indeed, in these settings, no extra parameters are needed. For the assortative regime we also report the values of the activity and assortativity. As described above, for $V_1 = 3$, in general, we need to fix $W = V_1^2 - 1 = 8$ free parameters. $2V_1 - 1 = 5$ are defined by the assumption of our model (i.e., by setting P_α and q_α values). Hence, three extra values are required to fully define the generalized contact matrix. We set these values as $a_1 = 0.6$ (correspondent to $p_{i1,j2} = 0.048$ in Eq. C.12), $a_2 = 0.6$ (correspondent to $p_{i2,j1} = 0.12$ in Eq. C.12), and $a_3 = 0.5$ (correspondent to $p_{i3,j1} = 0.07$ in Eq. C.12) respectively for the first, second and third SES group. We refer the reader to Section C.1 for the details about the synthetic generalized contact matrices construction.

SES Group	1	2	3
Population distribution	35%	45%	20%
Assortativity (q_i)	60%	50%	65%
Activity (P_i)	20%	40%	40%
Free parameters (a_i)	0.6	0.6	0.5

Table C.1: List of parameters used to generate Figure 2 in the main text.

Figure 3.3

To demonstrate the flexibility of the proposed approach to model NPIs in Figure 3 we analyze three different scenarios: (i) Baseline, (ii) NPI 1 and (iii) NPI 2. In particular, in the scenario of NPI 1, we assume that all the individuals reduce their contacts by 35% at $t^* = 50$, while in the scenario of NPI 2, we assume that the introduction of NPI also affects the structure of the matrix. Thus, to model these scenarios, we first reduce by 35% the total number of contacts across age groups such that $R_{ij}^{NPI} = 0.65R_{ij}$. Then, we feed R_{ij}^{NPI} to our model to generate synthetic generalized contact matrices. For the first scenario (NPI 1), the parameters of the model are equal to one of the baseline scenarios, while the parameters of the second scenario (NPI 2) change according to the assumption that we made on the new structure of contacts. In Table C.2 we report the parameters used to generate the generalized contact matrices for this analysis. As for the previous figure, three additional free parameters are needed to define the generalized contact matrix. These are, for the baseline, $a_1 = 0.5$, $a_2 = 0.5$, $a_3 = 0.5$ and $a_1 = 0.5$, $a_2 = 0.4$, $a_3 = 0.4$ for NPI2.

Figure 3.4

We built generalized contact matrices stratified in two dimensions by using real data from the MASZK study.

Information on social interactions has been collected in two different ways i) in an aggregate form, such that each participant reported the number of contacts they had with individuals in each of the eight age brackets considered, ii) in a diary in which each participant listed one by one the interactions they had on a given day by reporting some meta information of the *contactee* such as their age and SES (See Section 1.3 and B.1

	Baseline			NPI 2		
SES Group	1	2	3	1	2	3
Population distribution	33,3%	33,3%	33,3%	33,3%	33,3%	33,3%
Assortativity (q_i)	33,3%	50%	60%	50%	60%	70%
Activity (P_i)	25%	45%	30%	37%	37%	26%
Free parameters (a_i)	0.5	0.5	0.5	0.5	0.4	0.4

Table C.2: List of parameters used to generate Figure 3 in the main text, for (i) baseline and, (ii) NPI 2 scenarios.

for more details on the data collection). In particular, the average number of contacts of an individual in age class i , and SES α with an individual in age class j , and SES β is $G_{\mathbf{a},\mathbf{b}}$ where $\mathbf{a} = (i, \alpha)$ and $\mathbf{b} = (j, \beta)$. However, the *MASZK* data provided us with diary information only for the adult population (individuals older than 15 years old).

For what concerns children, the *MASZK* study provided us with information on the participants' SES, but their average number of contacts is available only in the aggregate form. Thus, from the data, we could compute only $G_{i\alpha,j}$ which is the average number of contacts that an individual of age group i and SES α has with all the individuals of age group j . To disaggregate the average number of contacts they had with each SES group we introduced some assumptions. Namely, we assumed assortative mixing along SES. We assigned the average number of contactees to the different SES as follows: $G_{i\alpha,j\beta} = G_{i\alpha,j}u_{\alpha\beta}$ where $u_{\alpha\beta}$ is a parameter that controls how these contacts are distributed among individuals of different SES. To mimic assortativity patterns among SES, the values of $u_{\alpha\beta}$ are set as follows:

$$u_{\alpha\beta} = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{pmatrix} \quad (\text{C.24})$$

We note higher values in the diagonal, meaning that 70% of children's contacts are with children belonging to the same SES. The average number of contacts of these matrices includes the contacts of the family members. However, contacts have been added afterwards to the matrix by assuming that individuals belonging to the same family have the same SES.

C.5 Impact of generalized contact matrices on epidemic modeling

To provide evidence of the robustness of the proposed approach we analyze contact data from a second country. In particular, we consider the age-structured contact matrix $\mathbf{C}$ of Zimbabwe [78] (Fig. C.2-a). We note how $\mathbf{C}$ is different from the Hungarian one. Indeed, although the age assortative and the inter-generational interaction are still clear, the matrix shows a high activity rate for individuals aged 80+. By applying the same analysis that we did for the Hungarian data we study the impact of generalized contact matrices on epidemic modeling in (i) the random mixing (Fig. C.2-b) and, (ii) the assortative mixing (Fig. C.2-e) regimes. As in the main text, we imagine a simple case where the second dimension contains three groups: 1, 2, 3. We assume that 35%, 45%, and 20% of the population belong to these three categories across all age groups. For the

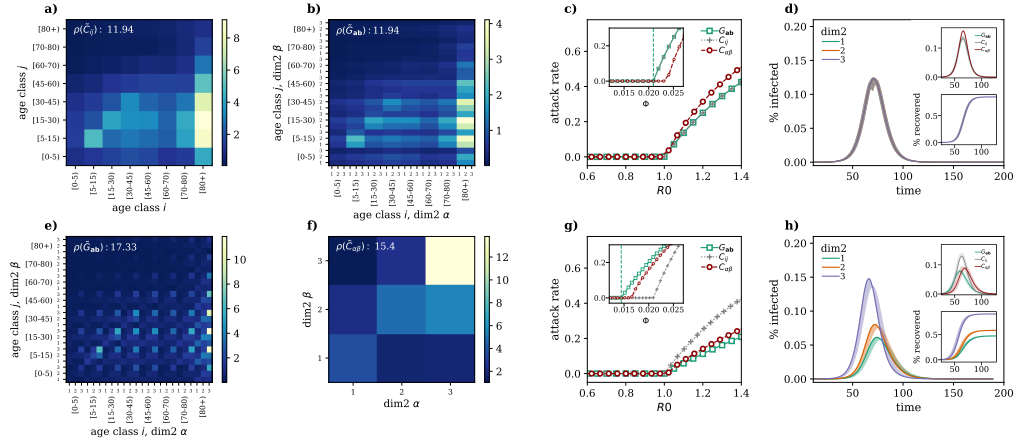


Figure C.2: Impact of generalized contact matrices on epidemic spreading. (a) Example of a classic age contact matrix C_{ij} (8×8); (b) and (e) generalized contact matrices with an additional dimension with three levels G_{ab} (24×24) in the case of random mixing among the second level groups (b) and in the case of assortative mixing and different activity levels among the second dimension (e); (f) depicts the case when we integrate the generalized contact matrix over all age classes; (c) attack rate as a function of R_0 (main figure) and disease transmissibility (inset). In panel (d) we show, in the case of random mixing, the prevalence (main figure and top inset) and the fraction of recovered as a function of time (bottom inset); (g) and (h) are as the previous two plots but for the case of assortative mixing with different levels of activity. Results refer to the median of 500 runs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, and $R_0 = 2.7$. Simulations start with a number $I_0 = 100$ of initial infectious seeds.

assortative mixing scenario, we assume that 60%, 50%, and 65% of the contacts in the first, second, and third groups of the additional category take place within each group. Additionally, we assume some levels of heterogeneity also in the activity of the different groups, setting to 20%, 40%, and 40% the share of contacts of the three groups. In Fig. C.2-f we show the matrix $C_{\alpha\beta}$ obtained by integrating the latter generalized contact matrix over all age classes.

We use these matrices to study the unfolding of a virus in the population. In Fig. C.2-c we plot the attack rate (i.e., epidemic size) as a function of R_0 and as a function of the transmissibility parameter Φ (inset) for 1) a model fed only with the contact matrix C_{ij} (grey crosses), 2) a model fed with the generalized contact matrix $G_{a,b}$ (green squares), 3) a model fed with the matrix $C_{\alpha\beta}$ where contacts are stratified only according to the second dimension (red circles). In Fig. C.2-d we show the prevalence of infected overtime of the three dimensions as predicted by the model fed with $G_{a,b}$ for the three groups. In the top inset, we show the overall prevalence of infected predicted by three models. In the bottom inset, we plot the fraction of recovered as a function of time for the three groups in the second dimension. We perform the same analysis for the assortative scenario and we show the results respectively in Fig. C.2-g and Fig. C.2-h. All the results are in perfect agreement with the ones discussed in the main text.

C.6 Robustness analysis

C.6.1 Sensitivity analysis on the spectral radius

In this section, we perform a sensitivity analysis of the model to build synthetic generalized contact matrices. In particular, we show how the spectral radius of G_{ab} differs

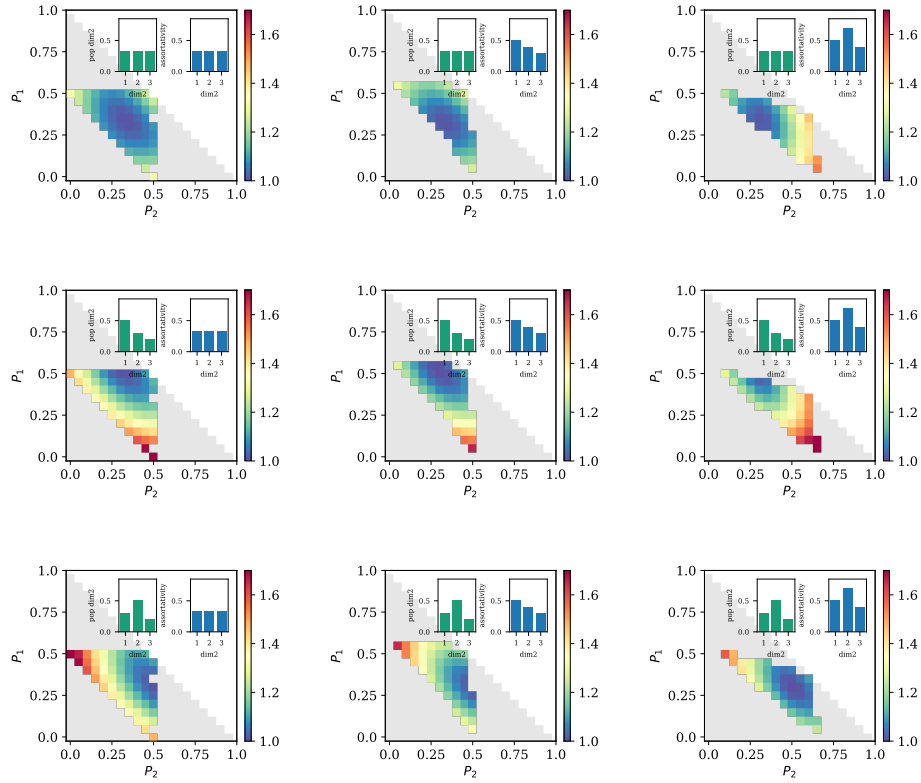


Figure C.3: Sensitivity analysis. Variation of the ratio of the $\rho(G_{ab})$ over $\rho(C_{ij})$ as a function of the activities P_1, P_2 . Given that $\sum_1^3 P_i = 1$ for each P_1 and P_2 , $P_3 = 1 - P_1 - P_2$. In each subplot G_{ab} is computed according to our synthetic model using different parameters distribution of population distribution of the second dimension (left inset) and assortativity along the second dimension (right inset). The grey area refers to the non-physical solutions of the model.

from the one computed on the aggregate age contact C_{ij} when we vary some of the parameters of the model.

We create nine different scenarios one for each combination of three sets of values for the population distribution of the second dimension and three sets of parameters for the assortativity. For each of these scenarios in Fig. C.3 we show the ratio of the $\rho(\tilde{G}_{ab})$ over $\rho(\tilde{C}_{ij})$ as function of the activity of individuals in dim1 (P_1) and, of individuals in dim2 (P_2). Given that $\sum_1^3 P_i = 1$ for each P_1 and P_2 , $P_3 = 1 - P_1 - P_2$, thus each pair of values of P_1, P_2 defines also the value of P_3 .

As shown above, $\rho(\tilde{G}_{ab})$ is always greater or equal to $\rho(\tilde{C}_{ij})$ as expected from the theory. From Fig. C.3 we can observe how the parameter sets of the different scenarios influence the magnitude of this relation, suggesting that different interaction patterns arising from correlation to an additional dimension effectively impact the estimation of the basic reproductive number.

C.6.2 Testing the model with three dimensions

To show the validity of our approach beyond the two dimensions here we present the results of our analysis when we stratify the matrix along three dimensions (age, dim2, and dim3). We assume that also the third dimension has three possible groups. Starting from a generalized contact matrix with two dimensions (Fig. C.4-b) we build a generalized contact matrix with the assumption of random mixing among the third dimension (Fig. C.4-f) and assortative assumption along the third dimension (Fig. C.4-e). We assume the same parameters as for the assortative regime of Fig. 2 in the main text to build the 2D matrix. For the third dimension, we assume that individuals are respectively 30%, 30% and 40% in the three dimensions. Additionally, we assume that 34%, 33% and 46% represent the respective proportion of assortative contacts for the three groups and that 40%, 40% and 20% indicate the relative activity of the three groups. Also in this case, we use these matrices to study the unfolding of a virus in the population. We can observe that, in the assortative mixing scenario, the spectral radius of the $G_{a,b}$ with three dimensions (Fig. C.4-e) is higher than the one of $G_{a,b}$ with two dimensions (Fig. C.4-b). This implies that, given $R_0 = 1$, the critical value of Φ predicted by the model with three dimensions is going to be smaller than the one predicted by the model with two dimensions.

In Fig. C.4-c we plot the attack rate (i.e., epidemic size) as a function of R_0 and as a function of the transmissibility parameter Φ (inset) for 1) a model fed only with the contact matrix C_{ij} (grey crosses), 2) a model fed with the generalized contact matrix $G_{a,b}$ with three dimensions (purple triangles). In Fig C.4-d we show the prevalence of infected overtime of the three dimensions as predicted by the model fed with $G_{a,b}$ with three dimensions for the three groups of $dim3$. In the top inset, we show the overall prevalence of infection predicted by the two models. In the bottom inset instead, we plot the fraction of recovered as a function of time for the groups in the third dimension. We perform the same analysis for the assortative scenario and we show the results respectively in Fig C.4-g and Fig C.4-h. All the results are in perfect agreement with the ones discussed in the main text.

C.7 Empirical generalized contact data

The MASZK questionnaire provided us with three diaries that we were able to use to generate the generalized contact matrices stratifying by age and SES of participants. Here, we show the results of the same analysis shown in the main text for two additional periods: April (Fig. C.5) and November (Fig. C.6) 2021. While the overall results confirm the one presented in the main text, models fed with standard and generalized contact matrices lead to estimations of the attack rates, for a given R_0 , which are closer to each other. Nevertheless, the model featuring generalized contact matrices allows capturing heterogeneities in the incidence across SES groups which are invisible to standard approaches.

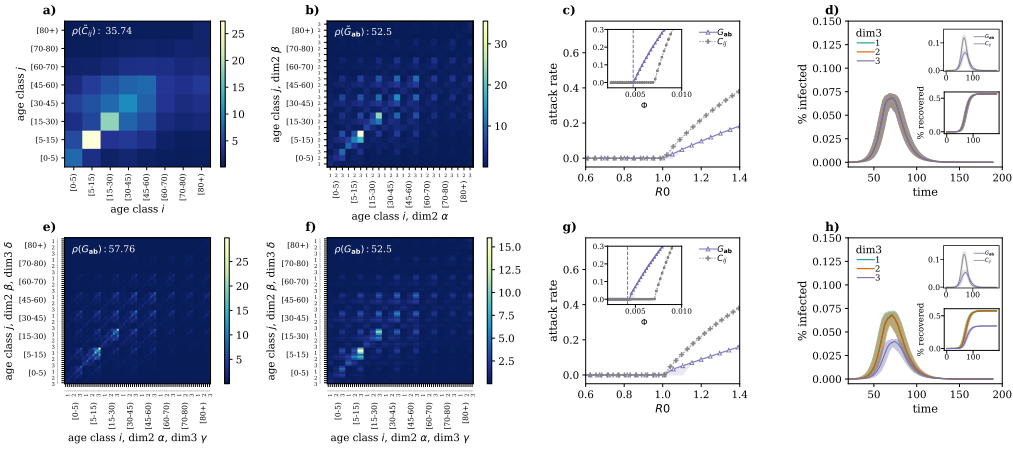


Figure C.4: Three dimensions generalized contact matrices. (a) Example of a classic age contact matrix C_{ij} (8×8); (b) Generalized contact matrices with an additional dimension with two levels G_{ab} (24×24) in the case of assortative mixing and different activity levels among the second dimension (e) and (f) Generalized contact matrices with three levels G_{ab} (72×72) (e) in the case of assortative mixing and different activity levels among the third dimension and (f) in the case of random mixing among the third dimension (e); (c) attack rate as a function of R_0 (main figure) and disease transmissibility (inset). In panel (d) we show, in the case of random mixing, the prevalence (main figure and top inset) and the fraction of recovered as function of time (bottom inset); (g) and (h) are as the previous two plots but for the case of assortative mixing with different levels of activity. Results refer to the median of 500 runs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, and $R_0 = 2.7$. Simulations start with a number $I_0 = 100$ of initial infectious seeds.

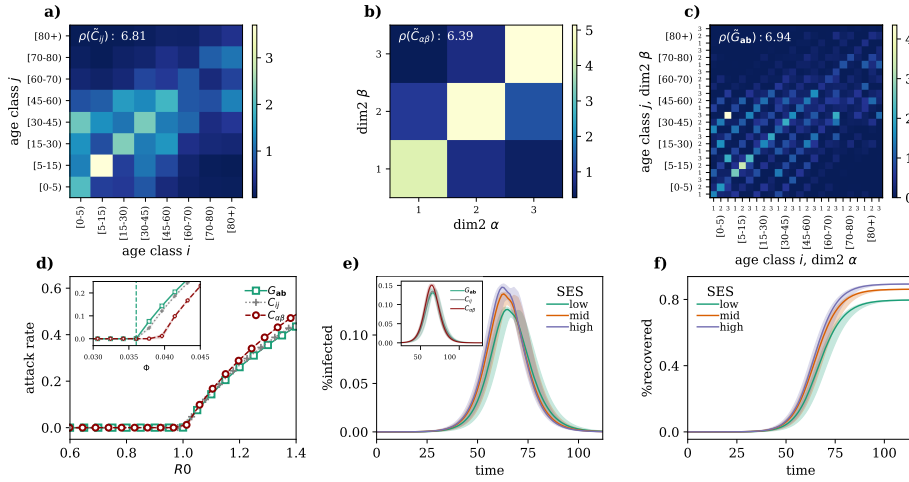


Figure C.5: Empirical generalized contact matrices. (a) Age contact matrix C_{ij} (8×8); (b) Socio-economic status (SES) contact matrix $C_{\alpha\beta}$ (3×3) and (c) Generalized contact matrix with age and socio-economic status G_{ab} (24×24); (d) attack rate as a function of R_0 , (e) disease prevalence as a function of time, and (f) fraction of recovered as function of time. Results refer to the median of 500 runs. Epidemiological parameters: $\mu = 0.25$, $\epsilon = 0.4$, $R_0 = 2.7$, $I_0 = 100$. The matrices have been computed using the contact diaries coming from the MASZK data collected in Hungary during April 2021.

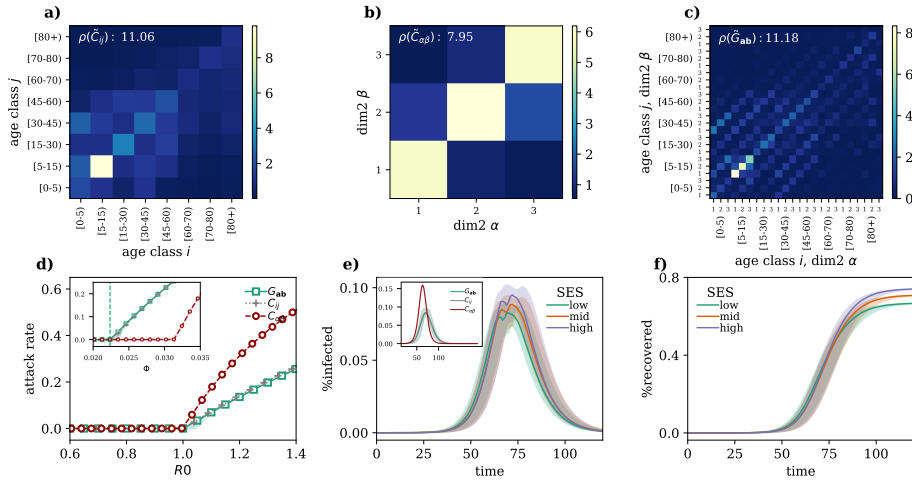


Figure C.6: Empirical generalized contact matrices(a) Age contact matrix C_{ij} (8×8); (b) Socio-economic status (SES) contact matrix $C_{\alpha\beta}$ (3×3) and (c) Generalized contact matrix with age and socio-economic status G_{ab} (24×24); (d) attack rate as a function of R_0 (e) disease prevalence as a function of time and (f) fraction of recovered as function of time. Results refer to the median of 500 runs. Epidemiological parameters: $\mu = 0.25$, $\epsilon = 0.4$, $R_0 = 2.7$, $I_0 = 100$. The matrices have been computed using the contact diaries from the MASZK data collected in Hungary during November 2021.

Appendix D

Appendix chapter 4

D.1 Vaccination data

Data on the number of vaccines administered have been sourced from Our World in Data [182]. Figures D.1 presents the number of vaccines administered daily by age for Hungary.

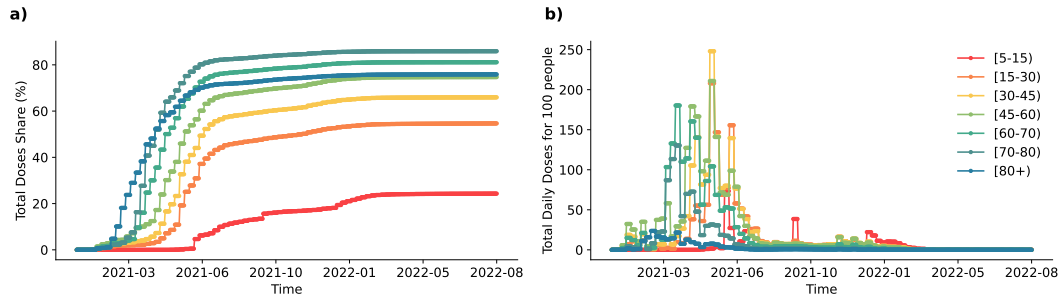


Figure D.1: Hungary COVID-19 vaccination **a.** Share of people who completed the initial COVID-19 vaccination protocol by age **b.** Total Daily Dosed for 100 people by age.

D.2 Computing attack rates

The attack rate is defined as the fraction of all individuals in a susceptible population who have been infected before a specific time t . In a SEIR model, the attack rate at time t can be computed as follows:

$$AR(t) = \frac{M_{S \rightarrow E}(t)}{N} \quad (\text{D.1})$$

where $M_{(S \rightarrow E)}$ indicates the number of transitions from the susceptible (S) to the exposed (E) state up to time t , and N represents the total number of individuals in the population.

However, when vaccination is introduced, the computation of the attack rate for *vaccinated* and *non-vaccinated* individuals separately becomes more complex. In particular, when vaccination is administered during the course of the epidemic, not only are susceptible individuals vaccinated but also those who have been exposed (i.e., already infected but not yet infectious) and those who have recovered. In epidemiological modeling, this

means that over the course of the epidemic, there are continuous movements between $S_{NV} \rightarrow S_V$ but also $E_{NV} \rightarrow E_V$ and $R_{NV} \rightarrow R_V$.

When computing the attack rate for *vaccinated* and *non-vaccinated* individuals separately, it is important to consider these transitions and account for the fact that vaccination is also administered to individuals in the E_{NV} and R_{NV} compartments. Therefore, the denominator of the attack rate is not simply the total number of vaccinated and non-vaccinated individuals at time t , but it should also account for these transitions.

Non-vaccinated population. For the non-vaccinated population, the denominator should include not only the total number of individuals not yet vaccinated at time t but also all the transitions from $E_{NV} \rightarrow E_V$ and $R_{NV} \rightarrow R_V$ up to time t . This is because these individuals, although vaccinated later, were not vaccinated when they were exposed or recovered. Thus, the adjusted denominator is:

$$N_{NV}(t) + M_{E_{NV} \rightarrow E_V}(t) + M_{R_{NV} \rightarrow R_V}(t) \quad (D.2)$$

Where $N_{NV}(t)$ is the total number of non-vaccinated individuals at time t , and $M_{E_{NV} \rightarrow E_V}(t)$ and $M_{R_{NV} \rightarrow R_V}$ indicate the number of individuals vaccinated up to time t in the exposed and recovered compartments, respectively. As $N_{NV}(t)$ can also be expressed as the difference between the total population and the vaccinated individuals, $N - N_V(t)$, and $N_V(t)$ can be expressed as the sum of all the vaccinated individuals, ie $M_{S_{NV} \rightarrow S_V}(t) + M_{E_{NV} \rightarrow E_V}(t) + M_{R_{NV} \rightarrow R_V}(t)$, we can rewrite Equation (D.2) as the difference between the total population and the number of individuals vaccinated in the susceptible compartment, the attach rate can be expressed as follow.

$$AR_{NV} = \frac{M_{S_{NV} \rightarrow E_{NV}}(t)}{N - M_{S_{NV} \rightarrow S_V}(t)} \quad (D.3)$$

Vaccinated population. Symmetrically, the denominator of the attack rate for vaccinated individuals should account for the fact that individuals moving from E_{NV} to E_V and R_{NV} to R_V were not vaccinated when they were initially infected. Therefore, these individuals should be removed from the count of the total number of vaccinated individuals at time t . Thus, the denominator becomes:

$$N_V(t) - M_{E_{NV} \rightarrow E_V}(t) - M_{R_{NV} \rightarrow R_V}(t) \quad (D.4)$$

By substituting $N_V(t)$ in Equation (D.4), we can compute the attach rate as follow:

$$AR_V(t) = \frac{M_{S_V \rightarrow E_V}(t)}{M_{S_{NV} \rightarrow S_V}(t)} \quad (D.5)$$

Entire population. Finally, the overall attack rate for the entire population can be computed as follows:

$$AR(t) = \frac{M_{S_{NV} \rightarrow E_{NV}}(t) + M_{S_V \rightarrow E_V}(t)}{N} \quad (D.6)$$

D.3 Synthetic population and mixing patterns

The age contact matrix and population distribution used are based on data for Hungary, as illustrated in Figure C.1. The model used to build the synthetic population and their mixing pattern (Figures 1 – 4) is the same as in Chapter 3. Here, we report the values of the activity, assortativity and the free parameters (see Table C.1).

SES Group	1	2	3
Population distribution	35%	45%	20%
Assortativity	60%	50%	65%
Activity	20%	40%	40%
Free parameters	0.6	0.6	0.5

Table D.1: List of parameters used to generate Figure 2 in the main text of Chapter 4.

D.4 Robustness analysis for Figures 1 and 3

In this section, we present a series of alternative simulation settings for Figures 1 and 3 of Chapter 4 to assess the robustness of the results. By exploring different assumptions and parameter variations, we aim to ensure that the findings remain consistent across a range of plausible scenarios. These robustness checks provide additional confidence in the conclusions drawn from the main simulations.

D.4.1 Vaccination stratified only by SES

Here, we simulate vaccination distributions that do not account for age, by only applying the vaccination distributions $VD1$, $VD2$, and $VD3$ across three socioeconomic status (SES) groups. This approach is applied to both Scenario 1 and Scenario 2, as shown in Figures D.2 and D.3, respectively.

In these simulations, the variations among the three vaccination distributions (VDs) are more pronounced in terms of overall prevalence across the population, as observed in panel *a* of Figures D.2 and D.3. Notably, in Scenario 2, the effectiveness of $VD2$ and $VD3$ increases substantially, nearly eliminating the epidemic wave.

Furthermore, differences among SES subgroups' attack rate become even more evident, as shown in panel *b* of Figures D.2 and D.3.

Finally, panels *c* of Figures D.2 and D.3 demonstrate that attack rates for specific subgroups are not accurately predicted by the age-stratified model. Indeed, the discrepancy between the two methodologies, $G_{ab} - f(C_{ij})$, is almost always non-zero.

Overall, these simulation results align with the findings discussed in Chapter 4, supporting the conclusions on the influence of vaccination distribution patterns on epidemic outcomes.

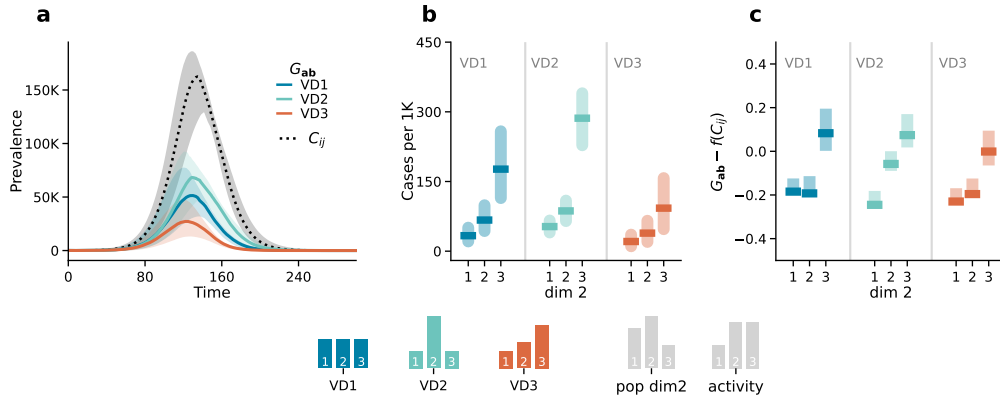


Figure D.2: Scenario 1. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1, VD2, VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25, \Psi = 0.4, g_1 = 0.6, R_0 = 1.6$. Simulations start with $I_0 = 100$ initial infectious seeds.

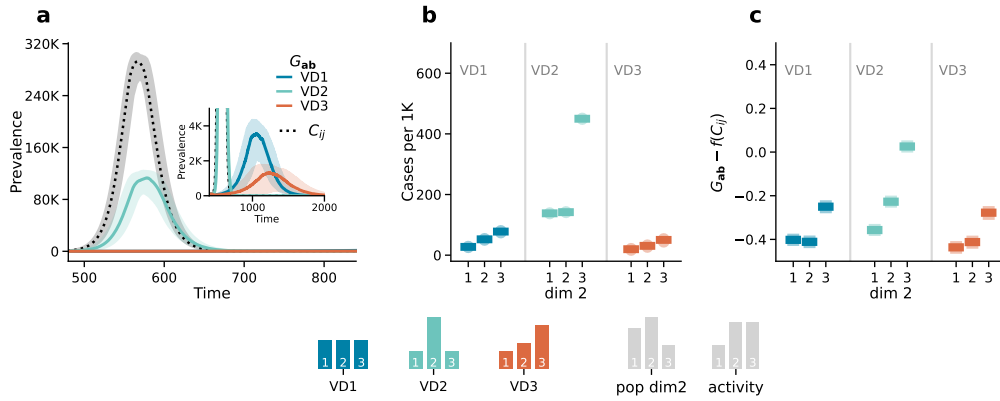


Figure D.3: Scenario 2. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1, VD2, VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25, \Psi = 0.4, g_1 = 0.6, R_0 = 3$. Simulations start with $I_0 = 100$ initial infectious seeds.

Homogeneous mixing

Keeping the same vaccination distributions, we run the same analysis as present in Figure 1 (Scenario 1) and Figure 3 (Scenario 2) of the main text on a simpler population mixing setting. Furthermore, we consider vaccination only by SES as in the previous section. In this case, we assume homogeneous (i.e., random) mixing where contacts along the second dimension are set proportional to the product of the population sizes in each group. The only set of parameters of interest is the population distribution, which is set to $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$. Indeed, in these settings, no extra parameters are needed. Again the model is the same as in Chapter 3.

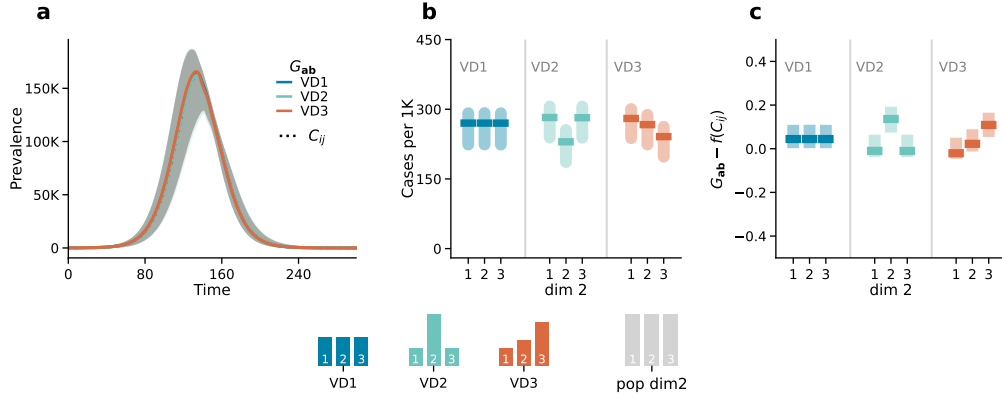


Figure D.4: Scenario 1. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1, VD2, VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25, \Psi = 0.4, g_1 = 0.6, R_0 = 1.6$. Simulations start with $I_0 = 100$ initial infectious seeds.

Scenario 1. The first observation is that, as expected from Ref. [150], homogeneous mixing leads to the same epidemic dynamics (i.e., same prevalence) across subgroups (see panels D.4a). Indeed, regardless of the vaccination distribution the coloured lines overlap. Furthermore, random mixing leads to the same total prevalence regardless of the inclusion of a second dimension in the contact matrices. Indeed, solid coloured and dashed grey lines representing respectively the prevalence estimated by modes fed with generalised or traditional contact matrices overlap.

In panels *b* of Figure D.4 we show the attack rate per one thousand (i.e., number of cases for 1000 individuals) for each subgroup, for each vaccination distribution (VDs). In a random mixing situation, differences in disease burden among subgroups arise solely from variations in vaccination uptake. When vaccination uptake is equal across all groups ($VD1$), infection rates are the same in all three SES groups. Under vaccination distribution $VD2$, the second SES group shows the lowest infection rate due to higher vaccination uptake. In the more unequal distribution of vaccines ($VD3$), the number of cases in each group inversely correlates with their respective vaccination uptake: the first group, having the lowest vaccination rate, becomes the most infected, while the

third group, with the highest vaccination uptake, experiences the fewest cases.

Finally, in panel D.4c we show the difference of the number of cases in each subgroup from the generalized model G_{ab} and as estimated from a simple age-stratified model $f(C_{ij})$. In this scenario, the differences among the two techniques, although small, are almost always different from zero.

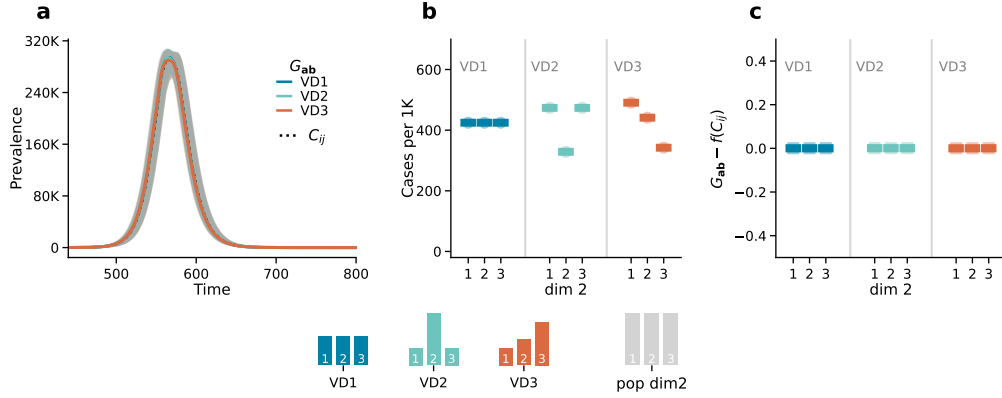


Figure D.5: Scenario 2. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1$, $VD2$, $VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, $R_0 = 3$. Simulations start with $I_0 = 100$ initial infectious seeds.

Scenario 2. Here, we report the results of the same simulation setting for Scenario 2, where the vaccination campaign rollout happened before the epidemic outbreak.

Panel *a* of Fig. D.5 shows that similar to Scenario 1, the coloured lines overlap regardless of the vaccination distribution, indicating consistent epidemic outcomes across different distribution strategies. Additionally, these coloured lines match the dashed grey lines, which represent outcomes based on traditional age contact matrices, highlighting that the introduction of an additional dimension does not alter the overall results, as also discussed in Chapter 3.

Panel D.5b shows the attack rate per one thousand (i.e., number of cases for 1000 individuals) for each subgroup, for each vaccination distribution (VDs). As before, in a random mixing situation, differences in disease burden among subgroups arise solely from variations in vaccination uptake. Thus the qualitative differences among attack rates in subgroups follow the ones seen in Scenario 1.

Panel D.5c shows the difference of the number of cases in each subgroup from the generalized model G_{ab} and as estimated from a simple age-stratified model $f(C_{ij})$. Interestingly, in this case, the difference is always zero, meaning that, in this case, it is possible to accurately predict the epidemic burden in the three additional subgroups starting from the output from age-stratified models. Indeed, in this scenario, since the vaccination is already completed before the epidemic begins, there is not a dynamic interaction between the vaccination uptake and the epidemic spreading, thus the attack

rate in the different subgroups is fully captured by the overall number of individuals in each subgroup, that were vaccinated before the epidemic started.

Sensitivity to varying t_{epi} and R_0

Finally, we run a sensitivity analysis for the difference between the predicted number of cases in each subgroup from the generalized model and as estimated from a simple age-stratified model ($G_{ab} - f(C_{ij})$). Specifically, we examine the differences between the overall attack rate for each subgroup as predicted by the generalized model G_{ab} and the attack rate predicted by the outputs of the age-stratified model aggregate at age level, as described in Equation (4.5) $f(C_{ij})$. In Figure D.6, we show how this difference changes (i) as a function of t_{epi} , the time at which the epidemic starts given that vaccination began at $t = 0$ with $R_0 = 2.5$, and (ii) as a function of R_0 , the reproduction number, with $t_{epi} = 0$. Again the vaccination is kept to be stratified only by SES. These analyses are done both under a homogeneous mixing setting (panel *a*) and an assortative mixing setting (panel *b*). Both the parameters t_{epi} and R_0 determine the overlap between the vaccination campaign and the epidemiological wave: with high t_{epi} , vaccination might be completed before the epidemic starts, whereas with high R_0 , the epidemic might peak early during the vaccination campaign, diminishing the vaccination's impact on the epidemic curve. The results indicate that only under random mixing, where vaccination is completed before the epidemic begins, can we accurately predict the attack rate within each subgroup based on the output of an age-stratified model (panel *a*, first column). Specifically, the difference ($G_{ab} - f(C_{ij})$) converges to zero as t_{epi} increases, indicating that in a fully vaccinated population with no interaction between epidemic and vaccination dynamics, the attack rate by subgroup aligns with predictions from the age-stratified model. However, as the overlap between the vaccination campaign and the epidemic increases, so does the error, as shown in panel *a*, first column, for low values of t_{epi} and in the second column for higher values of R_0 . In contrast, under assortative mixing, the interaction between differences in individual activity levels and vaccination uptake prevents accurate predictions (panel *b*).

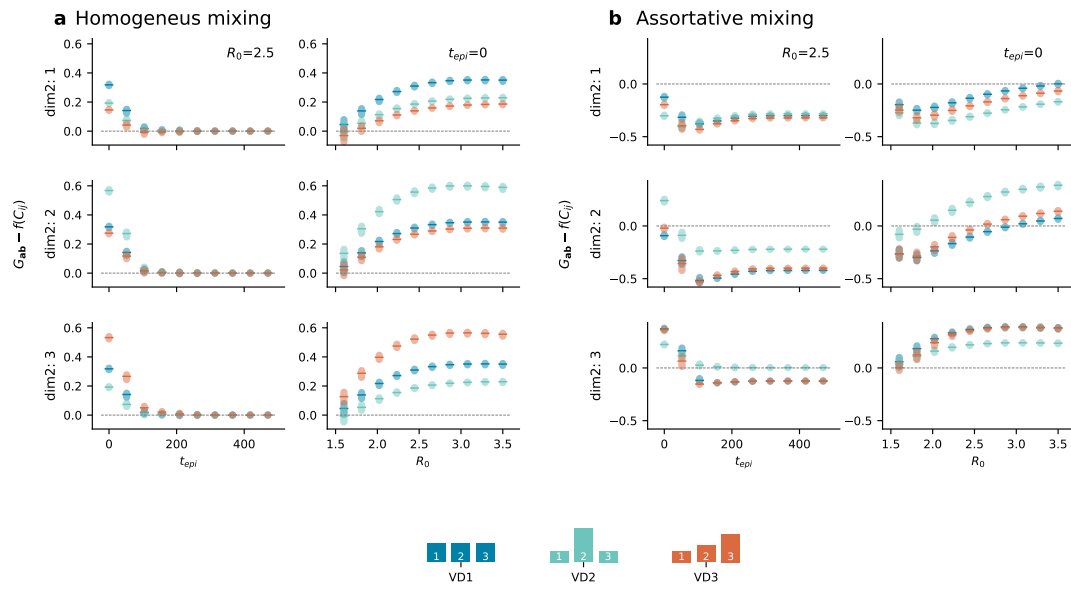


Figure D.6: Sensitivity analysis Differences between the overall attack rate for each subgroup as predicted by the generalized model G_{ab} and the attack rate predicted by the aggregate output of the age-stratified model, $f(C_{ij})$ as a function of (i) t_{epi} with $R_0 = 2.5$, and (ii) R_0 with $t_{epi} = 0$. These analyses are done both under a random mixing setting (panel a) and an assortative mixing setting (panel b)

D.4.2 Vaccination stratified by age and SES

Here, we explore a simulation setting with homogenous mixing while maintaining the vaccination modelling framework presented in the main analysis. Specifically, the distribution of vaccines follows a two-step approach: first by age, based on daily administration data, and then by socioeconomic status (SES), using three distinct vaccination uptake distributions ($VD1$, $VD2$, $VD3$).

Homogeneous mixing

We run the same analysis as presented in Figure 1 (Scenario 1) and Figure 3 (Scenario 2) of the main text on a simpler population mixing setting. As before, in the case of homogeneous mixing contacts along the second dimension are set proportional to the product of the population sizes in each group. Again the model is the same as in Chapter 3.

The results for Scenario 1 and Scenario 2 are shown in Figure D.7 and Figure D.8, respectively. These findings are consistent with the patterns discussed in the previous section, further validating the dynamics observed when varying vaccination strategy and mixing assumptions.

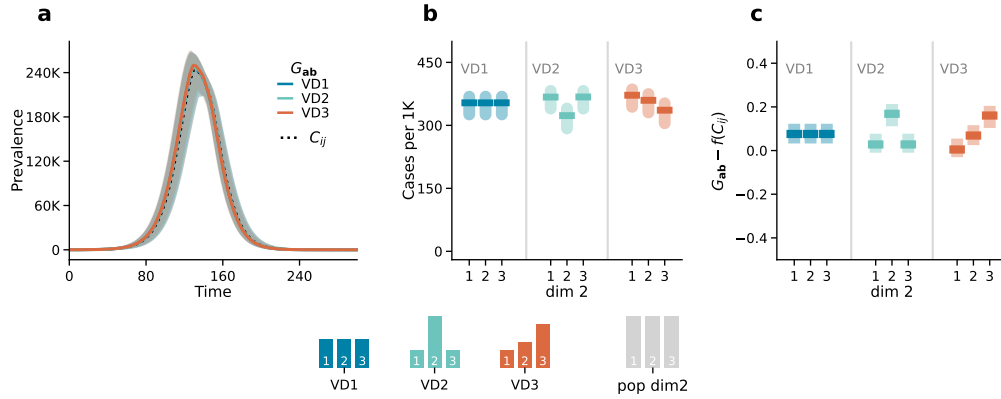


Figure D.7: Scenario 1. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1$, $VD2$, $VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, $R_0 = 1.6$. Simulations start with $I_0 = 100$ initial infectious seeds.

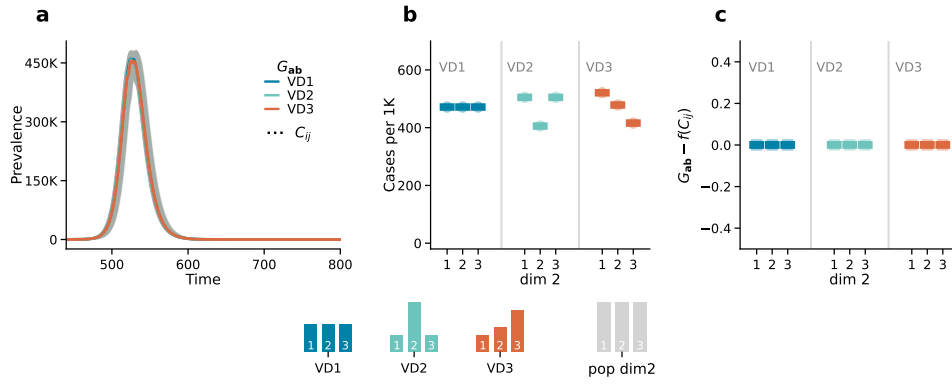


Figure D.8: Scenario 2. Epidemic outcomes and predictability: Panel *a* displays the number of newly infected individuals over time. The dotted black line corresponds to the outcome of the age-stratified model (C_{ij}), and the solid coloured lines correspond to the outcome of the generalized model (G_{ab}) in the three different vaccination strategies ($VD1, VD2, VD3$). Panel *b* shows the attack rate by the second dimension (dim 2) predicted by the generalized model (G_{ab}). Panel *c* shows the difference between the attack rate predicted by the generalized model G_{ab} and the one estimated from the aggregate output of the age-stratified model $f(C_{ij})$ by the second dimension (dim 2). Results refer to the median of 500 runs with IQRs. Epidemiological parameters: $\Gamma = 0.25, \Psi = 0.4, g_1 = 0.6, R_0 = 1.6$. Simulations start with $I_0 = 100$ initial infectious seeds.

Sensitivity to varying t_{epi} and R_0

We conducted the same sensitivity analysis for the difference between the predicted number of cases in each subgroup from the generalized model and as estimated from a simple age-stratified model ($G_{ab} - f(C_{ij})$), with results shown in Figure D.9. As before, we examine both a homogeneous mixing setting (panel *a*) and an assortative mixing setting (panel *b*).

Consistent patterns emerge, leading to the same qualitative results and outcomes. Specifically, the results reaffirm that under random mixing, predictions of subgroup attack rates align closely with those from an age-stratified model when vaccination concludes prior to the epidemic onset. However, as the overlap between the vaccination campaign and the epidemic grows, discrepancies increase. This effect is particularly pronounced under assortative mixing, where the interaction between subgroup activity differences and vaccination uptake prevents accurate predictions.

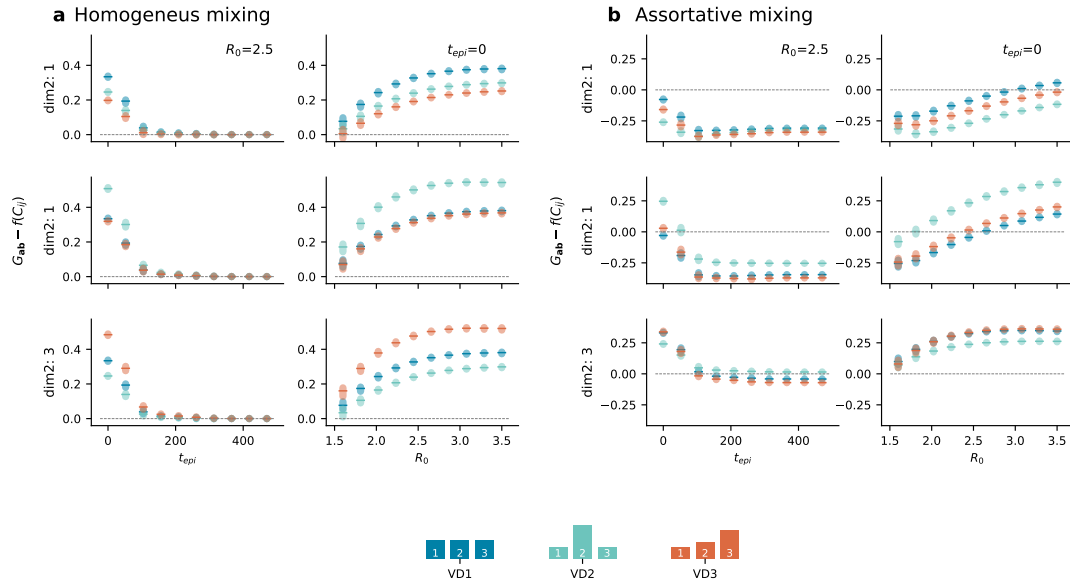


Figure D.9: Sensitivity analysis Differences between the overall attack rate for each subgroup as predicted by the generalized model G_{ab} and the attack rate predicted by the aggregate output of the age-stratified model, $f(C_{ij})$ as a function of (i) t_{epi} with $R_0 = 2.5$, and (ii) R_0 with $t_{epi} = 0$. These analysis are done both under a random mixing setting (panel *a*) and an assortative mixing setting (panel *b*)

D.5 Non pharmaceutical interventions

Table D.2 reports the values for activity, assortativity, and other free parameters used for both the tightened NPI simulation presented in Figure 2 (Scenario 1) and the relaxed NPI simulation presented in Figure 4 (Scenario 2) of Chapter 4.

SES Group	1	2	3
Assortativity	50%	60%	70%
Activity	40%	25%	35%
Free parameters	0.5	0.4	0.4

Table D.2: List of parameters used to simulate the NPIs

For completeness, we extend the model by adding a new compartment (D) to estimate the number of deaths under each vaccination distribution. In particular, we simulate the number of daily deaths by applying the age-stratified Infection Fatality Rate (IFR_i) estimated for COVID-19 by Ref.[181].

Section 1. We report the attack rate by socioeconomic status (SES) for both the Baseline setting and the setting with non-pharmaceutical interventions (NPI) in Figure D.10.

Figure D.11 presents the results of the model with deaths for Scenario 1. In particular, panels *a* and *b* show respectively the prevalence in the baseline case and the case with tightened NPIs where contacts are reduced by 20%. Similarly, panels *c* and *d* display the total number of deaths in the baseline case and under tightened NPIs, respectively. The mortality trend exhibits the same qualitative behaviour as the prevalence.

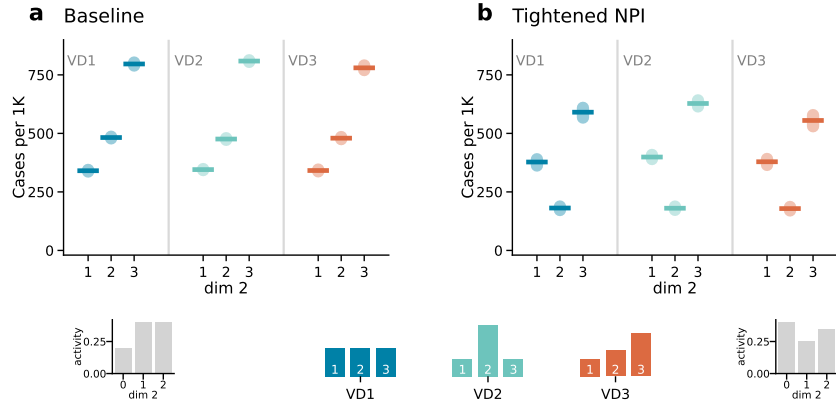


Figure D.10: Scenario 1: The impact of NPIs on SES groups under different vaccination distributions. Panel *a* refers to the baseline case and shows the attack rate scaled by 1000 for the three SES. Panel *b* follows the same structure but refers to the case where tightened NPIs are introduced, reducing contacts by 20%. NPIs are adjusted 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

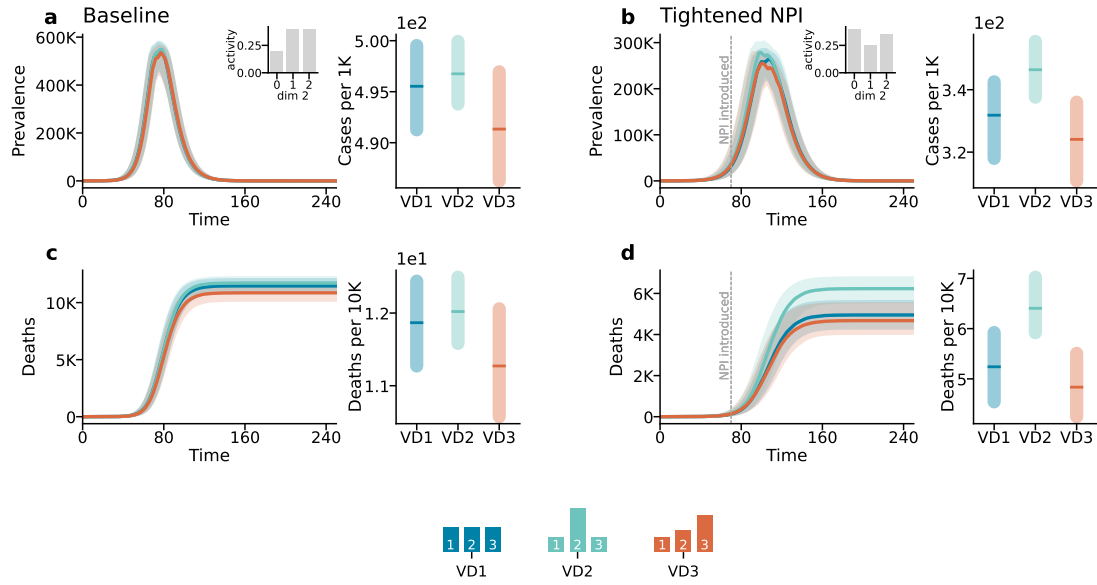


Figure D.11: Scenario 1: The impact of NPIs under different vaccination distributions. Panel *a* refers to the baseline case, with insets depicting the activity distribution along the second dimension. The first column shows the total number of infected individuals, while the second column shows the attack rate scaled by 1000. Panel *b* follows the same structure but refers to the case where tightened NPIs are introduced, reducing contacts by 20%. Panels *c* and *d* follow the same structure but refer to the total number of deaths (first column) and mortality rate scaled by 10K (second column). NPIs are adjusted 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

Scenario 2. Similarly, we report the attack rate by socioeconomic status (SES) for both the Baseline setting and the setting with non-pharmaceutical interventions (NPI) in Figure D.12.

While Figure D.13 presents the results of the model that includes mortality for Scenario 2 following the same structure of Figure D.11. Also in this case the mortality trend exhibits the same qualitative behavior as the prevalence.

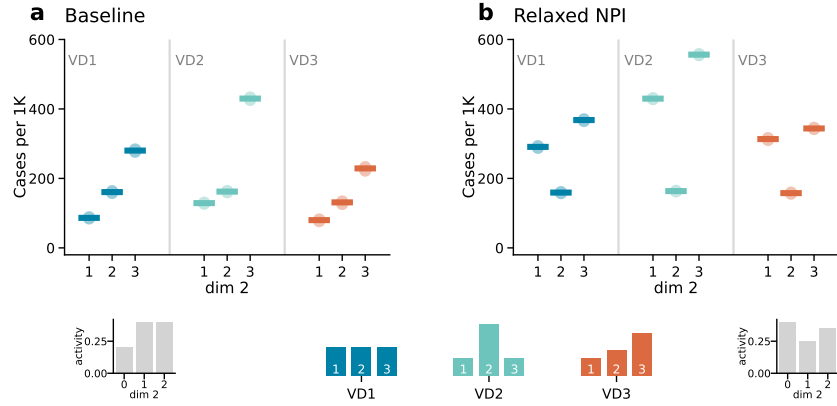


Figure D.12: Scenario 2: The impact of NPIs on SES groups under different vaccination distributions. Panel *a* refers to the baseline case and shows the attack rate scaled by 1000 for the three SES. Panel *b* follows the same structure but refers to the case where tightened NPIs are released and the contacts increase by 20%. NPIs are adjusted 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

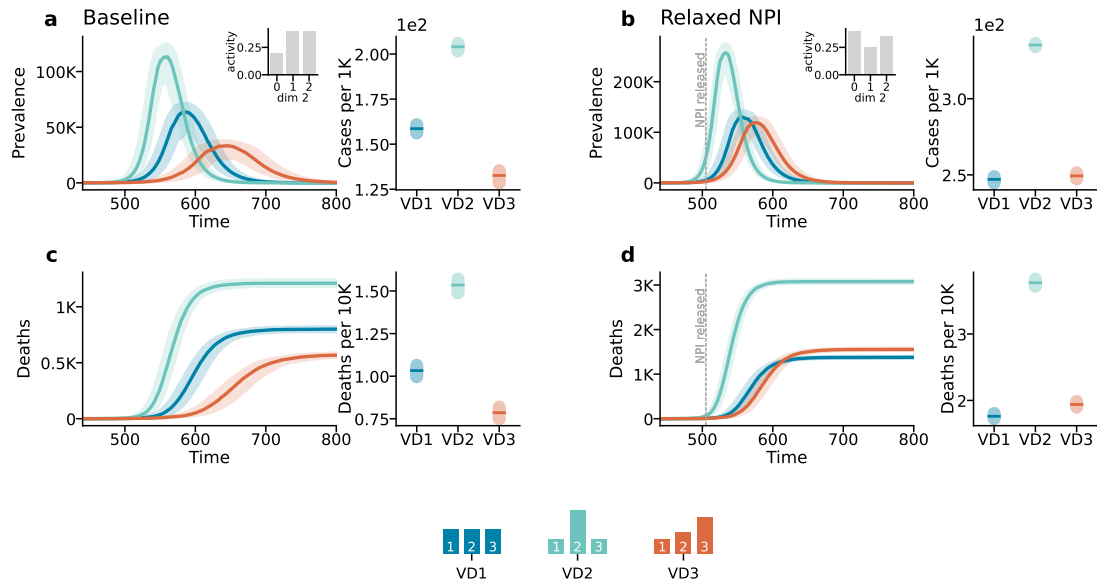


Figure D.13: Scenario 2: The impact of NPIs under different vaccination distributions. Panel *a* refers to the baseline case, with insets depicting the activity distribution along the second dimension. The first column shows the total number of infected individuals, while the second column shows the attack rate scaled by 1000. Panel *b* follows the same structure but refers to the case where tightened NPIs are released, increasing contacts by 20%. Panels *c* and *d* follow the same structure but refer to the total number of deaths (first column) and mortality rate scaled by 10K (second column). NPIs are adjusted 65 days after the epidemic onset. Results represent the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.5$, $g_2 = 0.8$, and $R_0 = 2.5$. Simulations begin with $I_0 = 100$ initial infected seeds.

D.6 Hungarian contact data

For completeness, also in this case we explore the number of deaths estimated by our model under the two different scenarios when the population distribution and the mixing patterns are taken from real data. Panels *a* and *c* of Fig.D.14 show the results for Scenario 1 while panels *b* and *d* of Fig.D.14 refers to Scenario 2.

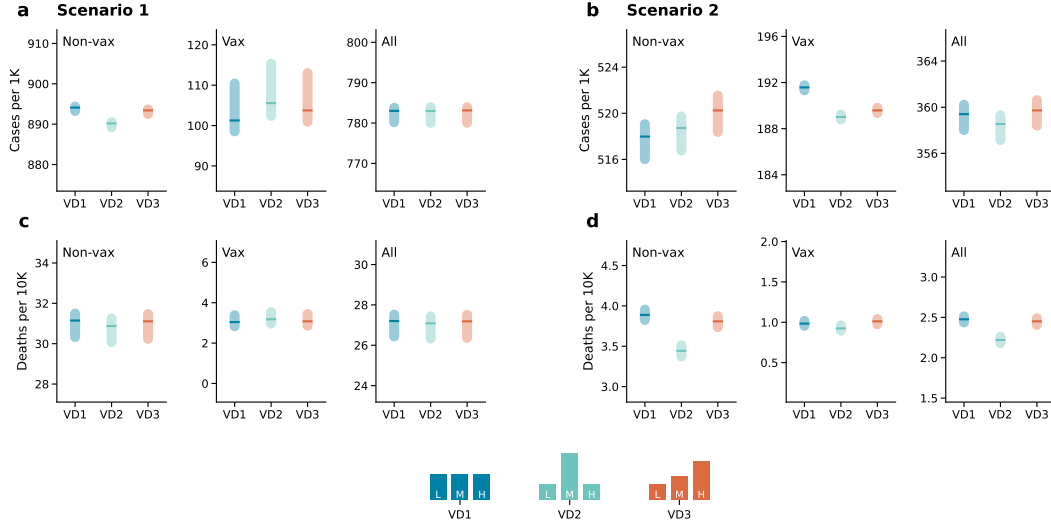


Figure D.14: Epidemic outcomes under different vaccination distributions and real contact data.. Panels in *a* and *c* refer to Scenario 1 and display the number of cases per 1000 (panel *a*) and the number of deaths per 10000 (panel *c*), stratified by vaccination status (non-vaccinated, vaccinated, and overall) and under different vaccination distribution scenarios (*VD0*, *VD2*, *VD3*). Panels in *b* and *d* show the corresponding number for Scenario 2. Results refer to the median of 500 runs with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$ and $g_2 = 0.8$. $R_0 = 2.7$ for the third wave and $R_0 = 3$ for the forth. Simulations start with $I_0 = 100$ initial

Additionally, here we present the same results stratified by age and vaccination status. These figures explore the dynamics of infection and mortality under different vaccination distribution scenarios (*VD1*, *VD2*, and *VD3*).

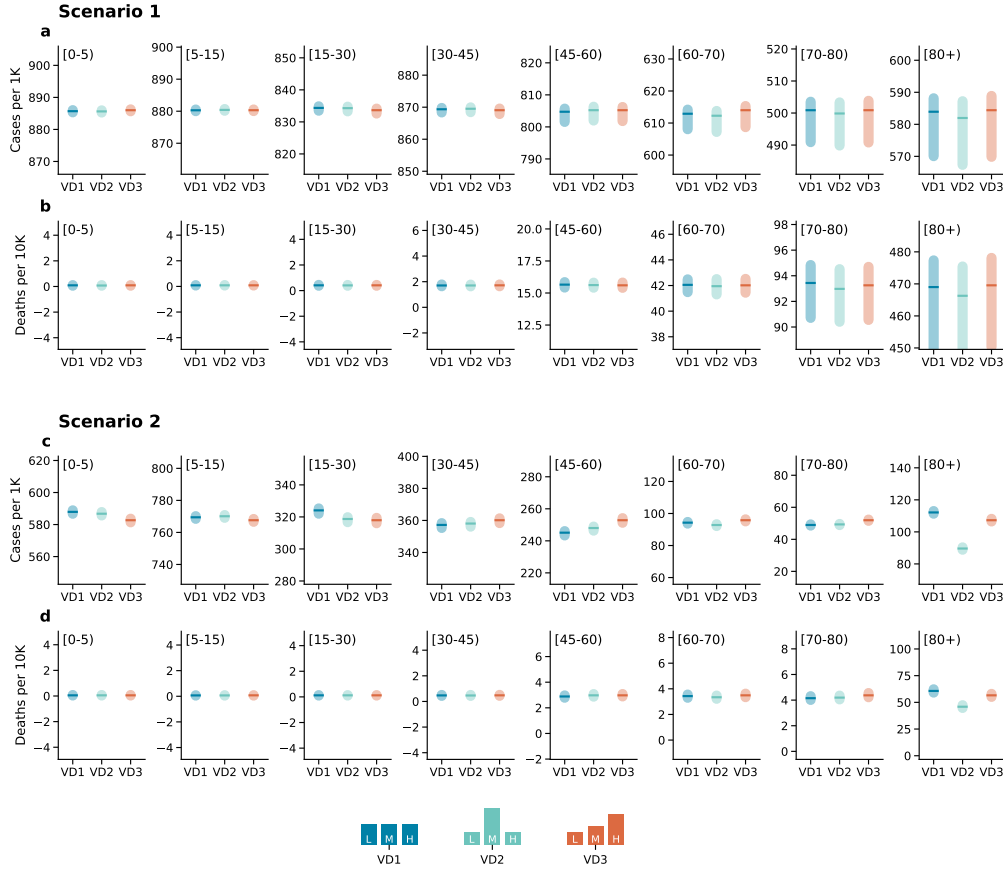


Figure D.15: Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the entire population. Panels (a) and (b) refer to Scenario 1. Panel (a) displays the number of cases per 1000 individuals under different vaccination distributions ($VD0, VD2, VD3$), while panel (b) shows the corresponding number of deaths per 10,000 individuals. Panels (c) and (d) present the same results for Scenario 2. Results represent the median of 500 simulations with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, and $g_2 = 0.8$. $R_0 = 2.7$ for the third wave and $R_0 = 3$ for the fourth. Simulations start with $I_0 = 100$ initial infectious cases.

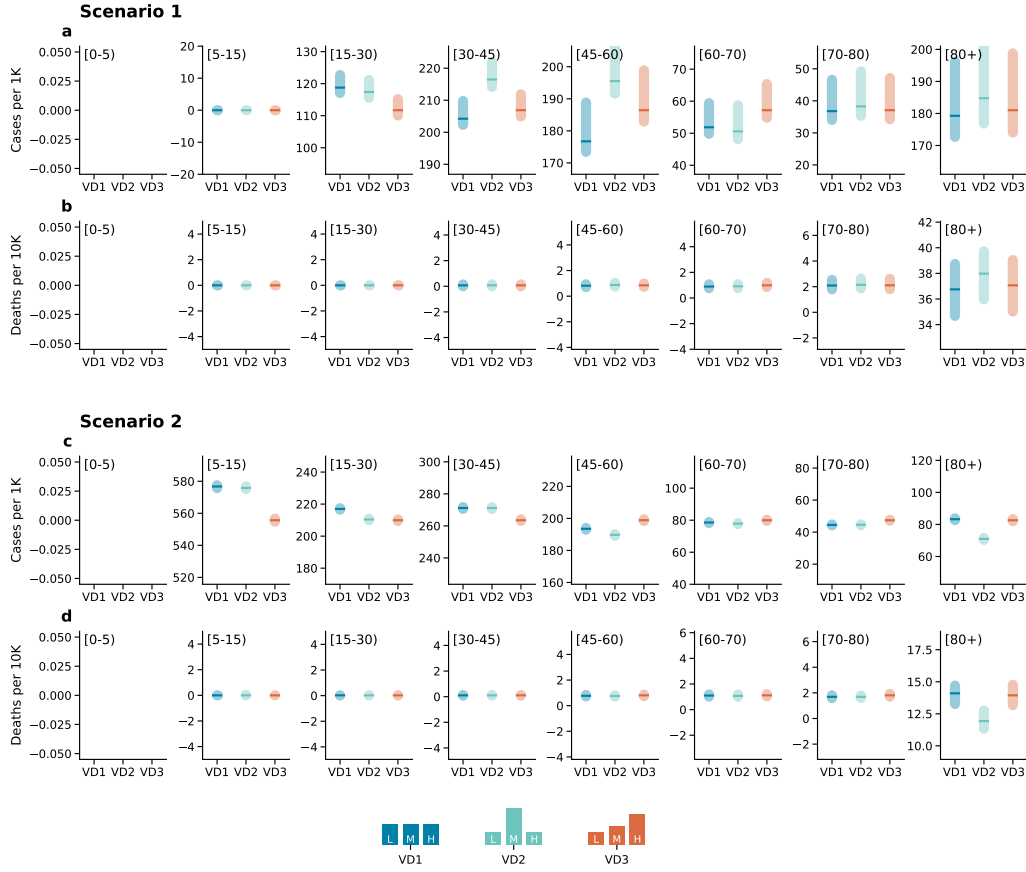


Figure D.16: Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the vaccinated population. Panels (a) and (b) refer to Scenario 1. Panel (a) displays the number of cases per 1000 individuals under different vaccination distributions ($VD0, VD2, VD3$), while panel (b) shows the corresponding number of deaths per 10,000 individuals. Panels (c) and (d) present the same results for Scenario 2. Results represent the median of 500 simulations with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, and $g_2 = 0.8$. $R_0 = 2.7$ for the third wave and $R_0 = 3$ for the fourth. Simulations start with $I_0 = 100$ initial infectious cases.

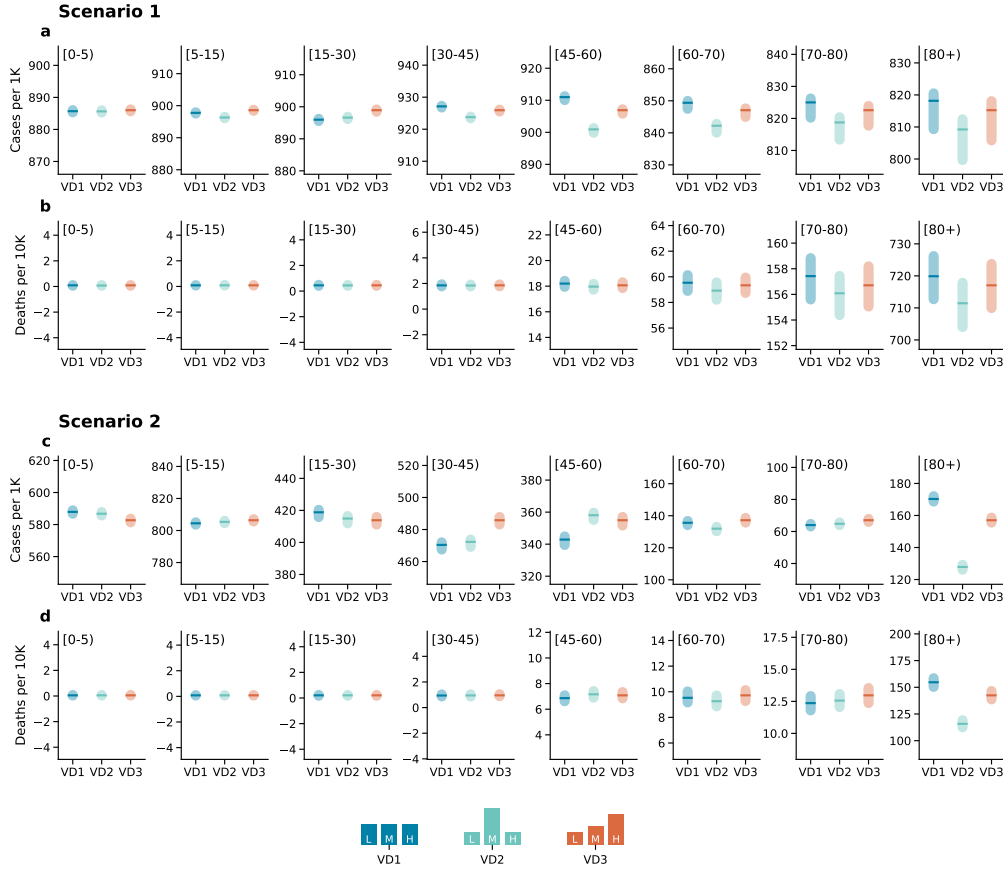


Figure D.17: Epidemic outcomes under different vaccination distributions and real contact data, stratified by age for the non-vaccinated population. Panels (a) and (b) refer to Scenario 1. Panel (a) displays the number of cases per 1000 individuals under different vaccination distributions ($VD0, VD2, VD3$), while panel (b) shows the corresponding number of deaths per 10,000 individuals. Panels (c) and (d) present the same results Scenario 2. Results represent the median of 500 simulations with confidence intervals. Epidemiological parameters: $\Gamma = 0.25$, $\Psi = 0.4$, $g_1 = 0.6$, and $g_2 = 0.8$. $R_0 = 2.7$ for the third wave and $R_0 = 3$ for the fourth. Simulations start with $I_0 = 100$ initial infectious cases.

Bibliography

- [1] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.
- [2] Roy M Anderson and Robert M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.
- [3] Stefano Merler and Marco Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681):557–565, 2010.
- [4] Kaiyuan Sun, Wei Wang, Lidong Gao, Yan Wang, Kaiwei Luo, Lingshuang Ren, Zhifei Zhan, Xinghui Chen, Shanlu Zhao, Yiwei Huang, Qianlai Sun, Ziyang Liu, Maria Litvinova, Alessandro Vespignani, Marco Ajelli, Cécile Viboud, and Hongjie Yu. Transmission heterogeneities, kinetics, and controllability of sars-cov-2. *Science*, 371(6526), 2021.
- [5] Michael Marmot, Sharon Friel, Ruth Bell, Tanja AJ Houweling, and Sebastian Taylor. Closing the gap in a generation: health equity through action on the social determinants of health. *The lancet*, 372(9650):1661–1669, 2008.
- [6] Svenn-Erik Mamelund, Clare Shelley-Egan, and Ole Rogeberg. The association between socioeconomic status and pandemic influenza: systematic review and meta-analysis. *PLoS One*, 16(9):e0244346, 2021.
- [7] Mariana Kikuti, Geraldo M Cunha, Igor AD Paploski, Amelia M Kasper, Monaise MO Silva, Aline S Tavares, Jaqueline S Cruz, Tássia L Queiroz, Moreno S Rodrigues, Perla M Santana, et al. Spatial distribution of dengue in a brazilian urban slum setting: role of socioeconomic gradient in disease risk. *PLoS neglected tropical diseases*, 9(7):e0003937, 2015.
- [8] Gonzalo E. Mena, Pamela P. Martinez, Ayesha S. Mahmud, Pablo A. Marquet, Caroline O. Buckee, and Mauricio Santillana. Socioeconomic status determines covid-19 incidence and related mortality in santiago, chile. *Science*, 372(6545):eabg5298, 2021.
- [9] Bo Burström and Wenjing Tao. Social determinants of health and inequalities in COVID-19. *European Journal of Public Health*, 30(4):617–618, 07 2020.
- [10] Ayan Paul, Philipp Englert, and Melinda Varga. Socio-economic disparities and covid-19 in the usa. *Journal of Physics: Complexity*, 2(3):035017, 2021.

- [11] H. Zhao, R. J. Harris, J. Ellis, and R. G. Pebody. Ethnicity, deprivation and mortality due to 2009 pandemic influenza a(h1n1) in england during the 2009/2010 pandemic and the first post-pandemic season. *Epidemiology and Infection*, 143(16):3375–3383, 2015.
- [12] Caroline Buckee, Abdisalan Noor, and Lisa Sattenspiel. Thinking clearly about social aspects of infectious disease transmission. *Nature*, 595(7866):205–213, 2021.
- [13] Michele Tizzoni, Elaine O Nsoesie, Laetitia Gauvin, Márton Karsai, Nicola Perra, and Shweta Bansal. Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nature Communications*, 13(1):1–7, 2022.
- [14] Jamie Bedson, Laura A Skrip, Danielle Pedi, Sharon Abramowitz, Simone Carter, Mohamed F Jalloh, Sebastian Funk, Nina Gobat, Tamara Giles-Vernick, Gerardo Chowell, et al. A review and agenda for integrated disease models including social and behavioural factors. *Nature human behaviour*, 5(7):834–846, 2021.
- [15] Jon Zelner, Nina B Masters, Ramya Naraharisetti, Sanyu A Mojola, Merlin Chowkwanyun, and Ryan Malosh. There are no equal opportunity infectors: epidemiological modelers must rethink our approach to inequality in infection risk. *PLoS computational biology*, 18(2):e1009795, 2022.
- [16] Kyra H Grantz, Madhura S Rane, Henrik Salje, Gregory E Glass, Stephen E Schachterle, and Derek AT Cummings. Disparities in influenza mortality and transmission related to sociodemographic factors within chicago in the pandemic of 1918. *Proceedings of the National Academy of Sciences*, 113(48):13839–13844, 2016.
- [17] Kathleen A Alexander, Claire E Sanderson, Madav Marathe, Bryan L Lewis, Caitlin M Rivers, Jeffrey Shaman, John M Drake, Eric Lofgren, Virginia M Dato, Marisa C Eisenberg, et al. What factors might have led to the emergence of ebola in west africa? *PLoS neglected tropical diseases*, 9(6):e0003652, 2015.
- [18] Nicola Perra. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Physics Reports*, 913:1–52, 2021.
- [19] Romain Garnier, Jan R Benetka, John Kraemer, Shweta Bansal, et al. Socioeconomic disparities in social distancing during the covid-19 pandemic in the united states: observational study. *Journal of medical Internet research*, 23(1):e24591, 2021.
- [20] Won Do Lee, Matthias Qian, and Tim Schwanen. The association between socioeconomic status and mobility reductions in the early stage of england’s covid-19 epidemic. *Health & place*, 69:102563, 2021.
- [21] Nicolò Gozzi, Michele Tizzoni, Matteo Chinazzi, Leo Ferres, Alessandro Vespignani, and Nicola Perra. Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. *Nature communications*, 12(1):1–9, 2021.
- [22] Samuel Heroy, Isabella Loaiza, Alex Pentland, and Neave O’Clery. Covid-19 policy analysis: labour structure dictates lockdown mobility behaviour. *Journal of the Royal Society Interface*, 18(176):20201035, 2021.

- [23] Laetitia Gauvin, Paolo Bajardi, Emanuele Pepe, Brennan Lake, Filippo Privitera, and Michele Tizzoni. Socio-economic determinants of mobility responses during the first wave of covid-19 in italy: from provinces to neighbourhoods. *Journal of The Royal Society Interface*, 18(181):20210092, 2021.
- [24] Joakim A Weill, Matthieu Stigler, Olivier Deschenes, and Michael R Springborn. Social distancing responses to covid-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences*, 117(33):19658–19660, 2020.
- [25] Jonathan Jay, Jacob Bor, Elaine O Nsoesie, Sarah K Lipson, David K Jones, Sandro Galea, and Julia Raifman. Neighbourhood income and physical distancing during the covid-19 pandemic in the united states. *Nature human behaviour*, 4(12):1294–1302, 2020.
- [26] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli, Andrea Flori, Alessandro Galeazzi, Francesco Porcelli, Ana Lucia Schmidt, Carlo Michele Valensise, Antonio Scala, Walter Quattrociochi, et al. Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, 117(27):15530–15535, 2020.
- [27] Spencer J Fox, Emily Javan, Remy Pasco, Graham C Gibson, Briana Betke, José L Herrera-Diestra, Spencer Woody, Kelly Pierce, Kaitlyn E Johnson, Maureen Johnson-León, et al. Disproportionate impacts of covid-19 in a large us city. *PLOS Computational Biology*, 19(6):e1011149, 2023.
- [28] Nicolò Gozzi, Matteo Chinazzi, Natalie E Dean, Ira M Longini Jr, M Elizabeth Haloran, Nicola Perra, and Alessandro Vespignani. Estimating the impact of COVID-19 vaccine inequities: a modeling study. *Nature Communications*, 14(1):3272, 2023.
- [29] Eugenio Valdano, Jonggul Lee, Shweta Bansal, Stefania Rubrichi, and Vittoria Colizza. Highlighting socio-economic constraints on mobility reductions during COVID-19 restrictions in France can inform effective and equitable pandemic response. *Journal of travel medicine*, 28(4):taab045, 2021.
- [30] Giulia Pullano, Eugenio Valdano, Nicola Scarpa, Stefania Rubrichi, and Vittoria Colizza. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the covid-19 epidemic in france under lockdown: a population-based study. *The Lancet Digital Health*, 2(12):e638–e649, 2020.
- [31] Kerry LM Wong, Amy Gimma, Pietro Coletti, Christel Faes, Philippe Beutels, Niel Hens, Veronika K Jaeger, Andre Karch, Helen Johnson, et al. Social contact patterns during the covid-19 pandemic in 21 european countries—evidence from a two-year study. *BMC infectious diseases*, 23(1):268, 2023.
- [32] Isolde Sommer, Ursula Griebler, Peter Mähknecht, Kylie Thaler, Kathryn Bouskill, Gerald Gartlehner, and Shanti Mendis. Socioeconomic inequalities in non-communicable diseases and their risk factors: an overview of systematic reviews. *BMC Public Health*, 15, 09 2015.
- [33] Clare Bambra, Ryan Riordan, John Ford, and Fiona Matthews. The covid-19 pandemic and health inequalities. *Journal of Epidemiology & Community Health*, 74(11):964–968, 2020.

- [34] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, Oxford, 1991.
- [35] Kathy Leung, Mark Jit, Eric HY Lau, and Joseph T Wu. Social contact patterns relevant to the spread of respiratory infectious diseases in hong kong. *Scientific reports*, 7(1):7974, 2017.
- [36] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, (3):e74, 2008.
- [37] Alessia Melegaro, Emanuele Del Fava, Piero Poletti, Stefano Merler, Constance Nyamukapa, John Williams, Simon Gregson, and Piero Manfredi. Social contact structures and time use patterns in the manicaland province of zimbabwe. *PLOS ONE*, 12(1):1–17, 01 2017.
- [38] Dina Mistry, Maria Litvinova, Ana Pastore y Piontti, Matteo Chinazzi, Laura Fumanelli, Marcelo F C Gomes, Syed A Haque, Quan-Hui Liu, Kunpeng Mu, Xinyue Xiong, M Elizabeth Halloran, Ira M Longini, Stefano Merler, Marco Ajelli, and Alessandro Vespignani. Inferring high-resolution human mixing patterns for disease modeling. *Nature Communications*, 12(1):323, 2021.
- [39] Kiesha Prem, Alex R Cook, and Mark Jit. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS computational biology*, 13(9):e1005697, 2017.
- [40] Carlos G Grijalva, Nele Goeyvaerts, Hector Verastegui, Kathryn M Edwards, Ana I Gil, Claudio F Lanata, Niel Hens, et al. A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru. *PloS one*, 10(3):e0118457, 2015.
- [41] Nicolò Gozzi, Matteo Chinazzi, Jessica T Davis, Kunpeng Mu, Ana Pastore y Piontti, Marco Ajelli, Nicola Perra, and Alessandro Vespignani. Anatomy of the first six months of covid-19 vaccination campaign in italy. *PLoS Computational Biology*, 18(5):e1010146, 2022.
- [42] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cécile Viboud, Alessandro Vespignani, Marco Ajelli, and Hongjie Yu. Changes in contact patterns shape the dynamics of the covid-19 outbreak in china. *Science*, 368(6498):1481–1486, 2020.
- [43] Andria Mousa, Peter Winskill, Oliver John Watson, Oliver Ratmann, Mélodie Monod, Marco Ajelli, Aldiouma Diallo, Peter J Dodd, Carlos G Grijalva, Moses Chapa Kiti, et al. Social contact patterns and implications for infectious disease transmission—a systematic review and meta-analysis of contact surveys. *ELife*, 10:e70294, 2021.
- [44] Thang Hoang, Pietro Coletti, Alessia Melegaro, Jacco Wallinga, Carlos G Grijalva, John W Edmunds, Philippe Beutels, and Niel Hens. A systematic review of social contact surveys to inform transmission models of close-contact infections. *Epidemiology (Cambridge, Mass.)*, 30(5):723, 2019.

- [45] Vittoria Colizza, Marc Barthélemy, Alain Barrat, and Alessandro Vespignani. Epidemic modeling in complex realities. *Comptes Rendus Biologies*, 330(4):364 – 374, 2007. L'épidémiologie : une science en développement / Epidemiology : a developing science.
- [46] Jacco Wallinga, Peter Teunis, and Mirjam Kretzschmar. Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents. *American Journal of Epidemiology*, 164(10):936–944, 09 2006.
- [47] Mackenzie Hamilton, Jesse Knight, and Sharmistha Mishra. Examining the Influence of Imbalanced Social Contact Matrices in Epidemic Models. *American Journal of Epidemiology*, page kwad185, 09 2023.
- [48] Sergio Arregui, Alberto Aleta, Joaquín Sanz, and Yamir Moreno. Projecting social contact matrices to different demographic structures. *PLOS Computational Biology*, 14(12):1–18, 12 2018.
- [49] Mark Newman. *Networks: An Introduction*. Oxford university press, 2010.
- [50] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.
- [51] Bryan Grenfell and John Harwood. (meta) population dynamics of infectious diseases. *Trends in ecology & evolution*, 12(10):395–399, 1997.
- [52] Alexei Vazquez. Epidemic outbreaks on structured populations. *Journal of Theoretical Biology*, 245(1):125–129, 2007.
- [53] Andrea Apolloni, Chiara Poletto, José J Ramasco, Pablo Jensen, and Vittoria Colizza. Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling*, 11:1–26, 2014.
- [54] Vittoria Colizza and Alessandro Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of Theoretical Biology*, 251(3):450 – 467, 2008.
- [55] Dennis L Chao, M Elizabeth Halloran, Valerie J Obenchain, and Ira M Longini Jr. Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS computational biology*, 6(1):e1000656, 2010.
- [56] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS medicine*, 4(1):e13, 2007.
- [57] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sapon Iamsirithaworn, and Donald S Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437(7056):209–214, 2005.
- [58] Ira M Longini Jr, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaworakul, Derek AT Cummings, and M Elizabeth Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

- [59] Wouter Van den Broeck, Corrado Gioannini, Bruno Gonçalves, Marco Quaggiotto, Vittoria Colizza, and Alessandro Vespignani. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC infectious diseases*, 11:1–14, 2011.
- [60] Daniel Bernoulli and Dominique Chapelle. Essai d’une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l’inoculation pour la prévenir. 1760.
- [61] Daniel Bernoulli. Réflexions sur les avantages de l’inoculation. *Mém. Paris*, pages 439–482, 1758.
- [62] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.
- [63] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.
- [64] P Rohani M Keeling. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.
- [65] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.
- [66] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.
- [67] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan Anton Jacob Metz. On the definition and the computation of the basic reproduction ratio r_0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28:365–382, 1990.
- [68] Odo Diekmann, JAP Heesterbeek, and Michael G Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the royal society interface*, 7(47):873–885, 2010.
- [69] Julie C Blackwood and Lauren M Childs. An introduction to compartmental modeling for the budding infectious disease modeler. *Letters in Biomathematics*, 5(1):195–221, 2018.
- [70] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton university press, 2011.
- [71] Zhen Wang, Chris T Bauch, Samit Bhattacharyya, Alberto d’Onofrio, Piero Manfredi, Matjaž Perc, Nicola Perra, Marcel Salathé, and Dawei Zhao. Statistical physics of vaccination. *Physics Reports*, 664:1–113, 2016.
- [72] Amelie Desvars-Larrive, Elma Dervic, Nina Haug, Thomas Niederkrotenthaler, Jiaying Chen, Anna Di Natale, Jana Lasser, Diana S Gliga, Alexandra Roux, Johannes Sorger, et al. A structured open dataset of government interventions in response to covid-19. *Scientific data*, 7(1):285, 2020.

- [73] Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Michele Tizzoni, Vittoria Colizza, and Alessandro Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one*, 6(1):e16591, 2011.
- [74] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.
- [75] Nicolò Gozzi, Paolo Bajardi, and Nicola Perra. The importance of non-pharmaceutical interventions during the covid-19 vaccine rollout. *PLoS computational biology*, 17(9):e1009346, 2021.
- [76] Laura Di Domenico, Giulia Pullano, Chiara E Sabbatini, Pierre-Yves Boëlle, and Vittoria Colizza. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC medicine*, 18(1):1–13, 2020.
- [77] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cécile Viboud, Alessandro Vespignani, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science*, 368(6498):1481–1486, 2020.
- [78] Alessia Melegaro, Emanuele Del Fava, Piero Poletti, Stefano Merler, Constance Nyamukapa, John Williams, Simon Gregson, and Piero Manfredi. Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe. *PLOS ONE*, 12(1):1–17, 01 2017.
- [79] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the national academy of sciences*, 107(51):22020–22025, 2010.
- [80] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.
- [81] Thomas Obadia, Romain Silhol, Lulla Opatowski, Laura Temime, Judith Legrand, Anne CM Thiébaud, Jean-Louis Herrmann, Eric Fleury, Didier Guillemot, Pierre-Yves Boelle, et al. Detailed contact data and the dissemination of staphylococcus aureus in hospitals. *PLoS Computational Biology*, 11(3):e1004170, 2015.
- [82] Surveillance definitions for COVID-19, European Centre for Disease Prevention and Control, date of access 2023.05.23.
- [83] Timo Smieszek, Elena U Burri, Robert Scherzinger, and Roland W Scholz. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology & infection*, 140(4):744–752, 2012.
- [84] Timo Smieszek, Victoria C Barclay, Indulaxmi Seeni, Jeanette J Rainey, Hongjiang Gao, Amra Uzicanin, and Marcel Salathé. How should social mixing be measured: comparing web-based survey and sensor-based methods. *BMC infectious diseases*, 14:1–13, 2014.

- [85] Beata Dávid, Eva Huszti, Ildikó Barna, and Yang chih Fu. Egocentric contact networks in comparison: Taiwan and hungary. *Social Networks*, 44(January 2016):253–265, 2016.
- [86] Hwai-Chung Ho Fu, Yang-chih and Hsiu Man Chen. Weak ties and contact initiation in everyday life: Exploring contextual variations from contact diaries. *Social Networks*, 35(3):279–287, 2013.
- [87] DL Paulhus and S Vazire. The self-report method. handbook of research methods in personality psychology, 2007.
- [88] H Russell Bernard, Peter Killworth, David Kronenfeld, and Lee Sailer. The problem of informant accuracy: The validity of retrospective data. *Annual review of anthropology*, 13(1):495–517, 1984.
- [89] Fred Collopy. Biases in retrospective self-reports of time use: An empirical study of computer users. *Management Science*, 42(5):758–767, 1996.
- [90] Filippo Trentini, Adriana Manna, Nicoletta Balbo, Valentina Marziano, Giorgio Guzzetta, Samantha O’Dell, Allisandra G Kummer, Maria Litvinova, Stefano Merler, Marco Ajelli, et al. Investigating the relationship between interventions, contact patterns, and sars-cov-2 transmissibility. *Epidemics*, 40:100601, 2022.
- [91] UHasselt. The comix study, 2024. Accessed: 2024-06-22.
- [92] Márton Karsai, Júlia Koltai, Orsolya Vásárhelyi, and Gergely Röst. Hungary in mask/maszk in Hungary. *Corvinus Journal of Sociology and Social Policy*, (2), 2020.
- [93] Sicheng Dai, Hélène Bouchet, Márton Karsai, Jean-Pierre Chevrot, Eric Fleury, and Aurélie Nardy. Longitudinal data collection to follow social network and language development dynamics at preschool. *Scientific Data*, 9(1):777, 2022.
- [94] Sicheng Dai, Mårton Karsai, Hélène Bouchet, Eric Fleury, Jean-Pierre Chevrot, and Aurélie Nardy. Interaction reconstruction methods for large-scale rfid social experiments. In *IC2S2 2019-5th International Conference on Computational Social Science*, volume 2, 2019.
- [95] James O’Connell, Manzar Abbas, Sarah Beecham, Jim Buckley, Muslim Chochlov, Brian Fitzgerald, Liam Glynn, Kevin Johnson, John Laffey, Bairbre McNicholas, et al. Best practice guidance for digital contact tracing apps: a cross-disciplinary review of the literature. *JMIR mHealth and uHealth*, 9(6):e27753, 2021.
- [96] Robert A Kleinman and Colin Merkel. Digital contact tracing for covid-19. *Cmaj*, 192(24):E653–E656, 2020.
- [97] Muhammad Shahroz, Farooq Ahmad, Muhammad Shahzad Younis, Nadeem Ahmad, Maged N Kamel Boulos, Ricardo Vinuesa, and Junaid Qadir. Covid-19 digital contact tracing applications and techniques: A review post initial deployments. *Transportation Engineering*, 5:100072, 2021.
- [98] Jesús A Moreno López, Beatriz Arregui García, Piotr Bentkowski, Livio Bioglio, Francesco Pinotti, Pierre-Yves Boëlle, Alain Barrat, Vittoria Colizza, and Chiara Poletto. Anatomy of digital contact tracing: Role of age, transmission setting, adoption, and case detection. *Science advances*, 7(15):eabd8750, 2021.

- [99] Kiesha Prem, Kevin van Zandvoort, Petra Klepac, et al. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the covid-19 era. *medrxiv* 2020; published online july 25. DOI, 10(2020.07):22–20159772, 2021.
- [100] Angela D’Adamo, Alina Schnake-Mahl, Pricila H Mullachery, Mariana Lazo, Ana V Diez Roux, and Usama Bilal. Health disparities in past influenza pandemics: A scoping review of the literature. *SSM-population health*, 21:101314, 2023.
- [101] C.T. Laurencin and A. McClinton. The covid-19 pandemic: a call to action to identify and address racial and ethnic disparities. *J. Racial and Ethnic Health Disparities*, page 398–402, 2020.
- [102] Merrill Singe. *Introduction to syndemic: a systems approach to public and community health*. San Francisco, CA: Jossey-Bass, 2009.
- [103] Jarvis T. Chen and Nancy Krieger. Revealing the unequal burden of covid-19 by income, race/ethnicity, and household crowding: Us county vs zip code analyses. *Harvard Center for Population and Development Studies Working Paper Series.*, 19, 04 2020.
- [104] J. Tom Mueller, Kathryn McConnell, Paul Berne Burow, Katie Pofahl, Alexis A. Merdjanoff, and Justin Farrell. Impacts of the covid-19 pandemic on rural america. *Proceedings of the National Academy of Sciences*, 118(1), 2021.
- [105] Karan Varshney, Talia Glodjo, and Jenna Adalbert. Overcrowded housing increases risk for covid-19 mortality: an ecological study. *BMC Research Notes*, 15(1):126, 2022.
- [106] Ahmad Khanijahani, Shabnam Iezadi, Kamal Gholipour, Saber Azami-Aghdash, and Deniz Naghibi. A systematic review of racial/ethnic and socioeconomic disparities in covid-19. *International journal for equity in health*, 20:1–30, 2021.
- [107] Noreen Goldman, Anne R Pebley, Keunbok Lee, Theresa Andrasfay, and Boriana Pratt. Racial and ethnic differentials in covid-19-related job exposures by occupational standing in the us. *PloS one*, 16(9):e0256085, 2021.
- [108] Yannick Leo, Eric Fleury, J Ignacio Alvarez-Hamelin, Carlos Sarraute, and Márton Karsai. Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598, 2016.
- [109] Xiaowen Dong, Alfredo J Morales, Eaman Jahani, Esteban Moro, Bruno Lepri, Burcin Bozkaya, Carlos Sarraute, Yaneer Bar-Yam, and Alex Pentland. Segregated interactions in urban and online space. *EPJ Data Science*, 9(1):20, 2020.
- [110] Ludovico Napoli, Vedran Sekara, Manuel García-Herranz, and Márton Karsai. Socioeconomic reorganization of communication and mobility networks in response to external shocks. *Proceedings of the National Academy of Sciences*, 120(50):e2305285120, 2023.
- [111] World Health Organization. Health inequities and their causes, 2023. Accessed: 2024-07-03.

- [112] Robert L Jin, Chandrakant P Shah, and Tomislav J Svoboda. The impact of unemployment on health: a review of the evidence. *CMAJ: Canadian Medical Association Journal*, 153(5):529, 1995.
- [113] Rino Rappuoli, Mariagrazia Pizza, Giuseppe Del Giudice, and Ennio De Gregorio. Vaccines, new opportunities for a new society. *Proceedings of the National Academy of Sciences*, 111(34):12288–12293, 2014.
- [114] Nuru Saadi, Y-Ling Chi, Srobana Ghosh, Rosalind M Eggo, Ciara V McCarthy, Matthew Quaife, Jeanette Dawa, Mark Jit, and Anna Vassall. Models of covid-19 vaccine prioritisation: a systematic literature search and narrative review. *BMC medicine*, 19:1–11, 2021.
- [115] World Health Organization et al. Who sage roadmap for prioritizing uses of covid-19 vaccines in the context of limited supply: an approach to inform planning and subsequent recommendations based on epidemiological setting and vaccine supply scenarios, first issued 20 october 2020, latest update 16 july 2021. Technical report, World Health Organization, 2021.
- [116] Richard L Oehler and Vivian R Vega. Conquering covid: how global vaccine inequality risks prolonging the pandemic. In *Open Forum Infectious Diseases*, volume 8, page ofab443. Oxford University Press US, 2021.
- [117] Malik Sallam. Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2):160, 2021.
- [118] Moosa Tatar, Jalal Montazeri Shoorekchali, Mohammad Reza Faraji, Mohammad Abdi Seyyedkolaei, José A Pagán, and Fernando A Wilson. Covid-19 vaccine inequality: a global perspective. *Journal of global health*, 12, 2022.
- [119] Nicole Bergen, Katherine Kirkby, Cecilia Vidal Fuertes, Anne Schlottheuber, Lisa Menning, Stephen Mac Feely, Katherine O’Brien, and Ahmad Reza Hosseinpour. Global state of education-related inequality in covid-19 vaccine coverage, structural barriers, vaccine hesitancy, and vaccine refusal: findings from the global covid-19 trends and impact survey. *The Lancet Global Health*, 11(2):e207–e217, 2023.
- [120] Mathew D Marques and Meru Sheel. Addressing the social inequities of vaccination: an imperative to close the gap. *The Lancet Global Health*, 11(2):e173–e174, 2023.
- [121] Anna Sára Ligeti, Beatrix Oroszi, Csaba Luca, Edit Bilics, József Ágoston, Gergely Röst, and Júlia Koltai. Socioeconomic determinants and reasons for non-acceptance to vaccination recommendations during the 3rd-5th waves of the covid-19 pandemic in hungary. *BMC Public Health*, 24(1):1796, 2024.
- [122] Adriana Manna, Júlia Koltai, and Márton Karsai. Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics. *Nature Communications*, 15(1):4137, 2024.
- [123] Beatrix Oroszi, Attila Juhász, Csilla Nagy, Judit Krisztina Horváth, Krisztina Eszter Komlós, Gergő Túri, Martin McKee, and Róza Ádány. Characteristics of the Third COVID-19 Pandemic Wave with Special Focus on Socioeconomic Inequalities in Morbidity, Mortality and the Uptake of COVID-19 Vaccination in Hungary. *Journal of personalized medicine*, 12(3):388, 2022.

- [124] Jennifer Attonito, Whitney Van Arsdale, Keren Fishman, Maral Darya, Mario Jacomino, and George Luck. Sociodemographic disparities in access to covid-19 vaccines upon initial rollout in florida: Study examines sociodemographic disparities in access to covid-19 vaccines upon initial rollout in florida. *Health Affairs*, 40(12):1883–1891, 2021.
- [125] Malorie Perry, Ashley Akbari, Simon Cottrell, Michael B Gravenor, Richard Roberts, Ronan A Lyons, Stuart Bedston, Fatemah Torabi, and Lucy Griffiths. Inequalities in coverage of covid-19 vaccination: A population register based cross-sectional study in wales, uk. *Vaccine*, 39(42):6256–6261, 2021.
- [126] Eulalia C Torres, Maribel Moreno, and María F Rivadeneira. Vaccination against covid-19 and socioeconomic inequalities: a cross-sectional study in ecuador. *Vaccine: X*, 15:100393, 2023.
- [127] Ruth Elizabeth Watkinson, Richard Williams, Stephanie Gillibrand, Caroline Sanders, and Matt Sutton. Ethnic inequalities in covid-19 vaccine uptake and comparison to seasonal influenza vaccine uptake in greater manchester, uk: A cohort study. *PLoS medicine*, 19(3):e1003932, 2022.
- [128] Amber Sacre, Clare Bambra, Josephine M Wildman, Katie Thomson, Natalie Bennett, Sarah Sowden, and Adam Todd. Socioeconomic inequalities in vaccine uptake: A global umbrella review. *Plos one*, 18(12):e0294688, 2023.
- [129] Majbritt Tang Svendsen, Carsten Kronborg Bak, Kristine Sørensen, Jürgen Pelikan, Signe Juul Riddersholm, Regitze Kuhr Skals, Rikke Nørmark Mortensen, Helle Terkildsen Maindal, Henrik Bøggild, Gitte Nielsen, et al. Associations of health literacy with socioeconomic position, health risk behavior, and health status: a large national population-based survey among danish adults. *BMC public health*, 20:1–12, 2020.
- [130] Nathalie Bajos, Alexis Spire, Léna Silberzan, and EPICOV Study Group. The social specificities of hostility toward vaccination against covid-19 in france. *PLoS One*, 17(1):e0262192, 2022.
- [131] Nathalie Bajos, Alexis Spire, Léna Silberzan, Antoine Sireyjol, Florence Jusot, Laurence Meyer, Jeanna-Eve Franck, Josiane Warszawski, and EpiCov Study Group. When lack of trust in the government and in scientists reinforces social inequalities in vaccination against covid-19. *Frontiers in Public Health*, 10:908152, 2022.
- [132] Marcel Salathé and Shashank Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199, 2011.
- [133] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140, 2020.
- [134] Yanay Gorelik and Michael Edelstein. Inequalities in initiation of covid19 vaccination by age and population group in israel-december 2020-april 2021. *MedRxiv*, pages 2021–06, 2021.

- [135] Sean AP Clouston, Douglas W Hanes, and Bruce G Link. Social inequalities and the early provision and dispersal of covid-19 vaccinations in the united states: A population trends study. *Vaccine*, 41(36):5322–5329, 2023.
- [136] Beatrix Oroszi, Attila Juhász, Csilla Nagy, Judit Krisztina Horváth, Martin McKee, and Róza Ádány. Unequal burden of covid-19 in hungary: a geographical and socioeconomic analysis of the second wave of the pandemic. *BMJ global health*, 6(9):e006427, 2021.
- [137] Dennis M Feehan and Ayesha S Mahmud. Quantifying population contact patterns in the united states during the covid-19 pandemic. *Nature communications*, 12(1):1–9, 2021.
- [138] Thomas Brambor, William Roberts Clark, and Matt Golder. Understanding interaction models: Improving empirical analyses. *Political analysis*, 14(1):63–82, 2006.
- [139] Carina Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82, 2010.
- [140] Paul D Allison. Comparing logit and probit coefficients across groups. *Sociological methods & research*, 28(2):186–208, 1999.
- [141] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton & Company, 2007.
- [142] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*, volume 488. Springer, Verlag London, 2005.
- [143] James Wambua, Neilshan Loedy, Christopher I Jarvis, Kerry LM Wong, Christel Faes, Rok Grah, Bastian Prasse, Frank Sandmann, Rene Niehus, Helen Johnson, et al. The influence of covid-19 risk perception and vaccination status on the number of social contacts across europe: insights from the comix study. *BMC Public Health*, 23(1):1350, 2023.
- [144] János Sándor, Ferenc Vincze, Maya Liza Shrikant, László Kőrösi, László Ulicska, Karolina Kósa, and Róza Ádány. Covid-19 vaccination coverage in deprived populations living in segregated colonies: A nationwide cross-sectional study in hungary. *Plos one*, 17(2):e0264363, 2022.
- [145] Chiara Cadeddu, Aldo Rosano, Leonardo Villani, Giovanni Battista Coiante, Ilaria Minicucci, Domenico Pascucci, and Chiara de Waure. Planning and organization of the covid-19 vaccination campaign: An overview of eight european countries. *Vaccines*, 10(10):1631, 2022.
- [146] Amanda Minter and Renata Retkute. Approximate bayesian computation for infectious disease modelling. *Epidemics*, 29:100368, 2019.
- [147] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.
- [148] Koronavírus elhunytakra és gyógyultakra vonatkozó településszintű adatok, átlászo.hu. Kimittud. Accessed: May 9, 2023.

- [149] Matthew Quaife, Kevin Van Zandvoort, Amy Gimma, Kashvi Shah, Nicky McCreesh, Kiesha Prem, Edwine Barasa, Daniel Mwanga, Beth Kangwana, Jessie Pinchoff, et al. The impact of covid-19 control measures on social contacts and transmission in kenyan informal settlements. *BMC medicine*, 18:1–11, 2020.
- [150] Adriana Manna, Lorenzo Dall’Amico, Michele Tizzoni, Márton Karsai, and Nicola Perra. Generalized contact matrices allow integrating socioeconomic variables into epidemic models. *Science Advances*, 10(41):eadk4606, 2024.
- [151] Kaiyuan Sun, Andrea Baronchelli, and Nicola Perra. Contrasting effects of strong ties on SIR and SIS processes in temporal networks. *The European Physical Journal B*, 88:1–8, 2015.
- [152] Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex Pentland. Mobility patterns are associated with experienced income segregation in large US cities. *Nature communications*, 12(1):4633, 2021.
- [153] Takahiro Yabe, Bernardo García Bulle Bueno, Xiaowen Dong, Alex Pentland, and Esteban Moro. Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters. *Nature Communications*, 14(1):2310, 2023.
- [154] Edouard Mathieu, Hannah Ritchie, Esteban Ortiz-Ospina, Max Roser, Joe Hasell, Cameron Appel, Charlie Giattino, and Lucas Rodés-Guirao. A global database of covid-19 vaccinations. *Nature human behaviour*, 5(7):947–953, 2021.
- [155] Valentina Marziano, Giorgio Guzzetta, Alessia Mammone, Flavia Riccardo, Piero Poletti, Filippo Trentini, Mattia Manica, Andrea Siddu, Antonino Bella, Paola Stefanelli, et al. The effect of covid-19 vaccination in italy and perspectives for living with the virus. *Nature communications*, 12(1):7272, 2021.
- [156] Sudam Surasinghe, Swathi Nachiar Manivannan, Samuel V Scarpino, Lorin Crawford, and C Brandon Ogbunugafor. Structural causal influence (sci) captures the forces of social inequality in models of disease dynamics. *arXiv preprint arXiv:2409.09096*, 2024.
- [157] Adriana Manna, Marton Karsai, and Nicola Perra. Social inequalities in vaccine coverage and their effects on epidemic spreading. *medRxiv*, 2024.
- [158] Monita Karmakar, Paula M Lantz, and Renuka Tipirneni. Association of social and demographic factors with COVID-19 incidence and death rates in the US. *JAMA network open*, 4(1):e2036462–e2036462, 2021.
- [159] Jon Zelner, Rob Trangucci, Ramya Naraharisetti, Alex Cao, Ryan Malosh, Kelly Broen, Nina Masters, and Paul Delamater. Racial disparities in coronavirus disease 2019 (COVID-19) mortality are driven by unequal infection risks. *Clinical Infectious Diseases*, 72(5):e88–e95, 2021.
- [160] Judy Yuen-man Siu. Health inequality experienced by the socially disadvantaged populations during the outbreak of COVID-19 in Hong Kong: An interaction with social inequality. *Health & social care in the community*, 29(5):1522–1529, 2021.
- [161] Sedona Sweeney, Theo Prudencio Juhani Capeding, Rosalind Eggo, Maryam Huda, Mark Jit, Don Mudzengi, Nichola R Naylor, Simon Procter, Matthew Quaife, Lela Serebryakova, et al. Exploring equity in health and poverty impacts of control measures for SARS-CoV-2 in six countries. *BMJ global health*, 6(5):e005521, 2021.

- [162] Kevin C Ma, Tigist F Menkir, Stephen Kissler, Yonatan H Grad, and Marc Lipsitch. Modeling the impact of racial and ethnic disparities on COVID-19 epidemic dynamics. *Elife*, 10:e66601, 2021.
- [163] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.
- [164] Alberto Aleta, David Martín-Corral, Michiel A Bakker, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E Dean, M Elizabeth Halloran, Ira M Longini Jr, et al. Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas. *Proceedings of the National Academy of Sciences*, 119(26):e2112182119, 2022.
- [165] Alberto Aleta, David Martin-Corral, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E Dean, M Elizabeth Halloran, Ira M Longini Jr, Stefano Merler, et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour*, 4(9):964–971, 2020.
- [166] Marco Pangallo, Alberto Aleta, R Maria del Rio-Chanona, Anton Pichler, David Martín-Corral, Matteo Chinazzi, François Lafond, Marco Ajelli, Esteban Moro, Yamir Moreno, et al. The unequal effects of the health–economy trade-off during the covid-19 pandemic. *Nature Human Behaviour*, 8(2):264–275, 2024.
- [167] Anna Machens, Francesco Gesualdo, Caterina Rizzo, Alberto E Tozzi, Alain Barrat, and Ciro Cattuto. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC infectious diseases*, 13(1):1–15, 2013.
- [168] Alberto Aleta, Guilherme Ferraz de Arruda, and Yamir Moreno. Data-driven contact structures: from homogeneous mixing to multilayer networks. *PLoS computational biology*, 16(7):e1008035, 2020.
- [169] Joel C Miller and Erik M Volz. Incorporating disease and population structure into models of SIR disease in contact networks. *PloS One*, 8(8):e69162, 2013.
- [170] Marco Ajelli, Bruno Gonçalves, Duygu Balcan, Vittoria Colizza, Hao Hu, José J Ramasco, Stefano Merler, and Alessandro Vespignani. Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. *BMC infectious diseases*, 10(1):1–13, 2010.
- [171] Júlia Koltai, Orsolya Vásárhelyi, Gergely Röst, and Márton Karsai. Reconstructing social mixing patterns via weighted contact matrices from online and representative surveys. *Scientific Reports*, 12(1):1–12, 2022.
- [172] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.
- [173] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science advances*, 6(23):eabc0764, 2020.

- [174] Eleonora Dal Grande, Catherine R. Chittleborough, Stefano Campostriani and Graeme Tucker, and Anne W. Taylor. Health estimates using survey raked-weighting techniques in an australian population health surveillance system. *American journal of epidemiology*, 182(6):544–556, 2015.
- [175] Max Roser. What is the covid-19 stringency index? *Our World in Data*, 2021. <https://ourworldindata.org/metrics-explained-covid19-stringency-index>.
- [176] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.
- [177] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20–28 january 2020. *Eurosurveillance*, 25(5):2000062, 2020.
- [178] Zoltán Vokó, Zoltán Kiss, György Surján, Orsolya Surján, Zsófia Barcza, Bernadett Pályi, Eszter Formanek-Balku, Gergő Attila Molnár, Róbert Herczeg, Attila Gyenesei, et al. Nationwide effectiveness of five sars-cov-2 vaccines in hungary—the hun-ve study. *Clinical Microbiology and Infection*, 28(3):398–404, 2022.
- [179] Julia Shapiro, Natalie E Dean, Zachary J Madewell, Yang Yang, M Elizabeth Halloran, and Ira Longini. Efficacy estimates for various covid-19 vaccines: what we know from the literature and reports. *MedRxiv*, pages 2021–05, 2021.
- [180] Statista. Hungary: Number of people vaccinated against covid-19. <https://www.statista.com/statistics/1196109/hungary-number-of-people-vaccinated-against-covid-19/>, 2023. Accessed: May 9, 2023.
- [181] Henrik Salje, Cécile Tran Kiem, Noémie Lefrancq, Noémie Courtejoie, Paolo Bosetti, Juliette Paireau, Alessio Andronico, Nathanaël Hozé, Jehanne Richet, Claire-Lise Dubost, et al. Estimating the burden of sars-cov-2 in france. *Science*, 369(6500):208–211, 2020.
- [182] Our World in Data. Share of people who completed the initial covid-19 vaccination protocol by age, 2024. Accessed: 2024-07-10.