# The role of social contacts on children's health: Statistical inference on infectious disease data.

Elisabetta De Cao* Emilio Zagheni† Piero Manfredi‡
Alessia Melegaro§

March 3, 2011

**Abstract**

The impact of public health intervention on children's wellbeing critically depends on how individuals mix and the social context in which this mixing occurs. Many countries lack of data on social mixing patterns, and rely on theoretical assumptions on population mixing to evaluate interventions. The aim of this work is to understand how different social interactions affect close-contact childhood infection processes. We propose a model which, by integrating different data sources (time use data and contact surveys), obtains mixing matrices that describe the social structure and reproduce infection profiles. We assume that potentially infectious contacts are proportional to self-reported number of social contacts and/or time of exposure in social activities. To evaluate the uncertainty of model outputs, we use the Bayesian Melding approach. We empirically analyze Italian data, where contact survey, time use data from early ages, and data on close-contact childhood infections are available.

*Department SEFEMEQ, University of Rome "Tor Vergata" and Dondena Centre for Research on Social Dynamics, Bocconi University. Email: elisabetta.decao@unibocconi.it.

†Max Plank Institute

‡University of Pisa

§Dondena Centre for Research on Social Dynamics, Bocconi University.

# 1 Introduction

Social mixing patterns are a relevant explanatory factor for the spread of close-contact infectious diseases (Anderson and May, 1991; Diekmann and Heesterbeek, 2000). The exposure and frequency of contacts between people belonging to different age groups is strongly dependent on demographic and social variables. Population age structure and household size are important demographic determinants of observed contact patterns. Social structure is another crucial element: for instance, the likelihood of interaction between people of different age groups depends on norms that influence the location where they spend time (e.g. work, school, home, restaurant, public transportaton, etc.) and the time slots of the day during which specific activities take place.

Three main approaches have been suggested in the literature to estimate mixing patterns using social data. A first approach relies on contact surveys in which the respondent self reports the number of contacts he has during a randomly sampled day, together with some additional information (i.e., age of contacted person, whether the contact is physical or not, etc.) (Wallinga J., 1999; Edmunds et al., 1997; Wallinga et al., 2006; Beutels et al., 2006; Mossong et al., 2008). A second approach relies on micro-simulation: contact and time of exposure matrices are obtained as output of a simulation informed by secondary data such as transportation data (Del Valle et al., 2007; Iozzi et al., 2010). A third approach relies on time use surveys: time of exposure matrices are estimated from time use diaries, assuming that proportional mixing holds at the level of single location and for short time slots (Zagheni et al., 2008).

In this paper, we first discuss the role of demographic and social structure in shaping mixing patterns. We then propose a model that combines information from time use and contact surveys. To evaluate the uncertainty of model outputs, we use the Bayesian Melding approach (Poole and Raftery, 2000). Finally, we test the ability of the model to fit serologic data and we use the model to evaluate the impact of specific public health interventions.

# 2 Data

We use data for Italy, for which we have one of the most comprehensive and up-to-date collections of social data (time use and contact surveys) and serologic data for varicella zoster virus and parvovirus B19.

Data on time use were collected by the Italian National Statistical Agency (ISTAT) in 2002-2003 on a sample of about 24 thousand households. Time use data were collected in the form of diaries in which the respondent records the activities that he did during the day and the location where the activity took place. We

decided to study Italy, because it is the only country in Europe that collected time use data from early ages (age 3), while most of the other countries have diaries starting from age 10 or 12. This is of extremely importance in our study, given that the childhood is a crucial period for acquiring the diseases considered.

The contact survey for Italy was collected as part of POLYMOD, a project funded by the European Union (see Mossong et al., 2008 for details). A sample of 849 respondents were asked to self report the number of contacts they had during a randomly sampled day, together with some additional information. The survey was conducted between May 2005 and September 2006. The respondents were recruited by random digit dialing using land line telephone. One person of the household was asked to participate and fill paper diaries. Diaries contain also demographic and socio-economic characteristics of the respondents. Participants had to record every person they had contact with between 5 a.m. and 5 a.m. of the following day, in a random day of the week. Contacts were defined as physical contact (skin-to-skin contact such as a kiss or handshake), or nonphysical contact (two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact). Participants had to provide information about the age (or age range) and sex of each contact person. For each contact, participants were asked to record location (home, work, school, leisure, transport, or other), the total duration of time spent together (less than 5 min, 5–15 min, 15 min to 1 h, 1–4 h, or 4 h or more) as well as the frequency of usual contacts with the individual (daily or almost daily, about once or twice a week, about once or twice a month, less than once a month, or for the first time).

The close-contact childhood infections considered are Varicella Zoster Virus (VZV) and parvovirus B19 (PVB19). Known as human herpes virus type 3, VZV causes varicella (or chickenpox), and it mainly occurs in childhood. Afterwards the virus is dormant in the body and may reactivate as herpes zoster (or shingles). Infection with VZV occurs through direct or aerosol contact with infected people. An infected person can transmit the virus for about 7 days. We ignore varicella cases resulting from contact with people who are suffering from zoster virus (Garnett and Grenfell, 1992; Whitaker and Farrington, 2004). Hence, zoster has not a large impact on transmission dynamics when considering large population with no immunization program (Ferguson et al., 1996). The first human parvovirus to be discovered in 1975, PVB19 infection, also known as 5th disease of childhood or slapped cheek syndrome, causes a mild rash illness (Anderson and Cherry, 2004). In adults, especially women, it is often complicated by acute arthritis (Cohen, 1995), and during pregnancy it is associated with intrauterine fetal death, fetal anemia, and hydrops fetalis (Tolfvenstam et al., 2001). Infection with PVB19 occurs through respiratory droplets. An infected person can transmit the virus for about 14 days. There is no vaccine available for PVB19.

In a period from 1997 and 2003, Italian serological samples were collected and tested for antibodies to VZV and PVB19 as part of the European Sero-Epidemiology Network (ESEN2) (Nardone et al., 2007) and POLYMOD (Mossong et al., 2008) projects. The sample size is 2517 and the age of participants ranges from 0 to 79 years. Children under 10 years old were oversampled in each country. In each country the same individuals were tested for both these infections, although they were not the same individuals who filled out the contact diaries. VZV and PVB19 are infections for which mass vaccination program were not in place in Italy. Therefore the data describe the natural history of the disease.

# 3 The Model

In epidemiology, a fundamental quantity is the age-specific force of infection ($\lambda_i$), that is the rate at which susceptible individuals[1] in the age group $i$ become infected. In standard infectious diseases modeling, the force of infection, $\lambda_i$, is proportional to the transmission rates between and within age groups, $\beta_{ij}$:

$$\lambda_i = \sum_j \beta_{ij} \times Y_j \tag{1}$$

where $Y_j$ is the number of infectives at steady state in age group $j$. In particular, $Y_j$ are derived by multiplying the proportion of infected by the population size, $w_j$, and the duration of infectiousness, $d$, and then dividing by the number of years spent in that age band ($a_j$ is an age band), as follows:

$$Y_j = [S(a_j) - S(a_{j-1})] \times w_j \times \frac{d}{(a_j - a_{j-1})}$$

Here the proportion of susceptible individuals in each age band $a_i$ is given by $S(a_i) = S(a_{i-1}) \times \exp\{-\lambda_i(a_i - a_{i-1})\}$

Traditionally, the transmission rates, which form the "who-acquires-infection-from-whom" matrix, are estimated 'indirectly' from epidemiological data, under suitable simplifying assumptions (Anderson and May, 1991). More recently, 'direct' approaches have been suggested: the transmission rates matrix is assumed to be proportional to either a contact matrix or a time of exposure matrix estimated from sample surveys (Wallinga et al., 2006; Zagheni et al., 2008).

[1] A susceptible individual (sometimes simply susceptible) is a member of a population who is at risk of becoming infected if she is exposed to the infectious agent, because he is naive to the infection or has lost his immunity.

## 3.1 Case A

To compare our study with the current literature, we consider as benchmark a transmission rate matrix proportional to a contact matrix. As in Melegaro et al. (2011), age-specific transmission parameters are estimated by multiplying each element of the social contact matrix $C$ by a proportionality factor $q$ which measures the disease-specific infectivity:

$$\beta_{ij} = qc_{ij} \tag{2}$$

This generates a "next generation matrix" $N$:

$$n_{ij} = \beta_{ij} \times w_j \times d \tag{3}$$

which provides the potential number of transmission events per person.[2] Following Diekmann et al. (1990) and Heesterbeek (1992), we obtain $R_0$, the basic reproduction number, as the leading eigenvalue of $N$:

$$\det(N - R_0 I) = 0 \tag{4}$$

$R_0$ is the average of the number of transmission events, or the average number of secondary infectious persons resulting from a single infectious person following his/her introduction into a totally susceptible population.

Hence, $q$ is the parameter that need to be estimated given the contact matrix and the proportions of samples testing positive in serological data for VZV and PVB19.

## 3.2 Case B

In the case of Italy, we have two independent data sources that give us information on time of exposure (i.e., time use survey) and number of contacts (i.e., contact survey). In this paper we propose a new model that combines these two data sources.

We assume that a fraction $q_2$ of the average time of exposure between groups $i$ and $j$, $(e_{ij})$, is suitable for transmission of the disease in terms of proximity of contact, physical condition of the location of contact, etc. People in the age groups $i$ and $j$ have a certain number of daily contacts on average, $(c_{ij})$, which differ in terms of duration. If we assume that a person randomly distributes her/his suitable minutes for transmission to people she/he has contact with, then some people may receive more than one suitable minute, whereas some others may not receive any of them. If the disease is highly transmissible, what matters for transmission between two people is that they have a contact with at least one suitable minute

[2] At endemic equilibrium, $N$ determines the force of infection $\lambda_i$ as in equation (1).

of exposure. We call this kind of contact 'suitable contact'. We do not know what kind of contacts are 'suitable', but we assume that the likelihood that people experience at least one such contact is positively related to the duration of their contacts.

Setting up our problem in these terms, we can use some results from classic probability problems, such as the 'occupancy problem'[3], to obtain the expected number of suitable contacts between age groups $i$ and $j$, $(u_{ij})$.

$$E[u_{ij}] = c_{ij}(1 - e^{-q_2 e_{ij}/c_{ij}}) \tag{5}$$

If we assume that the age-specific transmission rates are proportional to the age-specific number of suitable contacts, then, by multiplying the quantity in expression 5 by a parameter $q_1$ that represents a disease-specific infectivity parameter, we obtain:

$$\beta_{ij} = q_1 \times c_{ij}(1 - e^{-q_2 e_{ij}/c_{ij}}) \tag{6}$$

The parameters $q_1$ and $q_2$ can be interpreted as 'level' and 'shape' parameters, respectively. In this setting, high values of $q_2$ give little importance to the exposure matrix and more importance to the contact matrix. The level parameter $q_1$ then rescales the structure of suitable contacts to account for the degree of infectivity of the disease.

The force of infection can be obtained by plugging equation (6) into equation (1), while $R_0$ is the eigenvalue of the next generation matrix obtained plugging equation (6) into equation (3).

The parameters to estimate from serologic, time use, and contact data are $q_1$ and $q_2$. Instead of using a maximum likelihood technique, we adopt the so called Bayesian Melding approach.

# 4 Bayesian Melding approach

To estimate the parameters of interest we use the Bayesian Melding approach. The purpose of this approach is to take into full account information and uncertainty about inputs and outputs (Poole and Raftery, 2000).[4] For our application, the Bayesian Melding is the following. Consider a deterministic model $M$ that transforms a set of inputs $\theta$ into a set of outputs $\rho$: $\rho = M(\theta)$. The knowledge about the problem under study is translated into probabilistic statements, and hence, in

[3] A note on the occupancy problem is reported in the Appendix.

[4] Usually, the bayesian melding approach considers inputs and outputs of a deterministic model. Melegaro et al. (2011) are working on the specification of deterministic models for the spread of the close-contact infectious diseases, such as VZV and parvovirus B19. Therefore, for the moment we do not consider the deterministic models.

"direct" prior distributions for inputs and outputs: $p(\theta)$, $p(\rho)$. In this paper, the output is $R_0$, the inputs are $\theta = (q)$ in Case A, and $\theta = (q_1, q_2)$ in Case B. The model $M(\theta)$ is given by equations (2) ((6) in Case B), (3) and (4).

The prior distribution on the inputs implicitly defines a prior distribution on the outputs. The same way, a prior distribution on the outputs implicitly defines a prior distribution on the inputs. These implicitly defined priors are the so-called "induced" prior distributions: $p^*(\theta)$, $p^*(\rho)$. The literature provided more evidence on the output we are considering, the $R_0$, therefore we will apply the Bayesian Melding focusing on the induced prior distributions for inputs $p^*(\theta)$. Poole and Raftery (2000) propose a way to combine the two sets of priors, through logarithmic pooling:

$$\tilde{p}(\theta) \propto p^*(\theta)^{\alpha} p(\theta)^{1-\alpha}$$

Since we are going to use uniform distribution on intervals for the direct priors, we let $\alpha \uparrow 1$, and we obtain $\tilde{p}(\theta) \propto p^*(\theta)$.[5]

We define $W$ as the serologic data available. From the serological data we obtain a likelihood, $p(W|M(\theta))$, for the input $\theta$. We assume the prevalence to be equal to the seroprevalence, and we obtain the log-likelihood:

$$\log p(W|M(\theta)) = \sum_{i=1}^{N} \{Y_i \log(\pi(a_i)) + (1 - Y_i) \log(1 - \pi(a_i))\}$$

where $N$ is the size of the serological data set, $Y_i$ is a binary variable indicating if subject $i$ had experienced infection before age $a_i$, and the prevalence of immune individuals[6] is $\pi(a_i) = \Pr(Y_i = 1|a_i)$.

The posterior distributions for the parameters is obtained by combining priors and likelihoods (using the Bayes theorem):

$$p(\theta|W) \propto \tilde{p}(\theta) p(W|M(\theta))$$

When finding analytical solutions is not a viable option, the Sampling-Importance-Resampling algorithm (Rubin, 1987, 1988) is used to computationally calculate the posterior distributions. Considering the posterior distribution for the inputs, the algorithm works in four steps:

1. Sample $\{\rho^{(1)}, \ldots, \rho^{(n)}\}$ from the input prior $p(\rho)$ on $\rho$.

2. For each $\rho^{(i)}$ determine the corresponding series of inputs, $\theta^{(i)} = M^{-1}(\rho^{(i)})$,

[5]The limiting pooled prior obtained by setting $\alpha = 1$ will not be the same as $\alpha \uparrow 1$.

[6]The prevalence of immune individuals corresponds to the prevalence of individuals infected, since it is a SIR model.

by running the deterministic model.[7] This produces a sample of the induced priors on the inputs, $p^*(\theta)$.

3. Find the sampling importance weights for each $\theta^{(i)}$ based on the likelihood function.

4. Sample from the prior distribution $\{\theta^{(1)}, \ldots, \theta^{(n)}\}$ with probabilities equal to the weights to approximate the posterior distribution for the inputs.

## 4.1 Bayesian melding for Case A

In the benchmark case we consider only the contact data, and there is only one parameter to estimate: $q$. The direct priors are: for the input $q \sim U(0,1)$ and for the output $R_0 \sim U(1,8)$. We sample 100 different $R_0$'s and through our model we find set of values for $q$. The induced prior becomes $q \sim U(0.006, 0.054)$. The posterior median for $q$ is equal to 0.031. I recover the posterior distribution for R0, and its median is 4.653. AIC and BIC are respectively equal to 1242.722 and 1245.092.

## 4.2 Bayesian melding for Case B

In Case B we combine contact data and time use data, the input parameters to estimate are two: $\theta = (q_1, q_2)$. The direct priors for the inputs are $q_2 \sim U(0,1)$, and $q_1 \sim U(0,10)$, and for the output $R_0 \sim U(1,8)$. We know that the fraction $q_2$ is between 0 and 1, we want $R_0$ to be between 1 and 8. Hence we sample from 10000 possible combinations of $q_2$ and $R_0$ to obtain set of values for $q_2$. We find induced priors for the inputs equal to $q_1 \sim U(0.7, 5.89)$, and $q_2 \sim U(0.017, 0.3)$. The posterior median for $q_1$ is equal to 4.735, while for $q_2$ is 0.216. The posterior median for R0 is 6.265. AIC and BIC for the posterior median are respectively equal to 1261.968 and 1266.707.

[7] The model $M(\theta)$ is invertible. We can prove for Case A that:

$$\begin{aligned} R0 &= \text{eigenvalue}(N) \\ &= \text{eigenvalue}(q \times c_{ij} \times w_j \times d) \\ &= q \times d \times \text{eigenvalue}(c_{ij} \times w_j) \end{aligned}$$

Thus:

$$q = \frac{R0}{d} \times \frac{1}{\text{eigenvalue}(c_{ij} \times w_j)}$$

# 5 Preliminary Results and Discussion

Figure 1 shows a contour plot of the estimated average daily time that people in the age group $i$ spend with people in the age group $j$ in Italy. The estimates are obtained, respectively, by using the approach developed in (Zagheni et al., 2008) on Italian time use data and from the contact survey for Italy (Mossong et al., 2008). In both cases, the highest values are on the main diagonal, implying assortativeness. In the case of the contact matrix, the highest values are more concentrated along the main diagonal, compared to the estimated time of exposure.

Figures 2 and 3 show the fit of the model to serologic data, based on posterior medians for the parameters $q$ in the Case 1 where only contact data are used, and for $q_1$ and $q_2$ in the Case 2 where also time use data are used: we observe a good fit to the seroprevalence data.

Both data on time use and number of contacts are available for single locations (e.g., school, home, workplace): we can thus use this information to obtain a more detailed representation of contact patterns and we can write the elements of the $\beta$ matrix as a combination of suitable contacts in the $n$ different settings considered:

$$\beta_{ij} = q_1 \times [c_{ij,1}(1 - e^{-q_{2,1}e_{ij,1}/c_{ij,1}}) + \cdots + c_{ij,n}(1 - e^{-q_{2,n}e_{ij,n}/c_{ij,n}})] \qquad (7)$$

The representation in equation 7 allows us to evaluate the impact of specific interventions such as school closure (e.g., elimination of the setting 'school' and adjustment of the other settings to levels observed during vacation time) or behavioral changes (e.g., modification of the $q_2$ parameter for a specific setting). The effect of these interventions will be discussed in the paper.

# 6 Figures

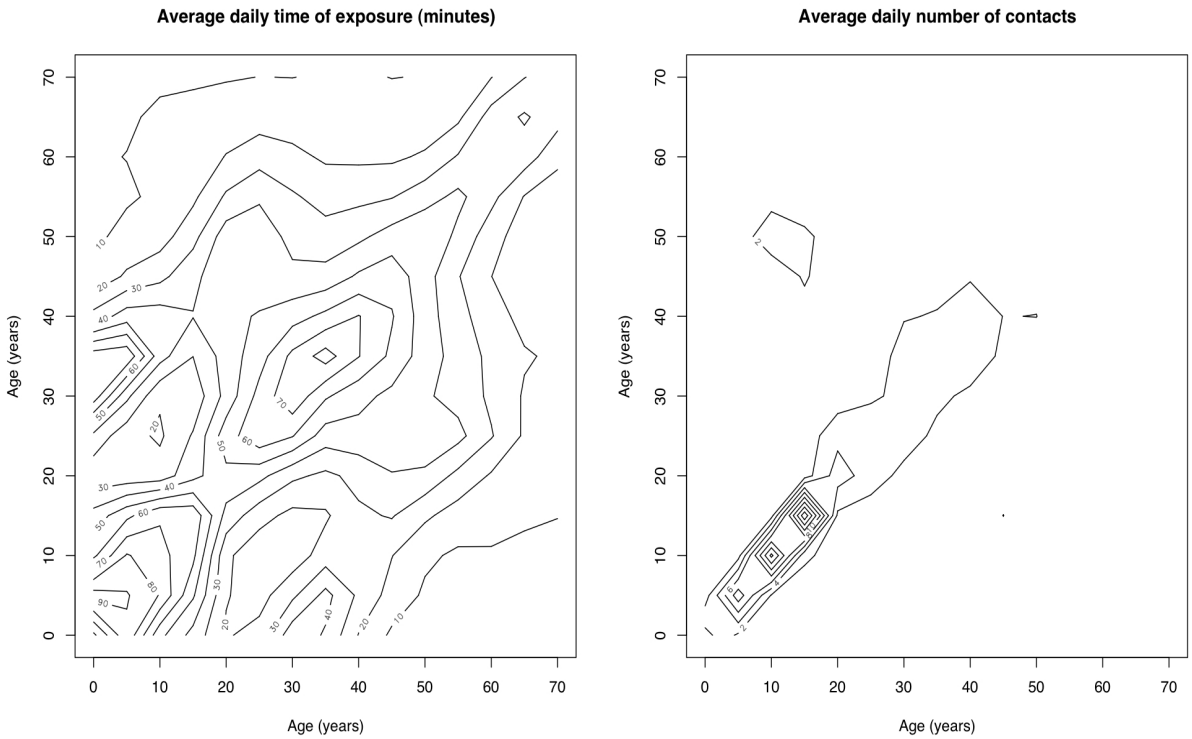


Figure 1: Mixing patterns for Italy estimated from time use and contact surveys.

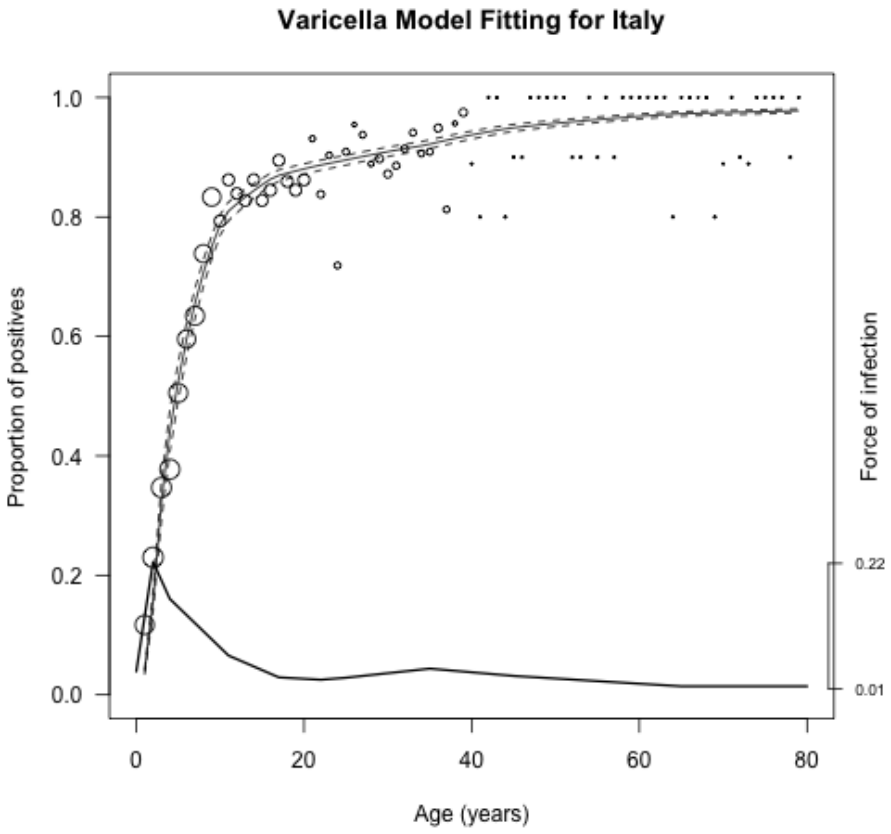


Figure 2: Case 1: contact data. Fit of the model to serologic data for Italy. Points are serologic data with size proportional to the corresponding sample size; solid line is the median of the posterior distribution; dashed lines are the 2.5% and 97.5% quantiles of the posterior distribution. The solid line at the bottom of the graph is the force of infection with its minimum and maximum values reported.

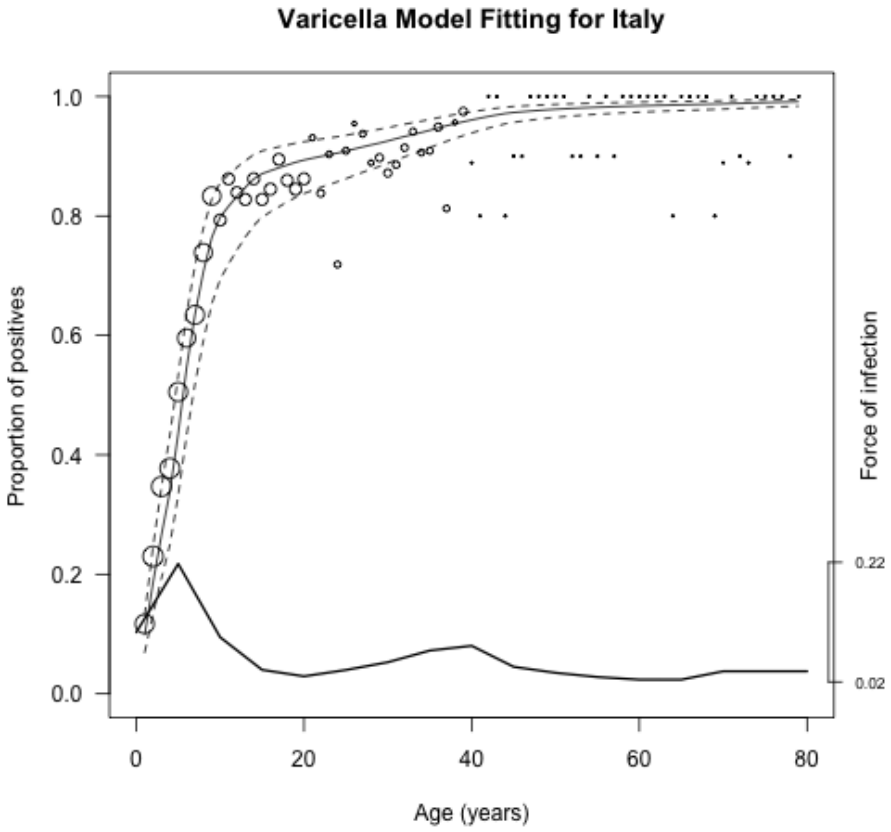


Figure 3: Case 2: combination of contact data and time use data. Fit of the model to serologic data for Italy. Points are serologic data with size proportional to the corresponding sample size; solid line is the median of the posterior distribution; dashed lines are the 2.5% and 97.5% quantiles of the posterior distribution. The solid line at the bottom of the graph is the force of infection with its minimum and maximum values reported.

# References


Anderson, M. and J. Cherry, 2004. *Textbook of pediatric infectious diseases*, chapter 17. Philadelphia, PA: Sauders.

Anderson, R. M. and R. M. May, 1991. *Infectious diseases of humans: Dynamics and control*. Oxford, United Kingdom: Oxford University Press.

Beutels, P., Z. Shkedy, M. Aerts, and et al., 2006. Social mixing patterns for transmission models of close contact infections: Exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiol Infect*, 134(6):1158–66.

Cohen, B., 1995. Parvovirus B19: An expanding spectrum of disease. *British Medical Journal*, 311:1549–1552.

Del Valle, S. Y., J. M. Hyman, H. W. Hethcote, and et al., 2007. Mixing patterns between age groups in social networks. *Soc Networks*, 29:539–554.

Diekmann, O., J. A. Heesterbeek, and J. A. Metz, 1990. On the definition and the computation of the basic reproduction ratio $R_0$ in models for infectious diseases in heterogeneous populations. *J Math Biol*, 28(4):365–382.

Diekmann, O. and J. A. P. Heesterbeek, 2000. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. New York, NY: John Wiley and Sons, Inc.

Edmunds, W. J., C. J. O'Callaghan, and D. J. Nokes, 1997. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proc R Soc Lond B*, 264:949–957.

Ferguson, N. M., R. M. Anderson, and G. P. Garnett, 1996. Mass vaccination to control chickenpox: the influence of zoster. *Proceedings of the National Academy of Science of the United States of America*, 93:7231–7235.

Garnett, G. P. and B. T. Grenfell, 1992. The epidemiology of varicella-zoster virus infections: A mathematical model. *Epidemiology and Infection*, 108:495–511.

Heesterbeek, J. A., 1992. $R_0$. Ph.D. thesis, University of Leiden.

Iozzi, F., F. Trusiano, F. C. Billari, and et al., 2010. Little Italy: An agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. *PLoS Computational Biology*, 6(12).

Melegaro, A., M. Jit, E. Zagheni, and et al., 2011. What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Epidemics (forthcoming)*.

Mossong, J., N. Hens, and M. e. a. Jit, 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *Plos Med*, 5(3):74.

Poole, D. and A. E. Raftery, 2000. Inference for deterministic simulation models: The Bayesian melding approach. *J. Amer. Statist. Assoc*, 95:1244–1255.

Tolfvenstam, N., T. Papadogiannakis, O. Norbeck, and et al., 2001. Frequency of human parvovirus B19 in intrauterine fetal death. *Lancet*, 357:1494–1497.

Wallinga, J., P. Teunis, and M. Kretzschmar, 2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol.*, 164:936–944.

Wallinga J., K. M., Edmunds W. J., 1999. Human contact patterns and the spread of airborne infectious diseases. *Trends Microbiol.*, 7:372–377.

Whitaker, H. L. and C. P. Farrington, 2004. Infections with varying contact rates: Application to varicella. *Biometrics*, 60:615–623.

Zagheni, E., F. C. Billari, P. Manfredi, and et al., 2008. Using time use data to parameterize models for the spread of close-contact infectious diseases. *American Journal of Epidemiology*, 168(9):1082–1090.

# A Occupancy problem

The result can be derived as follows. If we think of the number of suitable minutes of contact between groups $i$ and $j$, $(q_2 e_{ij})$, as 'balls', and the number of contacts between groups $i$ and $j$, $(c_{ij})$, as 'boxes', then the expected number of suitable contacts between the two age groups can be thought of as the expected value of occupied boxes from randomly assigned balls.

To compute this expected value, define the indicator function

$$Z_i = \begin{cases} 1 & \text{if the contacted person } i \text{ receives } \textit{zero} \text{ suitable minutes} \\ 0 & \text{otherwise} \end{cases}$$

We have

$$E[Z_i] = Pr[Z_i = 1] = (1 - \frac{1}{c_{ij}})^{q_2 e_{ij}} \approx e^{-q_2 e_{ij}/c_{ij}}$$

Consider now $Z = \sum_{i=1}^{c_{ij}} Z_i$. The variable $Z$ is the total number of contacted people who do not receive any suitable minute of transmission. Its expected value is:

$$E[Z] = \sum_{i=1}^{c_{ij}} E[Z_i] \approx \sum_{i=1}^{c_{ij}} e^{-q_2 e_{ij}/c_{ij}} = c_{ij} e^{-q_2 e_{ij}/c_{ij}}.$$