Epidemics 11 (2015) 14–23

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

The social contact hypothesis under the assumption of endemic equilibrium: Elucidating the transmission potential of VZV in Europe

E. Santermans[a,*], N. Goeyvaerts[a,b], A. Melegaro[c], W.J. Edmunds[d], C. Faes[a], M. Aerts[a], P. Beutels[b,e], N. Hens[a,b]

[a] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[b] Centre for Health Economic Research and Modelling Infectious Diseases, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium
[c] Department of Policy Analysis and Public Management and Dondena Centre for Research on Social Dynamics, Universit Commerciale L. Bocconi, Milan, Italy
[d] London School of Hygiene & Tropical Medicine, London, United Kingdom
[e] School of Public Health and Community Medicine, The University of New South Wales, Sydney, Australia

ARTICLE INFO

Article history:
Received 14 June 2014
Received in revised form
23 December 2014
Accepted 30 December 2014
Available online 10 January 2015

Keywords:
Mathematical model
Mixing
Contact data
Varicella
Risk factors

ABSTRACT

The basic reproduction number R_0 and the effective reproduction number R are pivotal parameters in infectious disease epidemiology, quantifying the transmission potential of an infection in a population. We estimate both parameters from 13 pre-vaccination serological data sets on varicella zoster virus (VZV) in 12 European countries and from population-based social contact surveys under the commonly made assumptions of endemic and demographic equilibrium. The fit to the serology is evaluated using the inferred effective reproduction number R as a model eligibility criterion combined with AIC as a model selection criterion. For only 2 out of 12 countries, the common choice of a constant proportionality factor is sufficient to provide a good fit to the seroprevalence data. For the other countries, an age-specific proportionality factor provides a better fit, assuming physical contacts lasting longer than 15 min are a good proxy for potential varicella transmission events. In all countries, primary infection with VZV most often occurs in early childhood, but there is substantial variation in transmission potential with R_0 ranging from 2.8 in England and Wales to 7.6 in The Netherlands. Two non-parametric methods, the maximal information coefficient (MIC) and a random forest approach, are used to explain these differences in R_0 in terms of relevant country-specific characteristics. Our results suggest an association with three general factors: inequality in wealth, infant vaccination coverage and child care attendance. This illustrates the need to consider fundamental differences between European countries when formulating and parameterizing infectious disease models.

Introduction

One of the key measures of infectious disease transmission is the basic reproduction number R_0: the expected number of secondary cases per primary case in a "virgin" population (Diekmann et al., 1990). If R_0 is larger than 1 the infection may become endemic and the larger R_0, the more effort is required to eliminate the infection from the population. Although R_0 is a useful theoretical measure, it is rarely observed in practice. The effective reproduction number R takes pre-existing immunity into account and thus reflects the average number of secondary cases that can be observed in a partially immune population. There are several methods to estimate R_0 and R (Vynnycky and White, 2010). In this article, we focus on deriving R_0 from transmission rates that can be estimated from serological data under the assumption of endemic equilibrium (Anderson and May, 1991). A disease in endemic equilibrium, or steady state, may undergo cyclical epidemics, but fluctuates around a stationary average over time. Whitaker and Farrington (Whitaker and Farrington, 2004a) have shown that the impact of regular epidemic cycles, displayed by many childhood infections, can be ignored when estimating R_0. In this equilibrium setting, each infectious individual infects one other individual on average, hence

* Corresponding author. Tel.: +32 474927849.
E-mail address: eva.santermans@uhasselt.be (E. Santermans).

http://dx.doi.org/10.1016/j.epidem.2014.12.005

R is expected to be equal to 1 (Diekmann et al., 1990). Again, if $R > 1$ the infection will continue to spread in the population whereas if $R < 1$ the infection will die out.

We consider pre-vaccination serological data for the varicella zoster virus (VZV) from 12 different European countries (Nardone et al., 2007). VZV is one of the eight known herpes viruses that affect humans. Primary infection with VZV results in varicella (chickenpox) and mainly occurs in childhood. In general, the disease is benign, however, symptoms may be more severe in adults and complications may occur when varicella is acquired during pregnancy. VZV is highly contagious and transmitted through direct close contact with lesions or indirectly through air droplets containing virus particles. The incubation period following VZV infection ranges from 13 to 18 days and each infected person transmits the virus for about 7 days. The antibody response following primary infection with VZV is believed to induce lifelong protection against chickenpox. However, the virus remains dormant within the body and may reactivate and give rise to herpes zoster (or shingles) after years to decades (Miller et al., 1993). In this article, we will focus on primary infection and ignore reactivation leading to zoster.

Estimating transmission rates for an airborne infection such as VZV requires assumptions on the underlying age-specific mixing patterns and R_0 has been shown to be highly sensitive to these assumptions (Greenhalgh and Dietz, 1994). Indeed, serological surveys do not provide complete information about these mixing patterns, since they reflect the rate at which susceptible individuals become infected, but not who is infecting whom. This indeterminacy prevents assessment of the validity of the mixing pattern. Recently, attempts have been made to deal with this unidentifiability by exploiting knowledge about the route of transmission (Farrington et al., 2001; Unkel et al., 2014). However, this relies on the strong assumption that infections are transmitted via the same route. The extent to which different routes of transmission compete may only be verified by additional data collection. In this article, we inform the mixing pattern with data from population-based social contact surveys and assume that transmission rates are proportional to contact rates. Social contact data have already proven to be a valuable additional source of information when estimating the 'Who Acquires Infection From Whom' (WAIFW) matrix and R_0 (see e.g. Wallinga et al., 2006; Ogunjimi et al., 2009; Goeyvaerts et al., 2010).

We use the inferred effective reproduction number as a model eligibility criterion combined with AIC as a model selection criterion. To our knowledge, Wallinga et al. (2001) were the first to use the effective reproduction number to asses the plausibility of different mixing patterns. However, this is the first time that R is explicitly used as a determinant in the model selection procedure. We evaluate how constant and age-specific proportionality factors affect the fit to the serology and the estimated R_0 values. Moreover, we assess the effect of age-specific heterogeneity related to infectiousness on model eligibility and fit. Further, from a selected set of demographic, socio-economic and spatio-temporal factors, we explore which factors best explain the between-country heterogeneity in R_0 using two non-parametric methods: the maximal information coefficient (MIC) and random forest.

The article is organized as follows. In Section "Materials and methods", a description of the serological and social contact surveys is provided, after which we elaborate on the dynamic model structure, estimation procedure and methods used to determine potential risk factors for varicella. In Section "Results", we present the estimates of R_0 and R under various model assumptions, and the results of the risk factor analysis. Finally in Section "Discussion", the models and results are discussed.

Materials and methods

Data

Serological data. In this article, we reanalyze the ESEN2 (European Sero-Epidemiology Network) data on VZV published by Nardone et al. (2007) together with newly available serology for Poland and Italy, totaling 13 serosurveys from 12 different countries including two samples from Italy (see Table 1). At the time of sera collection, which varied between 1995 and 2004, none of the participating countries had introduced a universal VZV vaccination program. Blood samples were tested using an enzyme-linked immunosorbent assay (ELISA), thereby classifying the samples as seropositive or seronegative (equivocal results were included as seropositive). Classification is based on the observed antibody level as compared to the cut-off level specified by the manufacturer of the test. Sample sizes range from 1268 for Poland to 4398 for Germany, with substantial variability between the surveyed age ranges.

Social contact survey. The spread of airborne or close-contact infections in a population is driven by social contacts between individuals. Recently, several studies were conducted to measure social mixing behavior, and Read et al. (2012) present a review of the different methodologies employed. The cross-sectional diary-based surveys that were conducted between May 2005 and September 2006 as part of the POLYMOD project, constituted the first large-scale prospective study to investigate social contact behavior in eight European countries (Mossong et al., 2008). Participants were recruited through random-digit dialing, face-to-face interviews or population registers, and completed a diary about their social contacts during one randomly assigned day. Participants were asked to record the age and gender of each contacted person, plus location, duration and frequency of the contact. Further, a distinction between two types of contact was made: non-close contacts, defined as two-way conversations of at least three words in each others proximity, and close contacts that involve any sort of physical skin-to-skin touching. For an extensive description of the survey, we refer to Mossong et al. (2008).

Estimating the basic and effective reproduction number

Force of infection and mass action principle. To describe VZV transmission dynamics, a compartmental MSIR (Maternal protection-Susceptible-Infected-Recovered) model for a closed population of size N with fixed duration of maternal protection A is considered, following Goeyvaerts et al. (2010) and Ogunjimi et al. (2009). Doing so, we explicitly take into account the fact that newborns are protected by maternal antibodies and do not take part in the transmission process. We assume that mortality due to infection can be ignored, which is plausible for VZV in developed countries, and that infected individuals maintain lifelong immunity to varicella after recovery. Further, demographic and endemic equilibria are assumed, which means that the age-specific population sizes remain constant over time and that the disease is in an endemic steady state at the population level. Under these assumptions the age-specific prevalence $\pi(a)$ is given by:

$$\pi(a) = 1 - e^{-\int_A^a \lambda(u)\,\mathrm{d}u},$$

where $\lambda(a)$ is the age-specific force of infection, i.e. the rate at which a susceptible person of age a acquires infection. There is a wide range of methods available to estimate $\lambda(a)$ from seroprevalence data, see Hens et al. (2010) for an historical overview.

Since we aim to estimate the basic and effective reproduction number for VZV, we disentangle the force of infection further to the level of age-specific transmission rates. Let $\beta(a, a')$ denote the

Table 1
Overview of the serological data and demographic parameters.

Country	Data collection	Age range (years)	Sample size	Life expectancy (years)	Population size
Belgium (BE)	2001–2003	0–71.5	3251	77.6	10,309,722
Germany (DE)	1995/1998	0–79	4398	77.1	82,050,377
Spain (ES)	1996	2–39	3590	77.5	39,427,919
England and Wales (EW)	1996	1–20.9	2032	76.0	51,125,400
Finland (FI)	1997–1998	1–79.8	2471	76.7	5,146,965
Ireland (IE)	2003	1–60	2430	77.6	3,963,814
Israel (IL)	2000–2001	0–79	1543	76.2	6,223,842
Italy (IT'97)	1996–1997	0.1–50	3110	78.2	56,872,349
Italy (IT'04)	2003–2004	1–79	2446	80.3	5,788,0478
Luxembourg (LU)	2000–2001	4–82	2640	77.2	438,723
The Netherlands (NL)	1995–1996	0–79	1967	77.0	15,493,889
Poland (PL)	1995–2004	1–19	1268	73.2	38,637,184
Slovakia (SK)	2002	0–70	3515	73.2	5,378,702

average per capita rate at which an infectious individual of age a' makes effective contact with a susceptible person of age a, per unit time. The key principle behind the estimation of $\beta(a, a')$ is the so-called mass action principle. If the mean infectious period D is short compared to the timescale on which transmission and mortality rates vary, the force of infection can be approximated by:

$$\lambda(a) \approx \frac{ND}{L} \int_A^\infty \beta(a, a')\lambda(a')s(a')m(a')\,\mathrm{d}a', \tag{1}$$

where N denotes the total population size, L the life expectancy, $s(a)$ the proportion of people in the population of age a that are susceptible, and $m(a) = \exp\{-\int_0^a \mu(t)\,\mathrm{d}t\}$ the survivor function at age a with age-specific mortality $\mu(a)$.

Given the transmission rates $\beta(a, a')$, following Diekmann et al. (1990), the basic reproduction number R_0 can be calculated as the dominant eigenvalue of the next generation operator given by

$$G(a, a') = \frac{ND}{L} m(a)\beta(a, a').$$

The effective reproduction number R takes into account the proportion of susceptible individuals and is the dominant eigenvalue of $G(a, a') \times s(a)$.

Mixing assumptions. Since $\lambda(a)$ is a one-dimensional function of age and $\beta(a, a')$ makes up a two-dimensional function, additional assumptions are necessary to estimate the transmission rates from seroprevalence data using the mass action principle. The traditional approach of Anderson and May (Anderson and May, 1991) stratifies the population into a small number of age classes and imposes different mixing patterns upon $\beta(a, a')$. This is the approach taken in the exploratory analysis of Nardone et al. (2007). However, the choice of the structure imposed on the WAIFW matrix as well as the choice of the age classes are ad hoc and impact the estimation of R_0 (Greenhalgh and Dietz, 1994; Van Effelterre et al., 2009). We will consider a more recent approach as proposed by Wallinga et al. (2006), by informing $\beta(a, a')$ with data on social contacts. This is also the approach taken by Goeyvaerts et al. (2010) who express $\beta(a, a')$ as

$$\beta(a, a') = q(a, a') \cdot c(a, a'),$$

where $c(a, a')$ is the per capita rate at which an individual of age a' makes contact with a person of age a, per unit of time, and $q(a, a')$ a proportionality factor that may capture, among other effects, age-specific susceptibility and infectivity.

In this article, we contrast the constant proportionality assumption, commonly used in the literature and referred to as "the social contact hypothesis" (Wallinga et al., 2006; Ogunjimi et al., 2009; Melegaro et al., 2011), against a log-linear function of the age of the susceptible individual, which entailed an improvement of model fit for VZV in Belgium (Goeyvaerts et al., 2010), that is respectively:

$$\log\{q(a, a')\} = \gamma_0 \quad \text{and} \quad \log\{q(a, a')\} = \gamma_0 + \gamma_1 a. \tag{2}$$

The contact rates $c(a, a')$ are estimated from the POLYMOD contact survey using a bivariate smoothing approach, considering those contacts with skin-to-skin touching lasting at least 15 min since these contacts have been shown to be most predictive for VZV (Goeyvaerts et al., 2010; Melegaro et al., 2011). For the countries who participated in the POLYMOD project, the corresponding contact rates were used, whereas for the other countries contact data of a neighboring country or a country with similar school enrollment ages were used (cf. Table 3 in Supplementary Material). We present a sensitivity analysis in the Supplementary Material to compare these ad-hoc choices with a more objective selection of contact data by means of AIC. In this analysis, we repeat the estimation procedure for each country seven times, each time for a different contact matrix, and select, per country, from these seven analyses, the one that results in the best fit to the serological data. We observe that the effect on R_0 remains within reasonable bounds, which indicates that the choice of contact data has limited influence on our estimates.

Estimation procedure. In this article we will estimate the force of infection using maximum likelihood estimation with the Bernouilli log-likelihood given by:

$$\ell(\lambda; \mathbf{y}, \mathbf{a}) = \sum_{i=1}^{n} y_i \log(1 - e^{-\int_A^{a_i} \lambda(u)\,\mathrm{d}u}) + (1 - y_i)(-\int_A^{a_i} \lambda(u)\,\mathrm{d}u). \tag{3}$$

Here, n denotes the size of the serological data set and y_i denotes a binary variable indicating whether subject i had experienced infection before age a_i. The transmission rates cannot be estimated analytically since the integral Eq. (1) has no closed form solution. However, it is possible to solve this numerically by turning to a discrete age framework, assuming a constant force of infection in each 1-year age interval. Now, estimation proceeds as follows: starting values for the parameters are provided after which the discretized mass action principle is iterated until convergence and finally, the resulting estimate of the force of infection is contrasted to the serology using the log-likelihood (3). To calculate 95% confidence intervals, non-parametric bootstraps are performed on both the contact data and the serological data to account for all sources of variability (Goeyvaerts et al., 2010). The number of bootstrap samples per country is fixed at 2000 with convergence rates varying between 62% and 100%.

Since some countries lack serological data on VZV in the older age groups, the original serology is augmented with simulated data to avoid excess variability of the bootstrap estimates (Goeyvaerts

et al., 2010). These simulations are drawn from a Bernouilli distribution with mean equal to the seroprevalence from the last 5 age categories with at least 20 observations available. The size of the simulated samples is determined by the demography of the population. This method is plausible from an epidemiological point of view since the VZV seroprofile is not expected to decline after 20 years of age. Based on the augmented data, post-stratification weights are calculated using census data and included in the likelihood. The life expectancy L and the age-specific mortality rates μ for every country are estimated based on demographic data from the year of serological data collection (Eurostat, the Office for National Statistics for England and Wales, Israeli Bureau of Statistics for Israel) using a Poisson model with log link and offset term (Hens et al., 2012). To ensure flexibility, a radial basis spline is used.

The duration of maternal immunity is fixed at $A = 0.5$ years, while the mean duration of infectiousness for VZV is taken as $D = 7/365$ years. Lastly, to reduce boundary irregularities induced by sparseness in the contact data for the elderly, the contact surface, and hence the serological data, are restricted to the 0–69 year age range. A sensitivity analysis showed little impact on the point estimates (results not shown).

Model eligibility and indeterminacy. The estimated effective reproduction number $\hat{R}$ and corresponding confidence interval allow us to check whether the above mixing patterns (2) conform with the assumption of endemic equilibrium. In this setting, each infectious individual infects one other individual on average, hence R is expected to be equal to 1 (Farrington, 2003). We use this property to exclude those models for which R is estimated to be significantly different from 1. Furthermore, the effective reproduction number allows us to make indirect inference about the age-specific heterogeneity related to infectiousness, assuming

$$\log\{q(a, a')\} = \gamma_0 + \gamma_1 a + \gamma_2 a', \tag{4}$$

where a' is the age of the infective individual. We refer to this model as the extended log-linear model, in which γ_2 is referred to as an infectiousness component. Direct inference can be troublesome, as shown by Goeyvaerts et al. (2010), since serological surveys do not provide information related to infectiousness. This indeterminacy can be illustrated as follows: assume for simplicity $\beta(a, a') = q(a, a')c(a, a') = q_0 q_1(a) q_2(a') c(a, a')$. Rewriting (1), this implies

$$q_0 q_1(a) = \frac{L\lambda(a)}{ND\int_A^\infty q_2(a')c(a, a')\lambda(a')s(a')m(a')\,da'},$$

where $\lambda(a)$, $s(a)$ and $c(a, a')$ can be estimated from serological data and social contact data, respectively. This implies that when $q_0 q_1(a)$ is flexibly modeled, the effect of $q_2(a')$ on the serological model is completely absorbed and the fit of this model does not change for varying infectivity curves. However, it does affect the estimated value of R_0 and R. We deal with this indeterminacy by letting γ_2 vary over a fixed interval and assessing the effect on $\hat{R}$. This way, the value of γ_2 can be determined such that R is not significantly different from 1. This is illustrated in Section "Results".

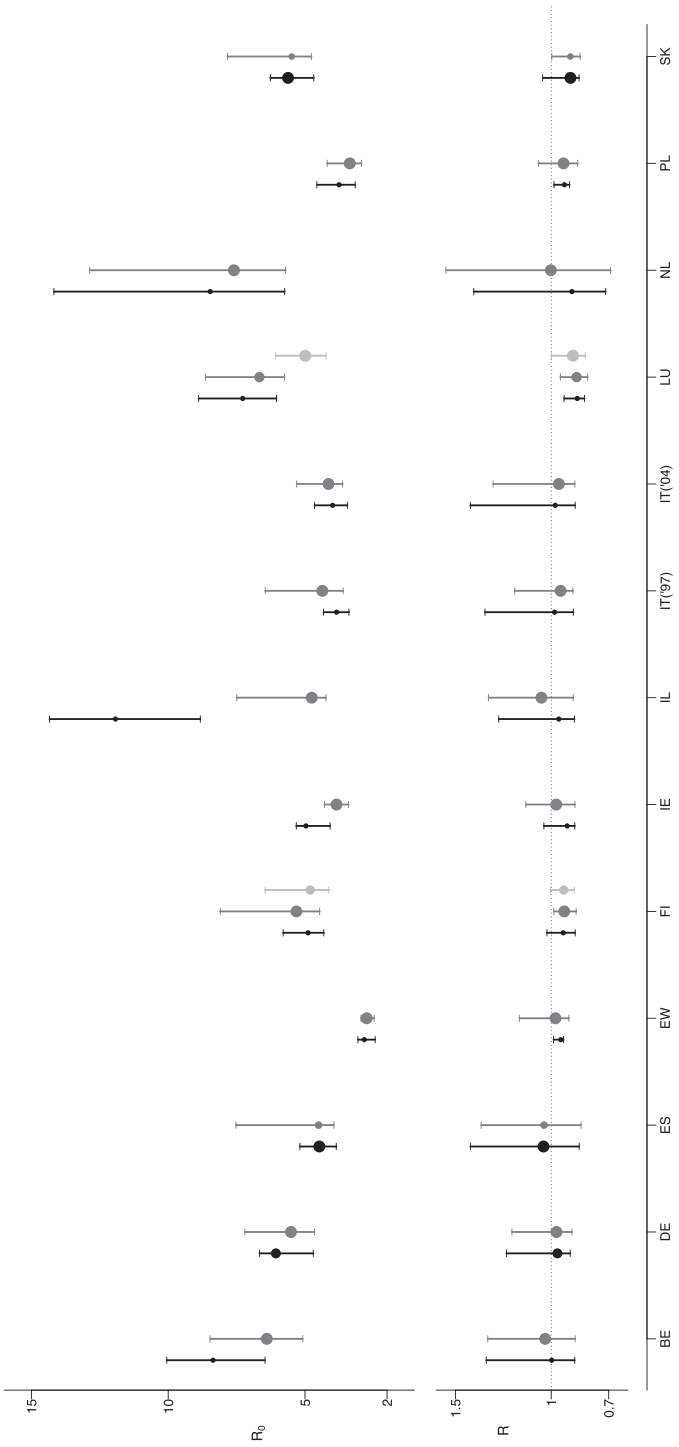


Fig. 1. Estimated basic and effective reproduction numbers with 95% bootstrap percentile confidence intervals for constant (black), log-linear (gray) and extended log-linear (light gray) proportionality factor. For each country, sizes of the dots are proportional to Akaike weights, hence larger dots correspond to smaller AIC values. The dotted horizontal line indicates the single eligible value for R under endemic equilibrium, which is one.

Elucidating potential risk factors

To address the differences in transmissibility between countries, a selection of 39 relevant country-specific variables was made, comprising data on demography, childcare, population density and weather (see Table 1 in Supplementary Material). To investigate associations between R_0 and these variables, two different non-parametric approaches are considered, which are briefly described below and more elaborately in the Supplementary Material.

Maximal information coefficient. The maximal information coefficient (MIC) (Reshef et al., 2011) is a measure of two-variable dependence, designed specifically for rapid exploration of high-dimensional data sets. The MIC is part of a larger family of maximal information-based non-parametric exploration statistics, which can be used not only to identify important relationships in data sets but also to characterize them.

Random forest approach. Secondly, a random forest approach for regression is used (Breiman, 2001), which is a class of ensemble

methods – methods that generate many classifiers and aggregate their results – specifically designed for classification and regression trees. Each tree is constructed using a different bootstrap sample of the data and each node is split using the best among a subset of predictors randomly chosen at each node. Compared to many other classifiers, this turns out to perform very well and is robust against overfitting (Breiman, 2001). In addition, it has only two parameters – the number of variables in the random subset at each node and the number of trees in the forest – and is usually not very sensitive to their values. We use the random forest algorithm from the randomForest package in R with the default number of trees (500). The number of split variables is selected such that the highest percentage explained variance is obtained. The package produces two measures of importance of the predictor variables: "mean decrease in accuracy" and "mean decrease in node purity".

Sensitivity analysis. To test the sensitivity of this risk factor analysis, we applied the MIC and random forest approach to estimates of R_0 when using the best fitting contact matrix. This sensitivity analysis is included in the Supplementary Material. We can conclude that the risk factor analysis is quite robust to changes in the contact matrix, as the most important influential factors do not change.

Results

Basic and effective reproduction number. We apply the social contact data approach with a constant and age-specific log-linear proportionality factor, as in (2), to the 13 serological data sets available for VZV. The estimated basic and effective reproduction numbers for both models are presented in Fig. 1 together with 95% bootstrap percentile confidence intervals (also in Table 3 in Supplementary Material). The size of the dots are proportional to the Akaike weights (see Supplementary Material), hence larger dots correspond to smaller AIC values. These estimates are supplemented with estimates of the mean age at infection in Table 3 in the Supplementary Material.

Models are classified as eligible based on the 95% confidence interval for the effective reproduction number, and eligible models are compared by means of AIC. When the model with lowest AIC value is eligible, this model is selected. This results in the age-specific log-linear proportionality factor being preferred for Belgium, Denmark, England and Wales, Ireland, Israel, Italy, The Netherlands and Poland. For Spain and Slovakia, the constant proportionality factor is sufficient to provide a good fit. For Finland, the log-linear model is preferred in terms of AIC, but this model is not eligible, whereas for Luxembourg, both models are not eligible. In both cases, the constant and basic log-linear model are not capable of providing a good fit to the data.

Therefore, we consider the extended log-linear model in (4) for Finland and Luxembourg. Fig. 2 presents the profile likelihood estimates of R_0 and R as a function of γ_2. We observe that by including an infectiousness component in the proportionality factor, the effective reproduction number R can be estimated closer to 1. Note that the estimate of R_0 decreases quite substantially with decreasing γ_2, in contrast to an increase in R. This reverse relation seems counter-intuitive, but is caused by an interplay between $q(a, a')$ and $s(a)$. Now, by performing a non-parametric bootstrap for every value of γ_2 on a specific grid, it is possible to determine the maximal value of γ_2 such that 1 is within the 95% bootstrap confidence interval of R. This is illustrated in Fig. 3.

The parameter estimates and confidence intervals for the extended log-linear model based on these maximal values of γ_2 are also displayed in Fig. 1. We observe the following: for Finland, the extended model has an improved fit compared to the constant model and is conform with the endemic equilibrium assumption.

Table 2
Ten factors with the largest MIC value of association with R_0, estimated from the final model selected for each country, and corresponding Spearman correlation coefficients ρ_S.

		MIC	ρ_S
1.	Inequality of income distribution	1.0	−0.64
2.	Poverty rate	1.0	−0.73
3.	% infants vaccinated against mumps	0.65	0.64
4.	average square meter living area pp	0.59	0.42
5.	% breast feeding at 3 months	0.47	−0.21
6.	% employed women 25–49 (min. 1 child 0–5)	0.46	0.38
7.	% infants vaccinated against pertussis	0.38	0.46
8.	% infants vaccinated against rubella	0.36	0.51
9.	% population aged 0–14	0.32	−0.22
10.	Total health expenditure	0.32	0.51

For Luxembourg, only the extended model is eligible, and in addition, it has the lowest AIC value. Note that the estimate of R_0 for Luxembourg decreases considerably.

The estimated seroprevalence curves based on the selected model for each country are presented in Figs. 4 and 5. The fitted seroprofiles show a similar pattern across countries, with most infections occurring during early childhood and the estimated prevalence approaching one as age increases. However, the prevalence does not reach one in all countries and, for example, Italy has a more particular profile. Looking at the FOI curves, the largest estimate is observed in the Netherlands (0.57 year^{-1}) at the age of 5, followed by Luxembourg (0.49 year^{-1}). The largest estimate of R_0 is obtained for The Netherlands (7.60) and the lowest for England and Wales (2.75). 11 out of 13 countries have R_0 estimated below 6.

Risk factors. There is considerable variation in estimated basic reproduction numbers, and hence in transmissibility, among the countries under consideration. Therefore, we aim to explain these differences by applying the MIC and random forest approach on a selected set of 39 relevant country-specific factors (Tables 1 and 2 in Supplementary Material). Table 2 in the Supplementary Material contains the pairs of potential risk factors with the strongest correlation given by the Spearman correlation coefficient. These correlations can be used to interpret the relation between R_0 and certain factors.

The ten factors with the largest MIC of association with R_0, are presented in Table 2 together with the corresponding Spearman correlation coefficients. This implies, for example, that the higher the inequality of income, the lower R_0. Results of the random forest analysis of R_0 are summarized in Table 3 where the ten highest scoring factors for both importance measures are given. Comparing the results of both analyses, we observe that factors related to the distribution of wealth (inequality of income and poverty rate), vaccination coverage in infants (e.g. mumps vaccination coverage) and child care attendance (e.g. the percentage of infants that receive no formal care) seem to be associated with the transmissibility of VZV.

Discussion

In this article, we investigated the transmissibility of VZV in 12 European countries using serological survey data and social contact data. We contrasted the social contact hypothesis, which is currently the most used approach in the literature, against an approach reflecting differences in characteristics related to susceptibility and infectivity. Furthermore, we introduced the effective reproduction number as a model eligibility criterion and we identified which country-specific socio-demographic factors are important in explaining differences in transmission potential between European countries using two non-parametric approaches: the maximal information coefficient and random forest.

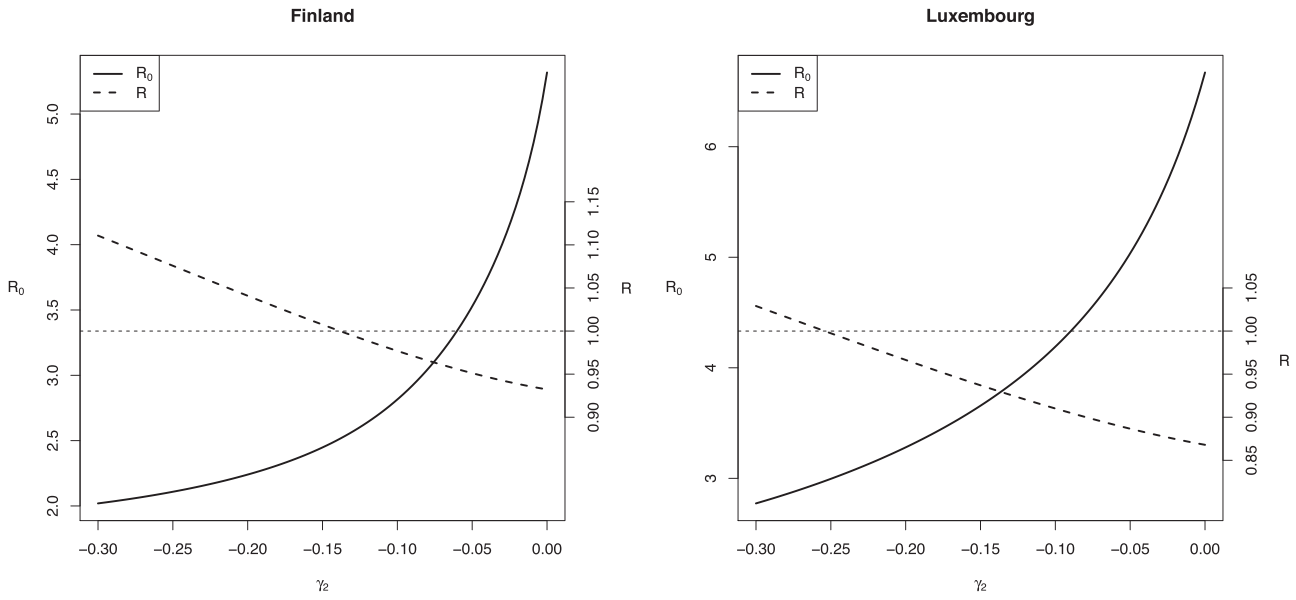


Fig. 2. Profile likelihood estimates of R_0 (left axis) and R (right axis) as a function of γ_2, the parameter related to infectiousness, for Finland and Luxembourg.

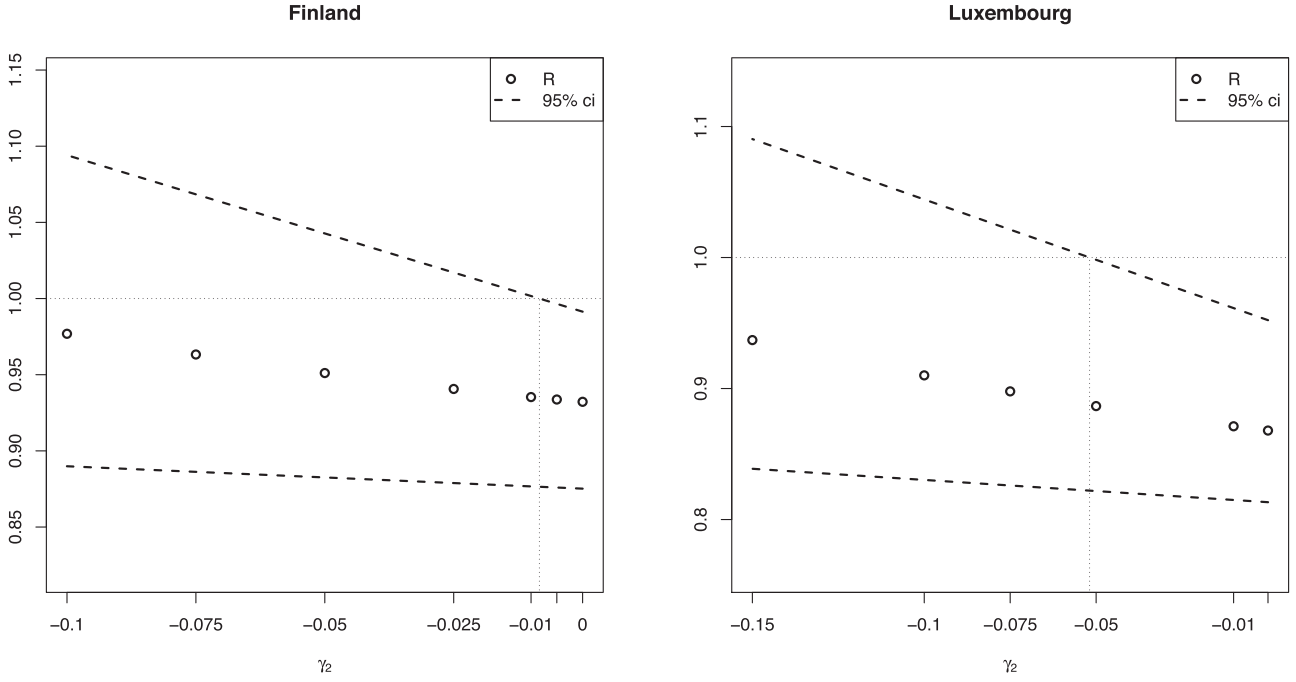
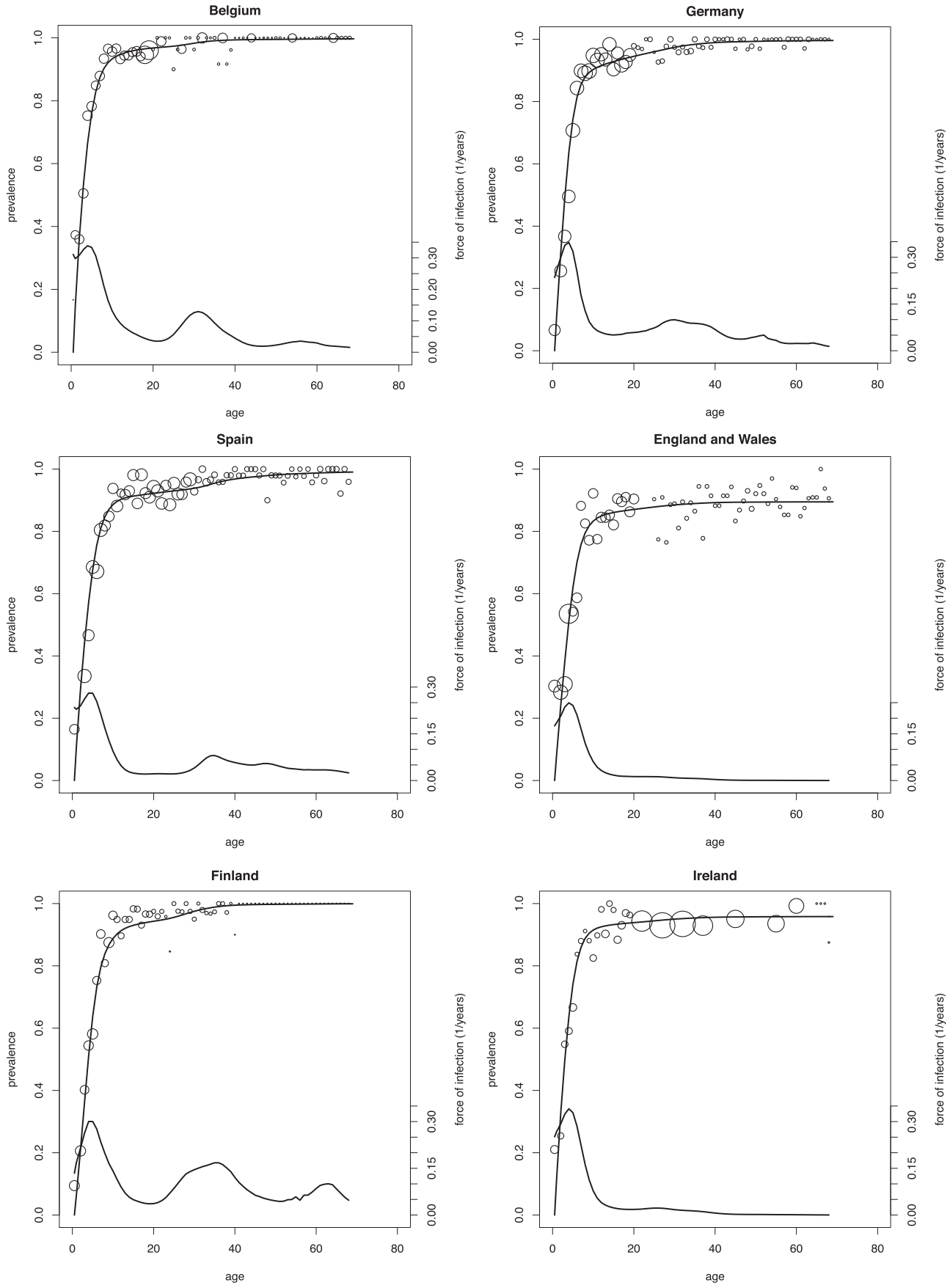


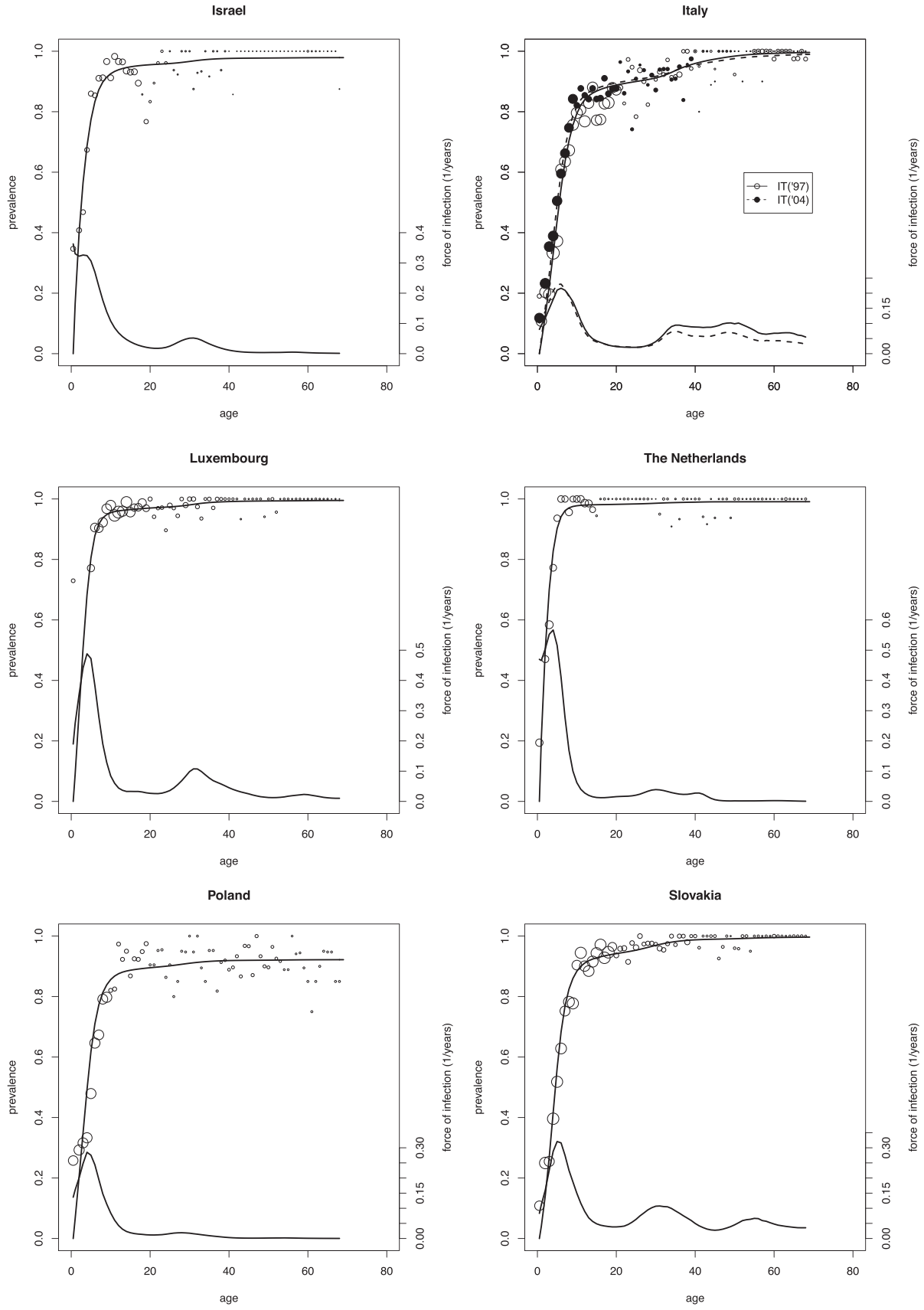
Fig. 3. Profile likelihood estimates of R (dots) with interpolated 95% bootstrap percentile confidence intervals (dashed lines) as a function of γ_2, the parameter related to infectiousness, for Finland and Luxembourg. The vertical dotted line indicates the value of γ_2 for which the upper confidence limit of R equals 1 (horizontal dotted line).

Table 3
Ten best scoring factors obtained by a random forest analysis of R_0, estimated from the final selected model for each country, and corresponding Spearman correlation coefficients ρ_S.

	% increase in MSE	ρ_S	Increase in node purity	ρ_S
1.	Inequality of income distribution	−0.64	Inequality of income distribution	−0.64
2.	Poverty rate	−0.73	Poverty rate	−0.73
3.	Total health expenditure	0.51	Average population density	0.33
4.	% 0–2 that receive no formal care	−0.29	% 0–2 that receive no formal care	−0.29
5.	% infants vaccinated against mumps	0.64	Unmet medical needs	−0.31
6.	% population aged 0–14	−0.22	Total health expenditure	0.51
7.	% employed women (min. 1 child 0–5)	0.38	Enrollment rates children 0–2	0.15
8.	Average square meter living area pp	0.42	Average square meter living area pp	0.42
9.	Average population density	0.33	% 65+ vaccinated against influenza	−0.19
10	Enrollment rates children 0–2	0.15	% infants vaccinated against mumps	0.64



Figs. 4 and 5. Observed age-specific VZV seroprevalence (dots) and the profile estimated from the final model selected for each country (solid line). The corresponding force of infection estimates are displayed by the lower solid line.



Figs. 4 and 5. (*Continued*).

The social contact hypothesis provided a good fit to the VZV seroprevalence for only 2 out of 12 countries. The other countries benefited from an extended approach by assuming an age-dependent proportionality factor, which supports and extends earlier findings of Goeyvaerts et al. (2010) for VZV in Belgium. This may reflect the additional importance of age-specific characteristics related to susceptibility and infectiousness, such as the mean infectious period. Furthermore, the social contact data are used as proxies for events by which an infection is transmitted. Hence, the proportionality factor can also be considered as an age-specific adjustment factor relating the true contact rates underlying infection to the social contact proxies. Alternatively, social data are difficult to collect from young children, with parents filling out the diary on their behalf. It may well be that they consistently underestimate the true number of contacts that young children make.

Our analysis directly improves upon the original analysis of the ESEN2 data on VZV by Nardone et al. (2007) who used the traditional Anderson and May approach by imposing a 3-parameter structure on the WAIFW matrix (Anderson and May, 1991). Our method of using R as a model eligibility criterion extends the approach of Goeyvaerts et al. (2010) by addressing the indeterminacy of the infectivity parameter. Our results complement those of Melegaro et al. (2011) who analyzed part of the VZV serology using the social contact hypothesis only. Comparing the estimated R_0 values, we notice that our results in general somewhat differ from the estimates obtained by Nardone et al. (2007) and Melegaro et al. (2011). This is not unexpected, since there are differences in methodology and it is known that transmission assumptions have a large impact on the estimation of R_0. See Table 4 in the Supplementary Material for a comparative overview of the results.

The results in Fig. 1 indicate that there are substantial epidemiological differences between European countries. This is important to consider when parametrizing mathematical models. Childhood vaccination coverage (for different vaccines), child care attendance, population density and average living area per person were positively associated with R_0, whereas income inequality, poverty, breast feeding, and the proportion of children under 14 years of age showed negative associations. While it seems intuitively logical that greater child care attendance and population density lead to more rapid spread of varicella, other associations are more difficult to interpret. Less poverty and income inequality, and higher vaccination coverages may be associated with more affluent societies in which women are more likely to be employed and children have more universal access to childcare and kindergarten from an early age on, facilitating the spread of VZV.

In our analyses, we relied on a few assumptions. First of all, we assumed that the serological status of an individual is a direct measure of his/her current immunity against VZV (Plotkin, 2010). Further, we considered physical contacts lasting longer than 15 min to be a good proxy for potential varicella transmission events as shown by Goeyvaerts et al. (2010) for Belgium. Finally, our use of R as a model eligibility criterion relied on the assumption of endemic equilibrium. This assumption is supported by the similarity in the results obtained for the two samples of Italy. In addition most surveys span two seasons, which partly captures any seasonal fluctuation. However, there are many factors that can cause changes in the age distribution of VZV cases over time, e.g. changes in demography, medical practice, socio-cultural factors etc. Looking at this more rigorously requires an additional in-depth analysis which is the topic of future research. However, to get a sense of the way $\hat{R}$ changes when demographic or endemic equilibrium are perturbed, we present a sensitivity analysis in the Supplementary Material. We observe that $\hat{R}$ increases when a percentage of the newborns would have been vaccinated and when the number of births would be increasing. It decreases when the annual number of births would decrease.

Since direct inference for the infectivity parameter is hindered by the lack of information regarding infectiousness in the serological data, we estimated this parameter via indirect inference using the effective reproduction number. This indeterminacy illustrates that the use of social contact data does not completely resolve the identifiability issues encountered when estimating mixing patterns from serological data. Hence, further research is necessary to obtain additional knowledge about the age-specific susceptibility and infectivity profiles in order to inform the proportionality factor in this social contact approach.

Acknowledgments

ES acknowledges support from a Methusalem research grant from the Flemish Government. NG is beneficiary of a postdoctoral grant from the AXA Research Fund. NH acknowledges support from the Antwerp University scientific chair in Evidence-Based Vaccinology, financed in 2009–2014 by an unrestricted gift from Pfizer. AM is currently receiving funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)/ERC Starting Grant [Agreement No. 283955]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. The computational resources and services used in this work were provided by the Hercules Foundation and the Flemish Government – department EWI.

This study was initiated as part of POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract number: SSP22-CT-2004-502084.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.epidem.2014.12.005.

References

Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J., 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, 365–382.

Vynnycky, E., White, R.G., 2010. An Introduction to Infectious Disease Modelling. Oxford University Press.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Whitaker, H.J., Farrington, C.P., 2004a. Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. Stat. Med. 23, 2429–2443.

Nardone, A., de Ory, F., Carton, M., Cohen, D., van Damme, P., Davidkin, I., Rota, M.C., et al., 2007. The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. Vaccine 25, 7866–7872.

Miller, E., Marshall, R., Vurdien, J., 1993. Epidemiology, outcome and control of varicella-zoster infection. Rev. Med. Microbiol. 4, 222–230.

Greenhalgh, D., Dietz, K., 1994. Some bounds on estimates for reproductive ratios derived from the age-specific force of infection. Math. Biosci. 124, 9–57.

Farrington, C.P., Kanaan, M.N., Gay, N.J., 2001. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. Appl. Stat. 50, 251–292.

Unkel, S., Farrington, C., Whitaker, H., Pebody, R., 2014. Time varying frailty models and the estimation of heterogeneities in transmission of infectious diseases. Appl. Stat. 63, 141–158.

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

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P., Beutels, P., 2009. Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. Math. Biosci. 218, 80–87.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Van Damme, P., Beutels, P., 2010. Estimating infectious disease parameters from data on social contacts and serological status. Appl. Stat. 59, 255–277.

Wallinga, J., Lévy-Bruhl, D., Gay, N., Wachmann, C., 2001. Estimation of measles reproduction ration and prospects for elimination of measles by vaccination in some Western European countries. Epidemiol. Infect. 127, 281–295.

Read, J., Edmunds, W., Riley, S., Essler, J., Cumming, D., 2012. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol. Infect. 140, 2117–2130.
Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., et al., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5 (3), 381–391, doi:10.1371/journal.pmed.0050074.
Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P., Beutels, P., 2010. Seventy-five years of estimating the force of infection from current status data. Epidemiol. Infect. 138, 802–812.
Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P., Beutels, P., 2009. Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. Epidemiol. Infect. 137, 48–57.
Melegaro, A., Jit, M., Gay, N., Zagheni, E., Edmunds, W.J., 2011. What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. Epidemics 3, 143–151.
Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., Beutels, P., 2012. Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective. Springer-Verlag New York Inc.
Farrington, C.P., 2003. Modelling Epidemics. The Open University.
Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C., 2011. Detecting novel associations in large data sets. Science 334, 1518–1524.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Plotkin, S., 2010. Complex correlates of protection after vaccination. Clin. Infect. Dis. 56, 1458–1465.