Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases

C. Paddy Farrington, Heather J. Whitaker, Steffen Unkel, Richard Pebody

ABSTRACT

New methods are proposed for investigating and quantifying the extent of heterogeneity in effective contact rates relevant to the transmission of infections, based on the correlations within individuals between times to infection for different infections. The methods are developed for serological surveys, which provide readily accessible sources of individual data on several infections, and are applied to a wide range of different infections. It is found that childhood infections are often highly correlated within individuals in early childhood, the correlations persisting into adulthood only for infections sharing a transmission route. The authors speculate that childhood correlation stems from confounding of different transmission routes, and represents heterogeneity in childhood circumstances, notably nursery attendance. In contrast, it is suggested that correlations in adulthood are route-specific. Two applications of the methods are discussed, to making inferences about routes of transmission when these are unknown or uncertain, and to the estimation of epidemiological parameters such as the basic reproduction number and critical immunization threshold. Two examples of such applications are presented, to elucidating the transmission route of polyomaviruses BKV and JCV, and to the estimation of the basic reproduction number and critical immunisation coverage of varicella zoster infection in four European countries.

KEYWORDS

Basic Reproduction Number; Communicable Diseases; Correlations; Disease Transmission, Infectious; Frailty; Heterogeneity; Mass Vaccination; Serologic Tests;

ABBREVIATIONS

B19: parvovirus B19; BKV: polyomavirus BKV; CI: confidence interval; CMV: cytomegalovirus; EBV: Epstein-Barr virus; HAV: hepatitis A virus; HPY: *Helicobacyer pylori*; HSV1: herpes simplex virus type 1; JCV: polyomavirus JCV; MMR: measles, mumps and rubella; MUM: mumps virus; RUB: rubella virus; TOX: toxoplasma infection; UK: United Kingdom.

INTRODUCTION

It has long been understood that the heterogeneity of a population with respect to factors that may enhance or inhibit the transmission of infections may affect the effectiveness of strategies to control them (1). As a general rule, the greater the heterogeneity, the greater the epidemic potential of the infection and the more difficult it is to control. Thus, estimates of the basic reproduction number and the critical immunization threshold, derived without accounting appropriately for heterogeneity, are likely to be biased.

Therefore, allowing for individual heterogeneity in statistical and mathematical models of infectious diseases is important. Such models often involve specifying contact rates between individuals. However, it is often difficult to decide what constitutes a contact between two individuals, and hence to specify what the relevant heterogeneities are, let alone measure them.

When what constitutes a contact is clear, relevant heterogeneities are relatively straightforward to measure, at least in principle. This is the case for sexually transmitted infections, where 'contact' means 'sexual contact' and relevant heterogeneities include frequency of sexual contacts and rate of partner change. Heterogeneity can then be quantified explicitly through surveys of sexual behaviour (2). However, for indirectly transmitted infections – for example infections transmitted by aerosol, or by the fecal-oral route via contaminated food or water, or by fomites – there are no contacts in any but a metaphorical sense. Additionally, while it may be known in broad terms which routes of transmission are involved, there is much less clarity about their relative importance. While some detailed studies of the relative importance of different transmission routes exist (3), these are uncommon. Often, the best that can be done is to seek information on a proxy variable which might be expected to be correlated with relevant behaviour. This is the rationale behind contact surveys, using proxies such as 'having a two-way conversation' (4-6). Inevitably, such an approach is approximate and may require *post-hoc* adjustments (7). However, it can also give important insights into transmission routes (8).

Here we use a different approach to quantifying relevant heterogeneities, using correlations between infections in individuals. The rationale of the approach is as follows. If two infections are transmitted by a similar route, one might expect that the extent of heterogeneity in behaviour relevant to the transmission of infection will be reflected by the strength of the correlation between the two infections. Thus, for example, an individual with a high 'activity

3

level' relevant to transmission of two infections (which might be low personal hygiene, in the case of fecal-oral infections, or a high rate of social interaction, for infections transmitted by droplets) will be more likely to acquire both infections by a given age than an individual with low activity levels, all other things being equal. One benefit of the approach is that it does not require an explicit definition of what exactly such 'activity levels' represent.

This idea was first explored in (9), and has been applied to study the transmission of hepatitides B and C viruses (10), and to Epstein-Barr and herpes simplex type I viruses (11). Typically, blood samples collected from a defined population are tested for antibodies to several antigens. This gives rise to multivariate current status data on several infections, on the same individuals, with additional information on age, gender etc..

In this paper we develop these ideas further. We show that the presence of correlations between different infections can be exploited to gain better understanding of how infections are transmitted, notably for infections with several possible transmission routes. We suggest that such correlations can be used to elucidate likely transmission routes when these are not known. Further, the degree of heterogeneity inducing the correlation can be modelled, and this information can then be used to improve the estimates of epidemiological parameters such as reproduction numbers, the estimation of which typically only accommodates the effect of directly measured heterogeneities. In brief, correlations between infections open a window on individual behaviours which are difficult to measure, regarding heterogeneities of contacts which are difficult to define.

We present new methodology to take into account the fact that activity levels – and hence the correlations they induce – vary with age. This enables us to describe and quantify heterogeneities and how they evolve over time. We apply the methods to a wide variety of different data sets, obtained in different serological surveys. We also explore two contrasting applications of this methodology, to the identification of routes of transmission of polyomaviruses BKV and JCV, and to estimation of the basic reproduction number and critical immunization level for varicella zoster virus infection.

STATISTICAL METHODS

The statistical framework is described in four subsections. The details are kept to a minimum; further details are available in the Web Material (available here at the end of the file).

Incorporating age-dependent heterogeneity via frailty models

Throughout, let x, y denote age. To begin with, consider a single infection, and suppose that age is the only measured attribute of that individual (the methods can readily be extended to include others). Suppose that all that individual's unmeasured attributes or behaviours which are relevant to the transmission of this infection at age x may be described by an *activity level*, which is a positive random variable u_x with density $f_x(u_x)$ and mean 1; for simplicity we shall assume that u_x is a deterministic function of x and a finite set of age-independent random variables. The variance of u_x,

$$\gamma(x) \quad = \quad \text{var}\{u_x\},$$

thus represents the degree of unmeasured individual heterogeneity in the population at age x. Our aim is to estimate $\gamma(x)$ for a range of infections and use these estimates to make inferences about the epidemiology of the infections.

Let $\beta_0(x;y)$ denote the *average effective contact rate* between an individual of age x and an individual of age y in this population. This is the contribution of a typical infectious individual of age y to the instantaneous rate of infection of a typical susceptible individual of age x; here 'typical' means 'average with respect to unmeasured heterogeneities'. We extend this notion to encompass activity levels by denoting $\beta(x, u_x; y, v_y)$ the contribution of an infectious individual of age x and activity level u_x to the instantaneous rate of infection of a susceptible individual of age y with activity level v_y. To make further progress, we assume that

$$\beta(x, u_x ; y, v_y) \quad = \quad u_x \, \beta_0(x ; y) \, v_y \qquad . \tag{1}$$

This model is an elaboration of one first proposed by (12) and implemented in (9). The assumption that individual activity levels combine multiplicatively as in equation 1 is a form of proportional mixing (13).

Now let $\lambda(x, u_x)$ denote the force of infection exerted on an individual of age x and activity level u_x. It follows from equation 1 that

$$\lambda(x, u_x) = u_x \, \lambda_0(x), \tag{2}$$

where $\lambda_0(x)$ is the baseline force of infection. This defines a frailty model for the force of infection, with age-varying multiplicative frailty u_x (9, 14).

5

Paired serological survey data

Consider two infections, labelled 1 and 2, conferring lasting immunity and for which long-lived serological markers are known. Serological tests on a blood sample collected at age x will determine whether the individual from which the sample was collected is seropositive or seronegative to each infection.

Suppose further that individual activity levels u_x are relevant to transmission of both infections 1 and 2. This will occur, in particular, if the two infections are transmitted by the same route. The forces of infection on an individual of age x with shared activity level u_x are then

$$\lambda_1(x, u_x) = u_x\, \lambda_{01}(x) \quad \text{and} \quad \lambda_2(x, u_x) = u_x\, \lambda_{02}(x), \tag{3}$$

the subscripts 1 and 2 referring to infections 1 and 2. Since the same frailty term u_x is shared by the two infections, equation 3 defines a shared frailty model (14). The test results obtained at age x can be as follows: seronegative for both infections, which occurs with probability denoted $S_{00}(x)$; seronegative for 1 and seropositive for 2, which occurs with probability $S_{01}(x)$; seronegative for 2 and seropositive for 1, with probability $S_{10}(x)$; and seropositive for both, with probability $S_{11}(x)$.

We have so far ignored variation with calendar time. This may be important, in particular, for infections transmitted via the fecal-oral route, owing to improvements in hygiene and sanitation over time. However, valid inferences about the shape of $\gamma(x)$ may be obtained from a single survey even when the baseline forces of infection decline with calendar time (see Web Material).

Displaying correlations between infections using bivariate serological survey data

The extent of heterogeneity in the population at age x of relevance to the transmission of both infections of interest can be estimated from the strength of association in the 2 x 2 tables of counts $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$, using the same notations as for the cell probabilities. Measures of association such as the odds ratio can be misleading and lack direct interpretation in the present context where the focus is specifically on age-specific heterogeneity. We use another measure, denoted $\varphi(x)$, whose properties approximate those of the cross-ratio function

(15,16). The value $\varphi(x) = 0$ corresponds to independence; $\varphi(x) > 0$ corresponds to positive association, notably that resulting from heterogeneity, and $\varphi(x) < 0$ to negative association, as may arise owing to cross-immunity.

We also use the following summary measure of association across age groups:

$$\overline{\varphi} \;=\; \frac{\sum\limits_{x=1}^{M} p_x \hat{\varphi}(x)}{\sum\limits_{x=1}^{M} p_x}, \qquad \mathrm{var}(\overline{\varphi}) \;=\; \frac{1}{\sum\limits_{x=1}^{M} p_x}, \tag{4}$$

where the hat denotes the estimated value of $\varphi(x)$ and p_x is its (estimated) precision, that is, the reciprocal of its variance. Zeroes in the 2 x 2 tables of counts at each age x were handled recursively as follows. When one of the four margins of the table was zero, we combined it with the data for age x-1, and allocated the average of the ages for the combined table. For tables with zero counts but four non-zero margins, we added 0.5 to all four cells.

Models for age-dependent heterogeneity

The baseline forces of infection $\lambda_{01}(x)$ and $\lambda_{02}(x)$ are estimated using piecewise constant functions. Our interest centres on the frailty term and its variance $\gamma(x)$. Our basic model for the frailty is of the form

$$u_x \;=\; \{1+(w_1\text{-}1)h(x)\}w_2, \tag{5}$$

where w_1 and w_2 are independent gamma random variables with mean 1 and variances γ_1 and γ_2, respectively, and $h(x)$ is a deterministic function, typically of the form

$$h(x) \;=\; \exp(-(\rho x)^2). \tag{6}$$

Note that $E(u_x) = 1$. These models were introduced in (17). Their rationale will be motivated further below; briefly, w_1 represents heterogeneity in childhood, which evolves according to $h(x)$, and w_2 represents heterogeneity in adulthood. For this model, the age-specific heterogeneity has variance

$$\gamma(x) \;=\; h(x)^2 \gamma_1(1+\gamma_2) + \gamma_2. \tag{7}$$

7

Suppose that paired serological data (n_{00x}, n_{01x}, n_{10x}, n_{11x}) are available at ages $x = 1,2,…,M$. A Dirichlet-multinomial model was used, to allow for overdispersion due to assay variability (11). The model parameters and hence the function $\gamma(x)$ and the baseline forces of infection may be estimated by maximising the log likelihood.

Impact on reproduction number and critical vaccination threshold

The methods of (9) may readily be extended to cover the present more general setting. Suppose for simplicity that an infection confers long-lasting immunity, has a short infectious period D, and is in endemic equilibrium in a population of size N with rectangular age structure on $[0, L]$. If $u(x)$ represents individual heterogeneity at age x, with variance $\gamma(x)$, then the basic reproduction number of the infection is the dominant eigenvalue of the operator $(ND / L)[1 + \gamma(x)]\beta_0(x, y)$. Increasing heterogeneity has the effect of increasing R_0 and the critical immunisation threshold for vaccination close to birth, $\pi_c = 1 - R_0^{-1}$.

DATA SOURCES

We used data from seven serological surveys, collected in England and Wales as part of a long-standing programme of serological surveillance in the United Kingdom (UK) (18), two European sero-epidemiology networks ESEN (19) and ESEN2 (20), and from the Europe-wide project POLYMOD (21). Details of the tests used may be found in the references listed below.

Survey 1, undertaken in 1986 in the UK (22), provides paired seroprevalence data on mumps (MUM) and rubella (RUB), the latter restricted to males owing to the selective rubella vaccination programme in adolescent girls.

Survey 2, undertaken in the UK in 1991, provides information on seroprevalence of parvovirus B19 (23), cytomegalovirus (CMV) (24), and rubella (18). Since universal rubella vaccination at 15 months was introduced in the UK in 1987-1988, we restricted the analysis of the rubella data to males aged 11+ years. The samples were also tested for antibodies to polyomavirus types BKV and JCV (25); these data will be discussed later in the paper.

Survey 3, undertaken in the UK in 1994 and 1995, provides paired data on seroprevalence to Epstein-Barr virus (EBV) and herpes simplex virus type I (HSV1) (26).

Survey 4, undertaken in the UK in 1996, provides information on seroprevalence of varicella zoster virus (VZV)(27), parvovirus B19 (28), *Helicobactyer pylori* (HPY) (29), and hepatitis A virus (HAV) (30). The samples were also tested for toxoplasma (TOX) using *Toxoplasma gondii* specific IgG by ELISA (CAPTIA Select Toxo-G ELISA, Trinity Biotech) (unpublished data).

Surveys 5, 6 and 7 were undertaken in Belgium, Italy and Poland, respectively, and provide paired seroprevalence data on parvovirus B19 and VZV as part of the POLYMOD project (8, 21, 31).

Details of all seven surveys are shown in Table 1. Excluding the polyomaviruses for the time being, these surveys provide 18 sets of paired data for analysis. Table 2 shows the major route of transmission for the 10 infections considered, based primarily on (32). Rubella is generally regarded as transmitted by droplets, though aerosol transmission is also mentioned in the literature (33). Close contact may be involved in much transmission of parvovirus B19 and VZV (8).

RESULTS

Descriptive analysis

The summary values of the association parameter described in equation 7, along with 95% confidence intervals, are presented in Table 3, stratified according to whether the main route of transmission is likely to be shared or not. This categorization was decided a priori based on Table 2.

Three main features emerge. First, associations are generally higher and significantly positive for pairs of infections sharing a major transmission route than infections not sharing such a route. Second, the associations between infections not sharing a major route of transmission are nevertheless often positive, though seldom significantly so. Third, the associations between infections transmitted by the respiratory route tend to be lower than between those transmitted by other routes.

These patterns in overall associations confirm that correlations between infections contain information on transmission routes, though the case is perhaps less compelling for respiratory infections. However, overall measures are crude, so we plotted the values of the association parameter $\varphi(x)$ at each age x (Figures 1-3). The areas of the points within each graph are proportional to the precisions p_x (see equation 4); the smooth lines are non-parametric precision-weighted estimates of trend.

These plots show that, in childhood, there is a strong correlation between infections irrespective of route of transmission, declining with age. For infections transmitted by different routes, the association declines to zero in adulthood, whereas for infections transmitted by the same route, the association generally declines to some positive, constant value – with some exceptions, particularly among respiratory infection pairs. This is shown in Table 4, which presents the association measure for ages 21+ years (when data are available beyond age 20) or ages 11+ years (when data are not available beyond age 20). The lower associations in adulthood between respiratory infections could be due to the lesser relevance of variations in individual behaviours to transmission of such infections, for example if aerosol spread is a major factor in transmission.

The patterns of association can be interpreted in terms of the changes in heterogeneity in the population that induce correlations via shared frailty terms: the stronger the association, the greater the heterogeneity. Note also that the strength of the associations should not be interpreted in terms of the magnitude of the forces of infection, but in terms of heterogeneities.

The data suggest the following broad interpretation (alternative interpretations are considered in the Discussion). In childhood, the routes of transmission considered here are confounded, owing to the nature of contacts at young ages, which include close mixing involving much direct contact. At young ages, the population is very highly heterogeneous with respect to such factors, reflecting variation in family circumstances and nursery attendance, as well as individual behaviour. At older ages, behaviour and circumstances change so that transmission routes gradually become differentiated, and common social factors intervene (such as school attendance), so that the heterogeneity drops (though the force of infection typically increases). For infections transmitted by the same route, the association eventually reflects the heterogeneity in behaviours associated with transmission solely via that specific route.

For infections transmitted by different routes, there is no common factor and the heterogeneity drops to zero.

Two further applications

Inference about routes of transmission

Our methods can be used to make inferences about routes of transmission, when these are uncertain or unknown. The idea is to test a panel of sera for antibodies to the infection of interest, and to several infections of know route of transmission. A shared route of transmission is revealed by a positive correlation in adulthood (positive correlations in childhood may reflect confounding of transmission routes).

We illustrate this idea with data on polyomaviruses BKV and JCV (26). This study found a strong negative correlation between BKV and JCV at younger ages, suggestive of cross-protection; this is reflected in the association plot between the two infections (Figure 4). The lack of association after age 50 may reflect reduced sensitivity of the test, resulting in misclassification of sera and thus bias towards the null (26).

The association plots of polyomaviruses BKV and JCV with cytomegalovirus, parvovirus B19 and rubella (the latter in males aged 11 – 69 years, to ensure they are unvaccinated) are shown in Figure 5. The odd behaviour of the plots at low ages may be attributable to the negative correlation between BKV and JCV. A further problem is the decline in test sensitivity at older ages. To mitigate these effects while retaining sufficient data, we calculated summary values of the association parameter in adulthood for the age range 21 – 55 years, using expression 7. These are shown in Table 5.

These estimates suggest a possible positive association in this age range between CMV and JCV, but not between CMV and BKV, though the association data are sparse. The data also suggest a possible positive association between parvovirus B19 and BKV in adulthood, but none with JCV (associations in childhood are clearly affected by the strong negative association between BKV and JCV). There is no compelling evidence of any associations with rubella in males.

These considerations suggest that the route of transmission of polyomavirus BKV may be shared with parvovirus B19, while that of polyomavirus JCV might be shared with cytomegalovirus. These inferences, though tentative, are in line with those of (26).

For completeness, Figure 6 shows the seroprevalence of the four infections, plotted to age 44 owing to sparsity of data at older ages. The force of infection of BKV is greater than that of parvovirus B19, and is more akin to that of varicella zoster virus (transmitted by a similar route as parvovirus B19). The force of infection of JCV is similar to that of cytomegalovirus. Note also the decline in measured antibodies to BKV at older ages.

Estimating the heterogeneity in transmission of varicella zoster virus and its impact on key epidemiological parameters

We illustrate the impact of individual age-dependent heterogeneity on estimates of the basic reproduction number R_0 and the critical immunisation threshold π_c for varicella zoster infection in four European countries: Belgium, Italy, Poland and the UK. For this purpose, we explicitly model the heterogeneity using the model described in equations 5 and 6.

This model presumes two distinct sources of heterogeneity, represented by the frailty terms w_1 and w_2. As discussed above, the frailty w_2, of variance γ_2, represents heterogeneity of behaviour and individual circumstances related to transmission by the route specific to VZV, namely exchange of respiratory secretions. This route-related heterogeneity is presumed to remain present throughout life, though it only becomes apparent in adulthood. The frailty w_1, on the other hand, represents heterogeneities in childhood behaviour and circumstances (such a nursery attendance) which are related to transmission of virtually any childhood infection. This childhood-related heterogeneity is presumed to decline with increasing age, as childhood behaviour and circumstances evolve, for example through the learning of personal hygiene.

To estimate the heterogeneity we used paired data on parvovirus B19 and VZV from the four countries. The empirical and fitted associations are shown in Figure 7. The association is particularly strong in Poland, less so in the three other countries. In Poland and the UK, there is more evidence of association at older ages – as represented by the positive asymptote – than in Belgium or Italy. These associations reflect the changing heterogeneity with age, shown in Figure 8, obtained from the estimated values of the parameters γ_1, γ_2 and ρ using equation 7.

To estimate reproduction numbers and critical immunization thresholds, we used the social contact matrices described in (34). The estimated values of R_0 and π_c of VZV are shown in Table 6. The R_0 estimates without allowing for additional heterogeneity lie in the range $3 - 7$, similar to those obtained in other studies using a variety of different methodologies (8, 31, 35). Our interest is primarily in the impact of heterogeneity, shown in the final column of Table 6. The impact is small for Belgium and Italy, but appreciable for the UK and substantial for Poland. Similar results were obtained assuming homogeneous mixing.

DISCUSSION

We have developed new methods for exploring unmeasured individual factors relevant to the transmission of infectious diseases. Our methods are based on interpreting correlations between infections in terms of heterogeneities. They are readily applied to paired data on presence of antibodies undertaken on the same sera, for different infections. We find support for such an approach from the fact that, as expected, infections sharing a common mode of transmission tend to display positive correlations between times to infection.

Our approach substantially extends previously published methodology (9) on analysing bivariate serological survey data, through a more appropriate representation of the association (16), better understanding and explicit modelling of age effects (17), and novel applications.

As illustrated with polyomaviruses, one practical application of these ideas is to the reverse inference, namely to shed light on routes of transmission of infections for which these are uncertain, by examining correlations with infections of known transmission route.

A second application is to improve estimation of epidemiological parameters, such as R_0 and π_c. This requires explicit modelling of the heterogeneity, a natural framework for which is provided by age-dependent frailty modelling. We find that ignoring such individual heterogeneity risks underestimating the immunization level required for effective control.

For most of the infection pairs we studied, we found a strong association in childhood, typically declining with age either to zero (typically for infections without a shared route of transmission, or for respiratory infections) or to a positive constant (typically for infections with a shared route). We interpreted this as a 'childhood heterogeneity' effect like to be greatly enhanced by age-dependent factors such as nursery attendance. However, it is

13

possible to conceive of alternative explanations. One is individual variation in the development of children's immune systems, resulting in individual variation in systemic susceptibility rather than in effective contact rates. Another is a selection effect induced by certain types of frailty distributions, though exhaustive study has so far not yielded any support for this (17). Finally, cross-reactions in the antibody assays used to test the samples could generate spurious dependences, though such a consistent pattern as that seen in these data would perhaps be difficult to explain.

Traditionally, serological survey data for different infections have been studied in isolation. We propose that they should be studied together, to exploit correlations between different infections. Further methodological work is indicated, to improve survey design to optimise estimation of the parameters describing the heterogeneity and to handle multivariate data of higher dimension.

REFERENCES

1. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.

2. Johnson AM, Mercer CH, Erens B et al. Sexual behaviours in Britain: Partnerships, practices, and HIV risk behaviours. *Lancet* 2001; 358: 1835-1842.

3. Teunis PFM, Brienen N, Kretzschmar MEE. High infectivity and pathogenicity of influenza A virus via aerosol and droplet transmission. *Epidemics* 2010; 2: 215-222.

4. Mossong J, Hens N, Jit M et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* 2008; 5: E74.

5. Del Valle SY, Hyman JM, Hethcote HW, Eubank SG. Mixing patterns between age groups in social networks. *Social Networks* 2007; 29: 539-554.

6. Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology* 2006; 164: 936-944.

7. Goeyvaerts N, Hens N, Ogunjimi B et al. Estimating infectious disease parameters from data on social contacts and serological markers. Journal of the Royal Statistical Society Series C 2010; 59: 255-277.

8. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. What contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Epidemics* 2011; 3: 143-151.

9. Farrington CP, Kanaan MN, Gay NJ. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with Discussion). *Journal of the Royal Statistical Society Series C* 2001; 50: 251-292.

10. Sutton AJ, Gay NJ, Edmunds WJ, Hope VD, Gill ON, Hickman M. Modelling the force of infection for hepatitis B and hepatitis C in injecting drug users in England and Wales. *BMC Infectious Diseases* 2006; 6: E93.

11. Farrington CP, Whitaker HJ. Contact surface models for infectious diseases: Estimation from serologic survey. *Journal of the American Statistical Association* 2005; 100: 370-379.

12. Coutinho FAB, Massad E, Lopez LR, Burattini MN, Struchiner CJ, Azevedo-Neto RS. Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling* 1999; 30: 97-115

13. Hethcote HW, Van Ark JW. Epidemiological models for heterogeneous populations: Proportionate mixing, parameter estimation, and immunization programs. *Mathematical Biosciences* 1987; 84: 85-118.

14. Aalen OO, Borgan Ø, Gjessing H. *Survival and Event History Analysis: A Process Point of View*. New York: Springer, 2008.

15. Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 1989; 84: 487-493.

16. Unkel S, Farrington CP. A new measure of time-varying association for shared frailty models with bivariate current status data. Open University Statistics Group Technical Report 10/14; available from http://statistics.open.ac.uk/technical_reports.

17. Farrington CP, Unkel S, Anaya-Izquierdo K. The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society Series B* 2012; 74: 1-24 (In Press).

18. Osborne K, Gay N, Hesketh L, Morgan-Capner P, Miller E. Ten years of serological surveillance in England and Wales: methods, results, implications and action. *International Journal of Epidemiology* 2000; 29: 362-368.

19. Osborne K, Weinberg J, Miller E. The European sero-epidemiology network. *Eurosurveillance* 1997; 2: 29-31.

20. Nardone A, Miller E. Serological surveillance of rubella in Europe: European sero-epidemiology network (ESEN2). *Eurosurveillance* 2004; 9: 5-7.

21. Mossong J, Hens N, Friedrichs V et al. Parvovirus B19 infection in five European countries: Seroepidemiology, force of infection and maternal risk of infection. *Epidemiology and Infection* 2008; 136: 1059-1068.

22. Morgan-Capner P, Wright J, Miller CL, Miller E. Surveillance of antibody to measles, mumps and rubella by age. *British Medical Journal* 1988; 297: 770-772.

23. Gay N, hesketh LM, Cohen BJ et al. Age specific antibody prevalence to parvovirus B19: How many women are infected during pregnancy? *Communicable Disease Report* 1994; 4: R104-R107.

24. Vyse AJ, Hesketh LM, Pebody RG. The burden of infection with cytomegalovirus in England and wales: How many women are infected in pregnancy? *Epidemiology and Infection* 2009; 137: 526-533.

25. Knowles WA, Pipkin P, Andrews N et al. Population-based study of antibody to the human polyomaviruses BKV and JCV and the simian polyomavirus SV40. *Journal of Medical Virology* 2003; 71: 115-123.

26. Morris MC, Edmunds WJ, Hesketh LM et al. Sero-epidemiological patterns of Epstein-Barr and Herpes Simplex (HSV-1 and HSV-2) viruses in England and Wales. *Journal of Medical Virology* 2002; 67: 522-527.

27. Vyse AJ, Gay NJ, Hesketh LM, Morgan-Capner P, Miller E. Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. *Epidemiology and Infection* 2004; 132: 1129-1134.

28. Vyse AJ, Andrews NJ, Hesketh LM, Pebody R. The burden of parvovirus B19 infection in women of childbearing age in England and Wales. *Epidemiology and Infection* 2007; 135: 1354-1362.

29. Vyse AJ, Gay NJ, Hesketh LM et al. The burden of *Helicobacter pylori* infection in England and Wales. *Epidemiology and Infection* 2002; 128: 411-417.

30. Morris MC, Gay NJ, Hesketh LM, Morgan-Capner P, Miller E. The changing epidemiological pattern of hepatitis A in England and Wales. *Epidemiology and Infection* 2002; 128: 457-463.

31. Nardone A, de Ory F, Carton M et al. The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine* 2007; 25: 7866-7872.

32. Heymann DL (Editor). *Control of Communicable Diseases Manual*. 19th Edition. Washington: American Public Health Association 2008.

33. Banatvala JE, Brown DWG. Rubella. *Lancet* 2004; 363: 1127-1137.

34. Farrington CP, Whitaker HJ, Wallinga J, Manfredi P. Measures of disassortativeness and their application to directly transmitted infections. *Biometrical Journal* 2009; 51: 387-407.

35. Van Effelterre T, Shkedy Z, Aerts M, Molenberghs G, Van Damme P, Beutels P. Contact patterns and their implied basic reproduction numbers: An illustration for varicella-zoster virus. *Epidemiology and Infection* 2009; 137: 48-57.

TABLES AND FIGURES

Table 1. Details of Surveys: Infection Pairs, Age Ranges, and Numbers of Paired Samples

Infection pair	Age range (y)	*n*	Infection pair	Age range (y)	*n*
	Survey 1			Survey 2	
MUM & RUB[1]	1 – 99	4585	B19 & CMV	1 – 96	1331
			B19 & RUB[1]	11 – 100	1700
	Survey 3		CMV & RUB[1]	11 – 100	726
EBV & HSV1	0 – 30	2893	B19 & BKV	1 – 69	1340
			B19 & JCV	1 – 69	1340
	Survey 4		BKV & CMV	1 – 69	665
B19 & VZV	1 – 20	1774	CMV & JCV	1 – 69	665
B19 & HAV	1 – 75	1128	BKV & RUB[1]	11 – 69	953
B19 & TOX	1 – 79	1516	JCV & RUB[1]	11 – 69	953
B19 & HPY	1 – 79	1829	BKV & JCV	1 – 69	2435
HAV & VZV	1 – 20	796			
TOX & VZV	1 – 20	1226		Surveys 5 – 7	
HPY & VZV	1 – 20	1262	Belgium	0 – 39	2380
HAV & TOX	1 – 81	2162	Italy	1 – 79	2434
HAV & HPY	1 – 84	3263	Poland	1 – 19	1150
HPY & TOX	1 – 84	3632			

[1] Males only

Table 2. Main Routes of Transmission for Infections, with References

Infectious agent	Acronym	Main route(s) of transmission (references)
Cytomegalovirus	CMV	Intimate mucosal contact (32)
Epstein-Barr virus	EBV	Oropharyngeal route via saliva (32)
Helicobacter pylori	HPY	Fecal-oral and/or oral-oral (32)
Hepatitis A virus	HAV	Fecal-oral (32)
Herpes simplex virus type 1	HSV1	Oropharyngeal route via saliva (32)
Mumps virus	MUM	Airborne, droplets or direct contact (32)
Parvovirus B19	B19	Close contact with respiratory secretions (8,32)
Toxoplasma	TOX	Oral ingestion of feline feces (32)
Rubella virus	RUB	Droplets or direct contact, aerosol (32,33)
Varicella zoster virus	VZV	Close contact with respiratory secretions, airborne or droplets (8,32)

Table 3. Association Between Paired Infections, with 95% Confidence Intervals (CI).

Likely shared main route of transmission			Unlikely shared main route of transmission		
Infections pair	$\bar{\varphi}$	95% CI	Infections pair	$\bar{\varphi}$	95% CI
HPY & TOX	0.634	(0.458, 0.810)	B19 & TOX	0.210	(-0.020, 0.440)
HAV & HPY	0.586	(0.454, 0.717)	HAV & VZV	0.202	(-0.066, 0.471)
HAV & TOX	0.531	(0.356, 0.706)	B19 & HAV	0.182	(0.010, 0.354)
B19 & VZV (Poland)	0.473	(0.335, 0.611)	B19 & HPY	0.159	(-0.023, 0.342)
EBV & HSV1	0.402	(0.316, 0.487)	B19 & CMV	0.129	(0.005, 0.253)
B19 & VZV (UK)	0.237	(0.138, 0.337)	TOX & VZV	0.075	(-0.270, 0.419)
MUM & RUB[1]	0.145	(0.084, 0.206)	CMV & RUB[1,2]	-0.017	(-0.174, 0.140)
B19 & VZV (Belgium)	0.106	(0.016, 0.197)	HPY & VZV	-0.050	(-0.409, 0.309)
B19 & VZV (Italy)	0.075	(-0.006, 0.157)			
B19 & RUB[1,2]	0.004	(-0.081, 0.089)			

[1]Males only [2] Age 11+ years

Table 4. Association Between Paired Infections at Age 21 Years and Older, with 95% Confidence Intervals (CI).

Likely shared main route of transmission			Unlikely shared main route of transmission		
Infections pair	$\bar{\varphi}$	95% CI	Infections pair	$\bar{\varphi}$	95% CI
HPY & TOX	0.518	(0.333, 0.704)	B19 & TOX	0.061	(-0.216, 0.338)
HAV & HPY	0.508	(0.373, 0.644)	HAV & VZV[1]	-0.111	(-0.489, 0.267)
HAV & TOX	0.437	(0.252, 0.622)	B19 & HAV	0.088	(-0.113, 0.288)
B19 & VZV[1] (Poland)	0.109	(-0.104, 0.322)	B19 & HPY	0.134	(-0.066, 0.334)
EBV & HSV1	0.219	(0.093, 0.346)	B19 & CMV	0.057	(-0.087, 0.202)
B19 & VZV[1] (UK)	0.200	(0.079, 0.322)	TOX & VZV[1]	0.004	(-0.461, 0.468)
MUM & RUB[2]	0.115	(0.038, 0.193)	CMV & RUB[1,2]	-0.024	(-0.199, 0.150)
B19 & VZV (Belgium)	0.084	(-0.091, 0.260)	HPY & VZV[1]	-0.148	(-0.594, 0.298)
B19 & VZV (Italy)	0.047	(-0.072, 0.167)			
B19 & RUB[1,2]	0.031	(-0.074, 0.136)			

[1]Age 11+ years   [2]Males only

Table 5. Associations with Polyomaviruses BKV and JCV in Individuals Aged 21 – 69 Years, with 95% Confidence Intervals (CI)

Infection		Polyomavirus BKV		Polyomavirus JCV	
		$\overline{\varphi}$	95% CI	$\overline{\varphi}$	95% CI
Cytomegalovirus	($n = 381$)	-0.0578	(-0.2393, 0.1237)	0.3239	(0.0769, 0.5710)
Parvovirus B19	($n = 755$)	0.1332	(0.0014, 0.2649)	0.0638	(-0.1003, 0.2279)
Rubella (males)	($n = 495$)	0.0562	(-0.1006, 0.2131)	-0.1287	(-0.3049, 0.0475)

Table 6. Impact of Heterogeneity on Basic Reproduction Number R_0 (and Critical Immunization Coverage π_c) for Varicella Zoster Virus in Four European Countries

Country	R_0 (π_c) ignoring heterogeneity	R_0 (π_c) with heterogeneity	Ratio of R_0 (odds ratio of π_c) with and without heterogeneity
Belgium	6.87 (0.85)	7.15 (0.86)	1.04 (1.08)
Italy	6.02 (0.83)	6.12 (0.84)	1.02 (1.08)
Poland	4.72 (0.79)	10.59 (0.91)	2.24 (2.69)
United Kingdom	3.25 (0.69)	5.19 (0.81)	1.60 (1.92)
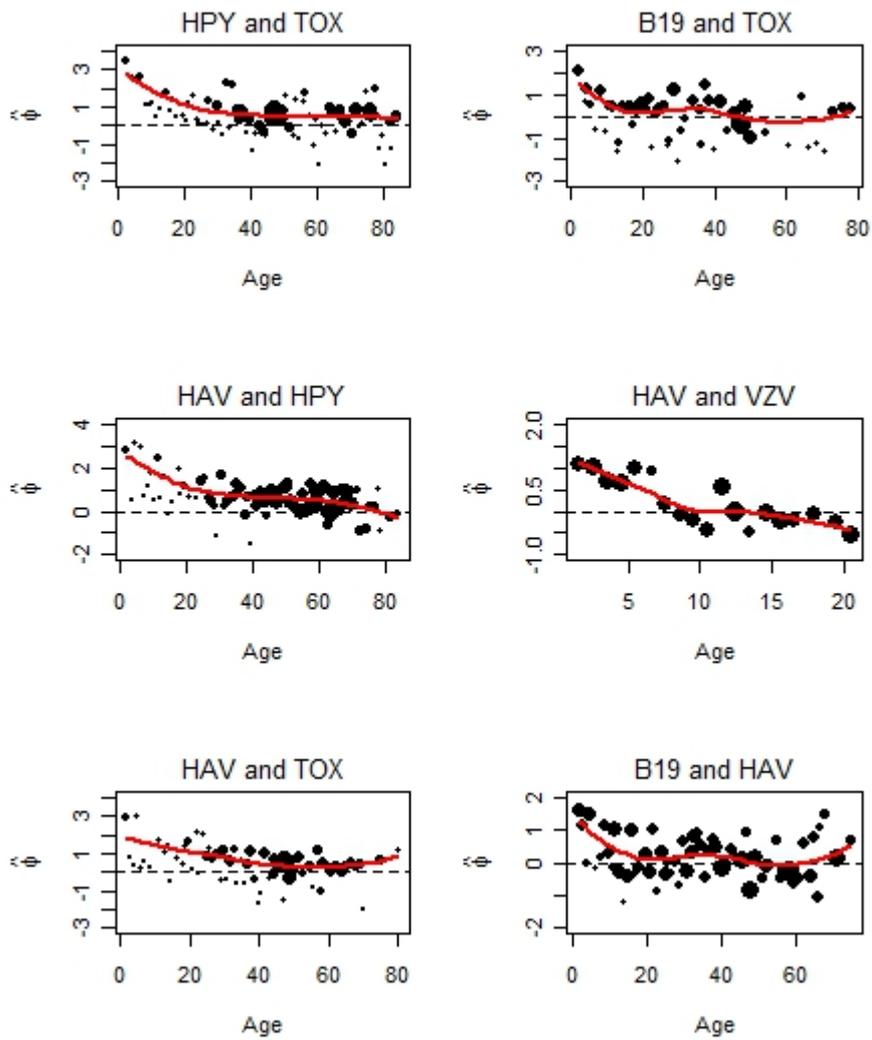
Figure 1. Association between times to infection. Top left: *Helicobacter pylori* and toxoplasma; top right: parvovirus B19 and toxoplasma; center left: hepatitis A virus and *Helicobacter pylori*; center right: hepatitis A virus and varicella zoster virus; bottom left: hepatitis A virus and toxoplasma; bottom right: parvovirus B19 and hepatitis A virus.
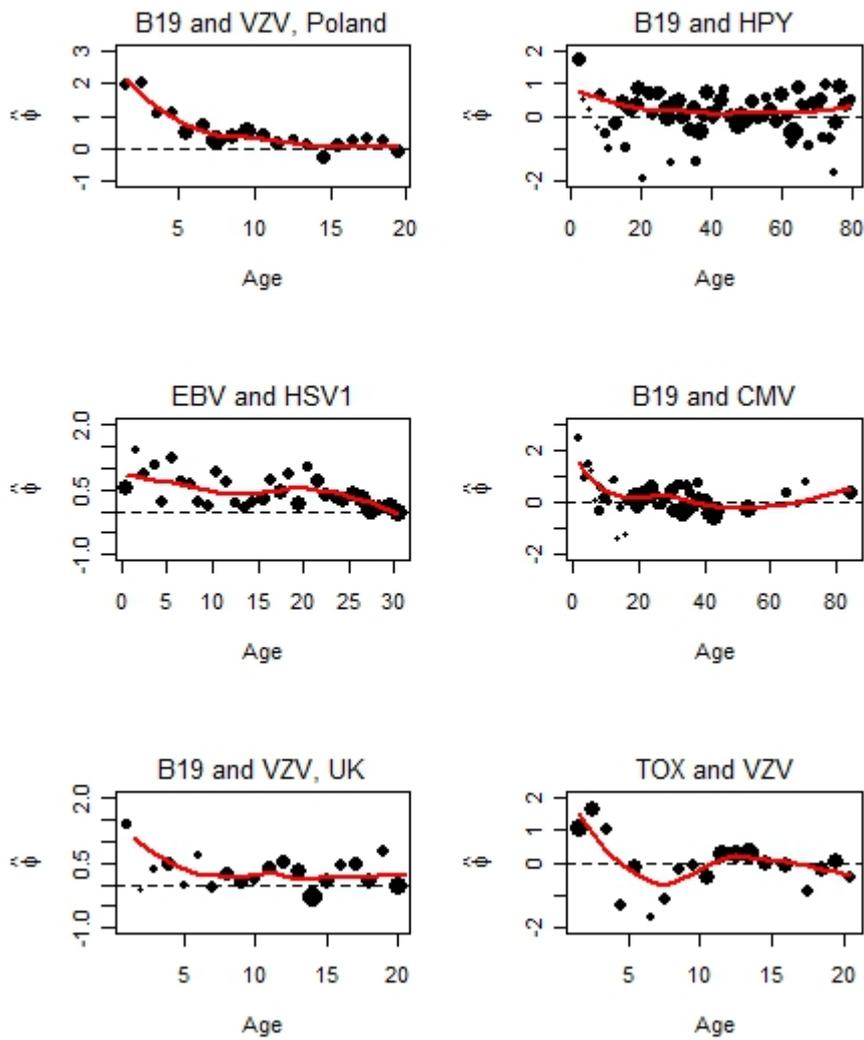
Figure 2. Association between times to infection. Top left: parvovirus B19 and varicella zoster virus in Poland; top right: parvovirus B19 and *Helicobacter pylori*; center left: Epstein-Barr virus and herpes simplex virus type I; center right: parvovirus B19 and cytomegalovirus; bottom left: parvovirus B19 and varicella zoster virus; bottom right: toxoplasma and varicella zoster virus.
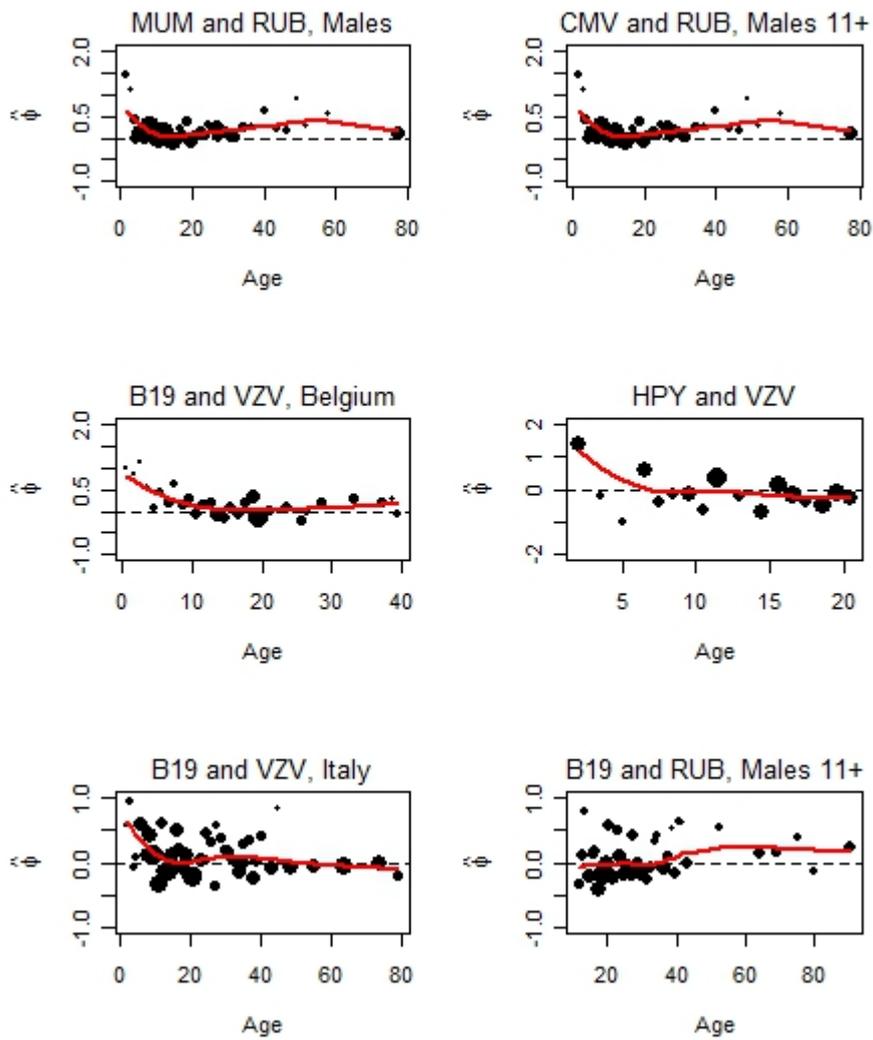
Figure 3. Association between times to infection. Top left: mumps virus and rubella virus in males; top right: cytomegalovirus and rubella in males aged 11 years or older; centre left: parvovirus B19 and varicella zoster virus in Belgium; centre right: *Helicobacter pylori* and varicella zoster; bottom left: parvovirus B19 and varicella zoster virus in Italy; bottom right: virus parvovirus B19 and rubella in males ages 11 years or older.
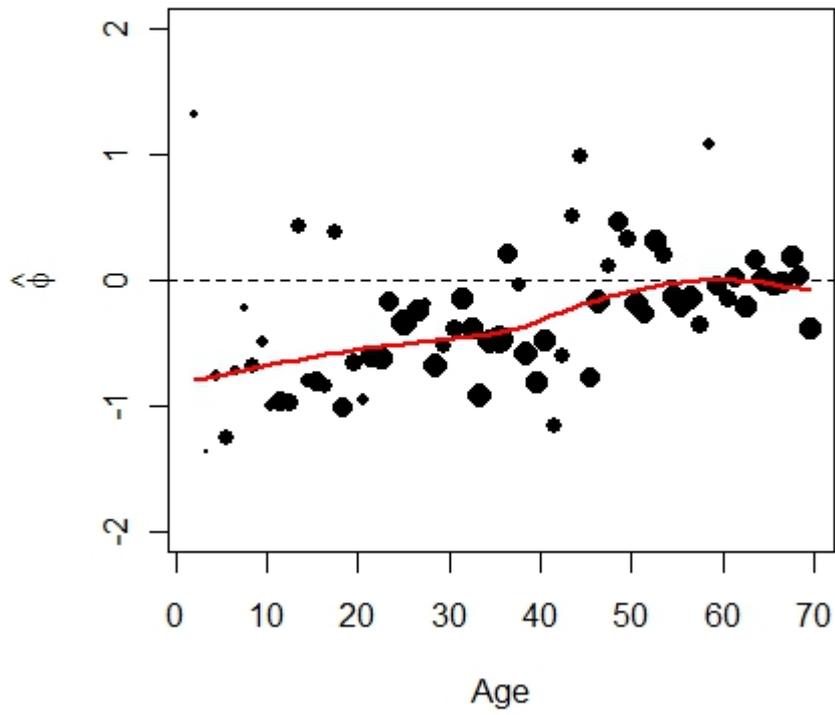
Figure 4. Association between times to infections for polyomavirus BKV and polyomavirus JCV.
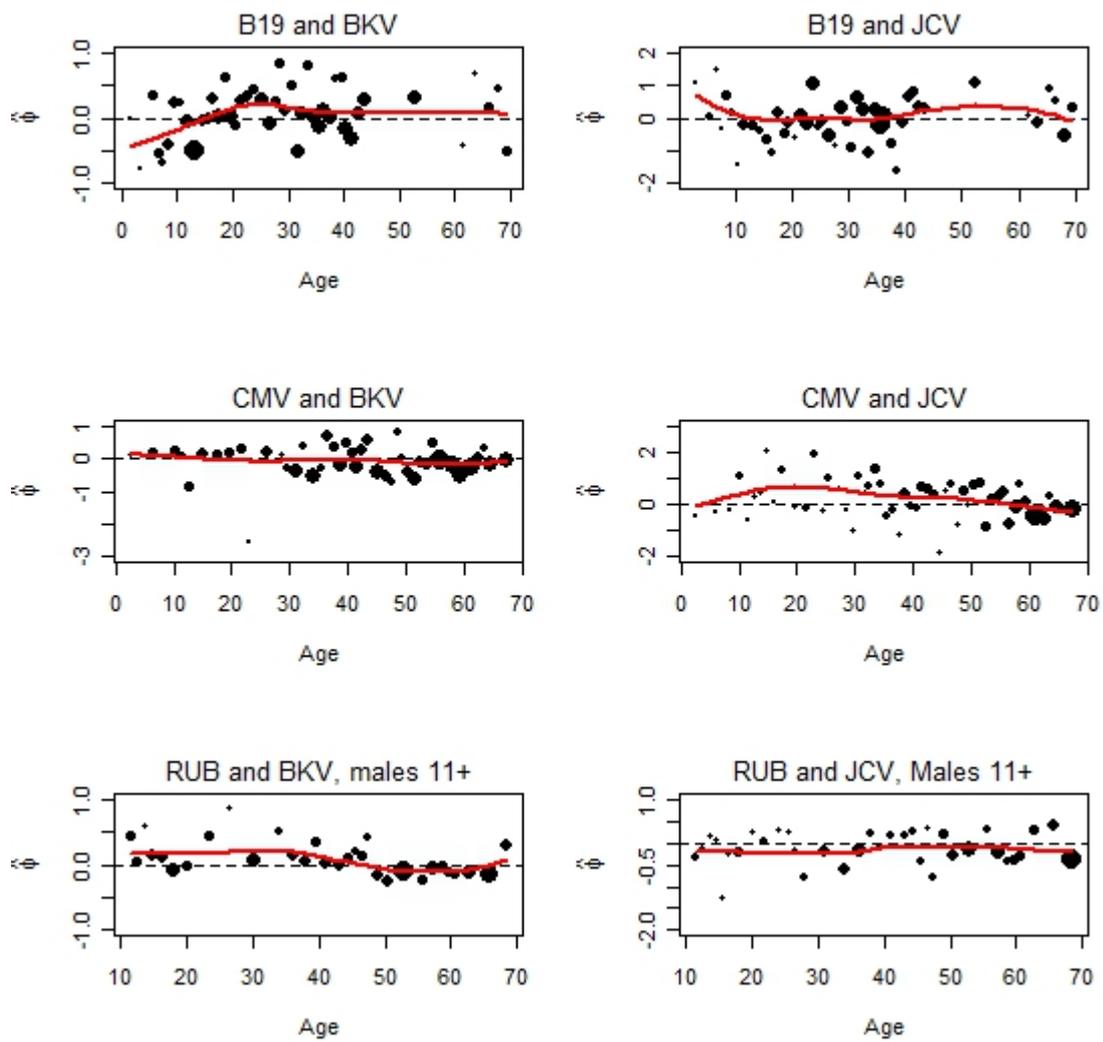
Figure 5. Association between time to infection for polyomaviruses BKV and JCV and selected infections. Top left: parvovirus B19 and Polyomavirus BKV; top right: parvovirus B19 and polyomavirus JCV; center left: cytomegalovirus and polyomarirus BKV; center right: cytomegalovirus and polyomavirus JCV; bottom left: rubella virus and polyomarirus BKV in males aged 11 years and over; bottom right: rubella virus and polyomavirus JCV in males aged 11 years and over.
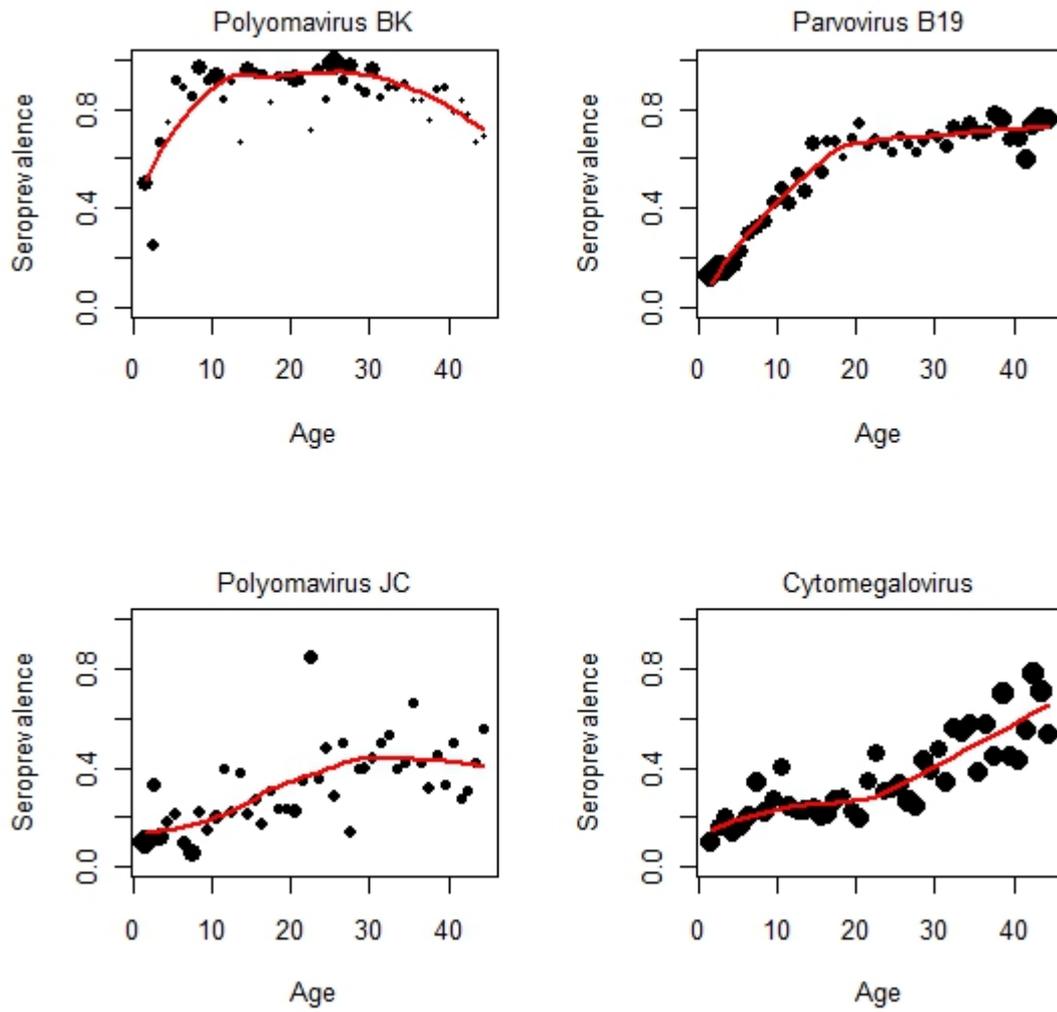
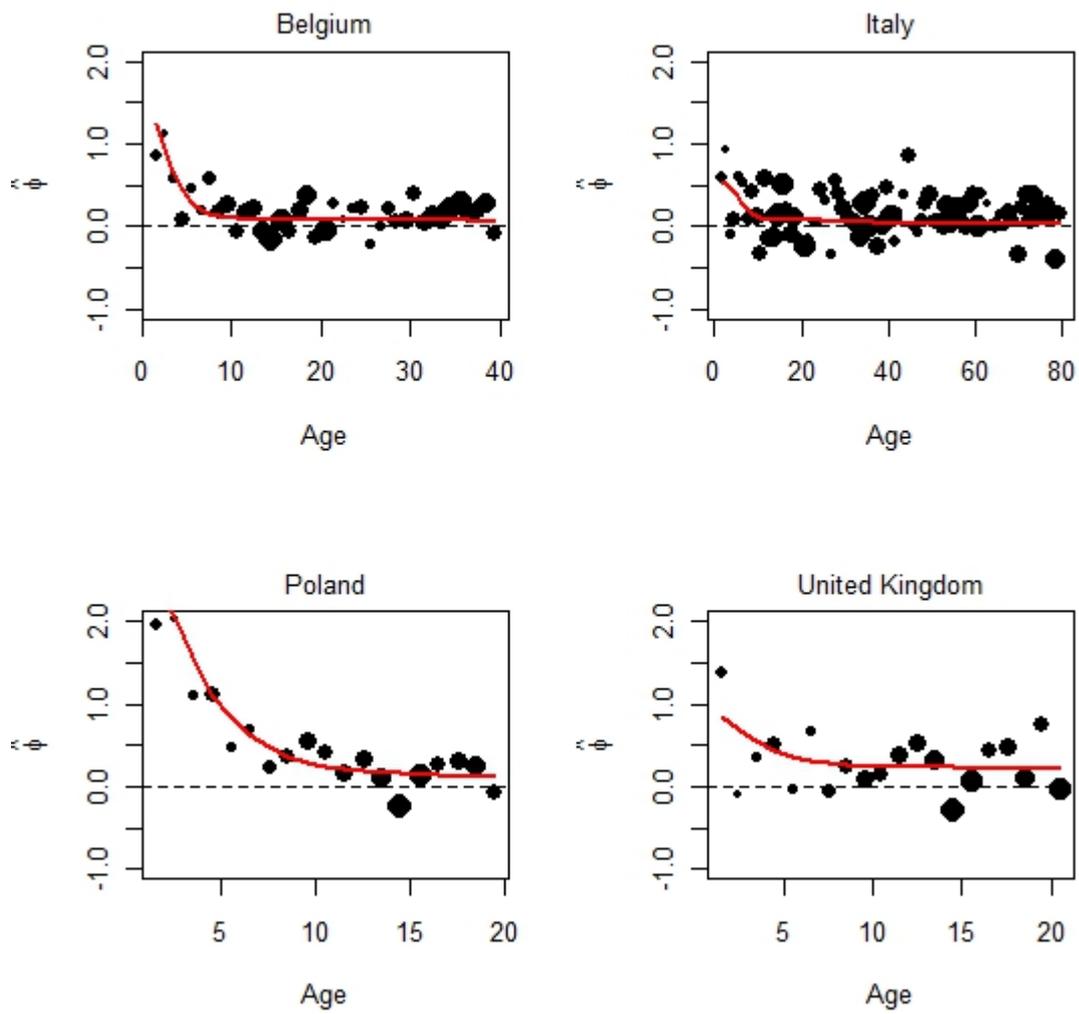Figure 6. Antibody prevalence by age (dots) and trend line (continuous curve).

Figure 7. Observed and fitted association between times to infection for parvovirus B19 and varicella zoster virus infections in four European countries.
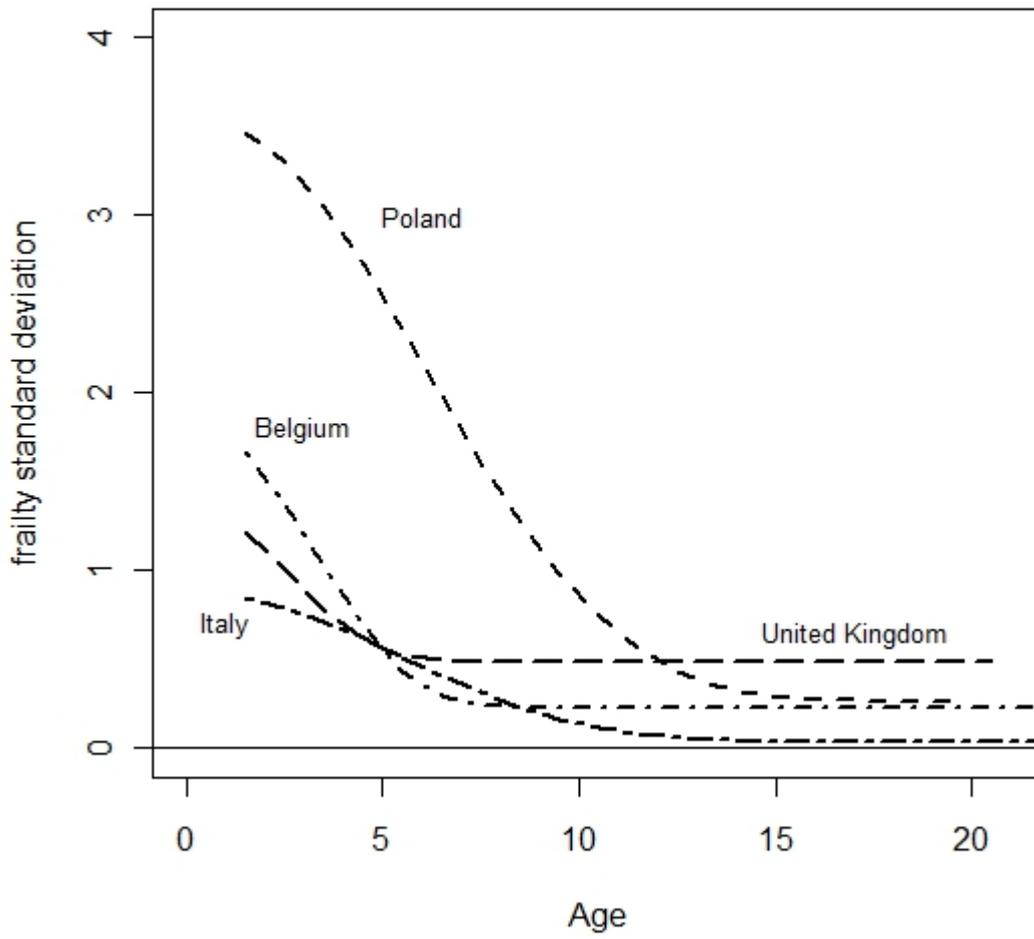
Figure 8. Standard deviation of the frailty associated with varicella zoster and parvovirus B 19 infection by age in four European countries.

WEB MATERIAL

This web material provides some more detail of the statistical methods used in the main text.

Incorporating age-dependent heterogeneity via frailty models

For simplicity assume throughout that u_x is a deterministic function of x and of a finite set of random variables $w_1, w_2, \ldots, w_k$. Thus $u_x = u(x, w_1, \ldots, w_k)$. Suppose that, as stated in equation 1, the effective contact rate takes the form

$$\beta(x, u_x ; y, v_y) = u_x \beta_0(x ; y) v_y .$$

If $I(y, u_y)$ is the number of infectious individuals of age y with activity level u_y, then the force of infection exerted on an individual of age x and activity level u_x is

$$\lambda(x, u_x) = \int_0^\infty \int_0^\infty \beta(x, u_x; y, u_y) I(y, u_y) du_y dy$$
$$= u_x \lambda_0(x),$$

where $\lambda_0(x)$ is the baseline force of infection and

$$\lambda_0(x) = \int_0^\infty \int_0^\infty \beta_0(x; y) u_y I(y, u_y) du_y dy.$$

Thus the assumption in equation 1 about the form of the contact rate leads to an age-dependent frailty model for the force of infection.

Displaying correlations between infections using bivariate serological survey data

The degree of heterogeneity in the population can be estimated from the strength of association in the 2 x 2 tables of counts at each age x. The details may be found in (16). For completeness, the association measure $\varphi(x)$ used in the paper is given here. It is defined as the value φ solving the implicit equation

$$\left(p_1(x)^{1-\exp(\varphi)} + p_2(x)^{1-\exp(\varphi)} - 1 \right)^{1/[1-\exp(\varphi)]} = p_{00}(x)$$

where $p_1(x)$ is the proportion of individuals of age x susceptible to infection 1, $p_2(x)$ is the proportion of individuals of age x susceptible to infection 2, and $p_{00}(x)$ is the proportion of individuals of age x susceptible to both infections.

Shared frailty models for paired serological survey data

Consider two infections labelled 1 and 2 with a shared age-dependent frailty and forces of infection described in equation 3. At each age x, the test results have probabilities

$$
\begin{aligned}
S_{00}(x) &= E\left\{\exp\left(-\int_0^x u_y[\lambda_{01}(y) + \lambda_{02}(y)]dy\right)\right\}, \\
S_{01}(x) &= E\left\{\exp\left(-\int_0^x u_y\lambda_{01}(y)dy\right)\right\} - S_{00}(x), \\
S_{10}(x) &= E\left\{\exp\left(-\int_0^x u_y\lambda_{02}(y)dy\right)\right\} - S_{00}(x), \\
S_{11}(x) &= 1 - S_{10}(x) - S_{01}(x) + S_{00}(x).
\end{aligned}
$$

The expectations in these expressions are with respect to the random variables $w_1, w_2,\ldots,w_k$ used to define u_y. The greater the heterogeneity represented by the variance of u_x, $\gamma(x)$, the greater the degree of association between the two serological outcomes.

Consider the model described in equation 5. Then

$$
S_{00}(x) = E_{w_1,w_2}\left\{\exp\left(-w_2\int_0^x[\lambda_{01}(y) + \lambda_{02}(y)]dy - w_2(w_1 - 1)\int_0^x h(y)[\lambda_{01}(y) + \lambda_{02}(y)]dy\right)\right\},
$$

with similar expressions for the other probabilities. For w_1 and w_2 independently gamma distributed with unit means and variances γ_1 and γ_2, respectively, the probabilities $S_{ij}(x)$ are obtained using a combination of analytical and numerical integration techniques.

For data $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$ on individuals of age x, the contribution to the Dirichlet-multinomial likelihood kernel from such individuals is

$$
\frac{\Gamma(\psi)}{\Gamma(n_x + \psi)} \prod_{i,j=0,1} \frac{\Gamma(n_{ijx} + \psi S_{ij}(x))}{\Gamma(\psi S_{ij}(x))}
$$

where $n_x = n_{00x} + n_{01x} + n_{10x} + n_{11x}$ and $\psi > 0$ is the dispersion parameter (in the limit as $\psi \to \infty$, the multinomial likelihood is retrieved; the model can be reparamaterised in terms of $\kappa = 1/(1+\psi)$ so that the multinomial likelihood corresponds to $\kappa = 0$). Individuals with data on only one infection (as occurred in some surveys) contributed a reduced likelihood kernel based on the appropriate 2-way margin. The observations in different age groups are treated as independent and the overall log-likelihood kernel is obtained by summing the terms over all ages x.

Impact of variation by calendar time

If the average effective contact rate varies with calendar time t, as is the case with some infections, notably those transmitted by the fecal-oral route, then equation 1 generalises to

$$\beta(x,u_x ; y,v_y ; t) = u_x \beta_0(x ; y ; t) v_y ,$$

the key additional assumption being that the distributions of the activity levels are not time-dependent. The number of infectious individuals is then of the form $I(y,u_y,t)$, and a similar argument to that used before leads to the frailty model

$$\lambda(x,t) = u_x \lambda_0(x,t),$$

where the baseline force of infection is now time-dependent. However, it is not possible to separate age and calendar time effects from a single serological survey: for this purpose serial surveys are required. Suppose that data from a single survey at time t_0 are available and that calendar-time variation is ignored in the analysis. Let $\lambda^*(x)$ denote the baseline force of infection and $h^*(x)$ the variation in the frailty variance estimated in this way. Then

$$\int_0^x \lambda(y,t_0 - x + y)dy = \int_0^x \lambda^*(y)dy,$$

$$\int_0^x h(y)\lambda(y,t_0 - x + y)dy = \int_0^x h^*(y)\lambda^*(y)dy.$$

Differentiating both sides with respect to x and rearranging gives

$$h^*(x) \quad = \quad h(x) + \frac{\int_0^x [h(x)-h(y)]\frac{\partial\lambda(y,t)}{\partial t}\Big|_{t_0-x+y}\,dy}{\lambda(x,t_0)-\int_0^x \frac{\partial\lambda(y,t)}{\partial t}\Big|_{t_0-x+y}\,dy}.$$

Suppose that $\lambda(x,t)$ declines with t, as is the case with most infections transmitted by the fecal-oral route in developed countries. Then if $h(x)$ is constant, so is $h^*(x)$ and $h^*(x) = h(x)$. More generally, the error is bounded by

$$\left|h^*(x)-h(x)\right| \quad \le \quad \frac{-\int_0^x \frac{\partial\lambda(y,t)}{\partial t}\Big|_{t_0-x+y}\,dy}{\lambda(y,t_0)-\int_0^x \frac{\partial\lambda(y,t)}{\partial t}\Big|_{t_0-x+y}\,dy}.$$

This upper bound is small for values x of interest (typically $5-10$ years) provided that the decline in the force of infection is small in the years immediately prior to t_0. This is the case for the infections considered here, for which the decline in incidence occurred well before the survey was undertaken.