Submitted to the Annals of Applied Statistics
2016 Vol. 0, No. 0, 1–23

EFFICIENT ESTIMATION OF AGE-SPECIFIC SOCIAL CONTACT RATES BETWEEN MEN AND WOMEN

By Jan van de Kassteele[*,†], Jan van Eijkeren[†] and Jacco Wallinga[‡]

Department of Statistics, Informatics and Mathematical Modeling. National Institute for Public Health and the Environment (RIVM)[†] and Epidemiology and Surveillance Unit. National Institute for Public Health and the Environment (RIVM)[‡]

Social contact patterns reveal with whom individuals tend to socialize, and therefore to whom they transmit respiratory infections. We infer highly detailed age-specific contact rates between the sexes using a hierarchical Bayesian model that smooths while simultaneously guarantees the inherent reciprocity of contact rates. Application of this approach to social contact data from a large prospective survey confirms a tendency that people, especially children and adolescents, mostly contact other people of their own age and sex, and reveals that women have more contact with children than men. These findings imply different exposure patterns between the two sexes for specific age groups, which agrees with available observations.

1. Introduction. The incidence and severity of most human respiratory infections such as influenza, tuberculosis, measles, rubella, mumps, human parvovirus B19 and cytomegalovirus depend on age and sex (Falagas, Mourtzoukou and Vardakas, 2007; Klein et al., 2010; Holmes, Hausler and Nunn, 1998; Borgdorff et al., 2000; Neyrolles and Quintana-Murci, 2009; Brown and Moss, 2010; Davis et al., 2010; Young and Brown, 2004; Pass et al., 2009). For the planning and evaluating of vaccination programs against these infectious diseases it is necessary to know how an infection spreads among age groups and among women and men. At-risk contacts for infection with respiratory pathogens are typically assessed via proxy measures, such as having a conversation or touching (Edmunds, O'Callaghan and Nokes, 1997; Wallinga, Teunis and Kretzschmar, 2006; Mossong et al., 2008; Kucharski et al., 2014). A quantitative understanding of the different contact behavior of men and women of various ages will thus contribute to a better understanding of how respiratory infections spread.

[*]Corresponding author

Keywords and phrases: Social contact patterns, Hierarchical Bayesian model, Gaussian Markov Random Field, Integrated Nested Laplace Approximations, Infectious disease transmission

Even though sex differences in contact behavior have been a topic of general interest, there is surprisingly little scientific evidence to quantify the stereotypical gender roles of men and women (Mehl et al., 2007). Mossong et al. (2008) found no evidence of the daily number of social contacts being different for men and women when averaging over all ages. Further stratification by age revealed a strong tendency of girls contacting girls rather than boys, and vice versa in elementary schools (Cauchemez et al., 2011; Conlan et al., 2011). A recent study reported strong evidence of within-sex preferential mixing for broad age groups (Dodd et al., 2016). However, to the best of our knowledge, there are no studies that offer precise estimates of the numbers of contacts by age and sex for a representative study population.

There is a large number of studies that collect data on at-risk contacts for infection using proxy measures such as having a conversation or touching (Edmunds, O'Callaghan and Nokes, 1997; Wallinga, Teunis and Kretzschmar, 2006; Mossong et al., 2008; Salath et al., 2010; Read et al., 2012; Danon et al., 2013; Kucharski et al., 2014; Kwok et al., 2014; Eames et al., 2015; Dodd et al., 2016, e.g.); for textbooks that cover this field see Vynnycky and White (2010); Hens et al. (2012). Most of these studies follow the format of Mossong et al. (2008) and collect data on social contact behavior stratified by age and sex. These data have been used to infer contact rates that are relevant to the spread of a respiratory infection has been explored using mathematical modeling studies (Medlock and Galvani, 2009; Miller et al., 2010; Rohani, Zhong and King, 2010; Keeling and White, 2011, e.g.). In all cases, the estimation has been limited to contact rates that are only age-specific, rather than age and sex specific.

The estimation of contact rates by both age and sex from these data sets has proven to be statistically challenging for two reasons. First, the stratification by age and sex of both the study participants and their contacts leads to a very large number of contact rates (model parameters) to be estimated. For example, if we would choose to estimate the contacts by men and women in 81 age cohorts (age 0 - 80) to cover the age range in the general population, this would require estimating $(2 \times 81)^2 = 26,244$ contact rates. Second, the contact rates should also meet a reciprocity requirement. This requirement arises from the definition of a contact as a reciprocal event where two individuals have a conversation or touch. At the individual level, this implies that if John contacts Mary, Mary must have contacted John. Both issues, the need for regularization of the number of parameters while constraining to meet reciprocity, could result into computational problems.

There are relatively few statistical approaches to inferring contact rates from social contact data (see Hens et al. (2012), ch. 15 for an overview).

Most statistical analyses of social contact data, such as the one presented in Mossong et al. (2008) do not guarantee reciprocity, and therefore risk generating internally inconsistent outcomes. A straightforward regularization approach is to aggregate contact data into wide age categories and apply a likelihood function with constraints such as to guarantee reciprocity of contacts (Wallinga, Teunis and Kretzschmar, 2006). Aggregating age into a few categories results in coarse-graining, with an inevitable loss of detail. A better regularization option is to use this plate regression splines with a smooth-then-constrain approach (Hens et al., 2009; Goeyvaerts et al., 2010). But also in this case there is a considerable risk of losing details.

Recent developments in computational intensive statistical methods present us the alternative possibility of using hierarchical Bayesian models. From a Bayesian point of view, regularization techniques correspond to imposing certain prior distributions on model parameters that give lower probability to more complex models. Here we estimate social contact rates by age and sex using an innovative estimation scheme based on a Gaussian Markov Random Field (GMRF) in a hierarchical Bayesian model, controlled by a few hyperparameters and with a non-Gaussian response variable (Rue and Held, 2005). A tailor-made construction of the precision matrix of the GMRF prior allows us to impose smoothness in contact rates while simultaneously accounting for the reciprocity of contacts. Such models are computationally very efficient when using Integrated Nested Laplace Approximations (INLA) (Rue, Martino and Chopin, 2009). We apply this method to contact data from a prospective survey of social contact patterns collected in the Netherlands with 825 participants reporting 11,225 contacts. We show that it is possible to extract contact rates from this dataset at an unprecedented level of detail. Finally, we explore the consequences for the spread of a respiratory infection via such contacts and examine how men and women differ in their age-specific risk of infection.

2. Methodology.

2.1. *Definitions and notation.* In this section we present the notation that is used to estimate the daily age and sex-specific contact rates from a contact survey. Following Mossong et al. (2008), we define a participant as someone who participated in the prospective contact survey. We define a contact as a conversation (i.e., at least one sentence) between a participant and another person at close physical proximity or touching the other person's skin (e.g., shaking hands or kissing). Following Hens et al. (2012), we use subscripts i as an index for a participant's age, where $i = 1$ and $i = 81$ corresponds to 0 and 80 years of age, respectively. Similarly, j is used as an

index for a contacted person's age between $j = 1$ and $j = 81$. We introduce superscripts MM as an index to refer to male-to-male contacts, FM to refer to female-to-male contacts, MF to refer to male-to-female contacts, and FF to refer to female-to-female contacts. Therefore, in total we have four blocks of 81×81 contacts. Let us, for the moment, only consider male-to-female contacts MF.

We denote the total number of unique individuals contacted by all participants by random variable Y. Then we can define Y_{ij}^{MF} as the total number of females of age j that are contacted by all male participants of age i during one day. We denote the total number of participants by t. Then we can define t_i^M as the total number of male participants of age i in the contact survey. We denote the contact intensity between two groups of individuals by m. Then we can define m_{ij}^{MF} as the mean number of females of age j that are contacted by one male participant of age i during one day. It is given by:

$$m_{ij}^{MF} = \frac{\mathrm{E}(Y_{ij}^{MF})}{t_i^M}. \tag{2.1}$$

We denote the total population size by w. Then we can define w_i^M as the male population of age i, and w_j^F as the female population of age j. Because contacts are reciprocal by nature, the total number of male to female contacts from age i to age j should be equal to the total number female to male contacts from age j to age i during one day:

$$m_{ij}^{MF} w_i^M = m_{ji}^{FM} w_j^F. \tag{2.2}$$

We denote the contact rate between two groups of individuals by c. Then we can define c_{ij}^{MF} as the mean number of female individuals of age j that are contacted by one male participant of age i during one day divided by the population number of females of age j. This is given by:

$$c_{ij}^{MF} = \frac{m_{ij}^{MF}}{w_j^F} = \frac{\mathrm{E}(Y_{ij}^{MF})}{t_i^M w_j^F}. \tag{2.3}$$

From equation 2.2 it follows that $c_{ij}^{MF} = c_{ji}^{FM}$. Similar arguments give $c_{ij}^{MM} = c_{ji}^{MM}$ and $c_{ij}^{FF} = c_{ji}^{FF}$. That is, contact rates c are symmetric, contact intensities m are not.

2.2. *Inferring contact rates from reported contacts.* We infer contact rates using a hierarchical Bayesian model with three levels. The first level, the observation level, refers to the total number of contacts of any age and

sex-specific combination in the dataset. A reasonable assumption is to use a Negative Binomial distribution (Wallinga, Teunis and Kretzschmar, 2006; Mossong et al., 2008; Goeyvaerts et al., 2010) with mean $\mathrm{E}(Y_{ij}^{MF})$ and global dispersion parameter θ:

$$Y_{ij}^{MF} \mid \mathrm{E}(Y_{ij}^{MF}), \theta \sim \mathrm{NegBin}[\mathrm{E}(Y_{ij}^{MF}), \theta]. \tag{2.4}$$

At the second level, from equation 2.3 it follows that the mean of the total number of contacts $\mathrm{E}(Y_{ij}^{MF})$ is the product of a known denominator $U_{ij}^{MF} = t_i^M w_j^F$ and the unknown contact rate c_{ij}^{MF}. The denominator represents the total number female contacts that all male participants of age i would have if they contacted all female individuals of age j in the population. Because contact rates are positive, it is natural to use the log-link function:

$$\log[\mathrm{E}(Y_{ij}^{MF})] = \log(U_{ij}^{MF}) + \beta + x_{ij}^{MF}. \tag{2.5}$$

Hence, $\log(c_{ij}^{MF})$ is modeled by $\beta + x_{ij}^{MF}$, where we can interpret β as a global intercept and x_{ij}^{MF} as deviations from it. These deviations have a smooth and symmetric structure. In our approach x_{ij}^{MM}, x_{ij}^{MF}, x_{ij}^{FM} and x_{ij}^{FF} are considered a realization of a zero mean two-dimensional Gaussian Markov Random Field (GMRF) over the ages i and j and both sexes. A GMRF is a random field following a multivariate Normal (Gaussian) distribution with conditional (Markov) independence assumptions (Rue and Held, 2005). This conditional independence is defined by a precision matrix $\boldsymbol{Q} = \tau \boldsymbol{R}$, where $\boldsymbol{R}$ is a sparse structure matrix that will be defined more precisely in section 2.3 and τ is the global precision parameter that controls the smoothness of the deviations. If we stack x_{ij}^{MM}, x_{ij}^{MF}, x_{ij}^{FM} and x_{ij}^{FF} in a vector $\mathrm{vec}(\boldsymbol{x})$, then we can write:

$$\mathrm{vec}(\boldsymbol{x}) \mid \tau \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{Q}). \tag{2.6}$$

At the third level, hyper priors are specified for intercept, the precision (smoothing) parameter of the GMRF and the dispersion parameter of the Negative Binomial distribution. We place a Normal prior with mean 0 and precision 0.001 on the parameter for the intercept β, a Gamma prior distribution with shape parameter 1 and rate parameter 0.0001 on precision parameter τ, and a Normal prior with mean 0 and precision 0.001 on the logarithm of the dispersion parameter θ:

$$\beta \sim \mathrm{Normal}(0, 0.001), \tag{2.7}$$

$$\tau \sim \mathrm{Gamma}(1, 0.0001), \tag{2.8}$$

$$\log(\theta) \sim \mathrm{Normal}(0, 0.001). \tag{2.9}$$

The contact rates and parameters for this highly structured hierarchical Bayesian models can be efficiently estimated by the recently established Integrated Nested Laplace Approximations technique (INLA) (Rue, Martino and Chopin, 2009). The implementation can be found in the Supplementary material.

2.3. *Smoothing while maintaining symmetry.* A tailor-made construction of the precision matrix $\boldsymbol{Q}$ of the GMRF prior allows us to impose smoothness in contact rates while simultaneously accounting for the reciprocity of contacts. It is essential that the contact rates are symmetric, such that $c_{ij}^{MM} = c_{ji}^{MM}$, $c_{ij}^{MF} = c_{ji}^{FM}$ and $c_{ij}^{FF} = c_{ji}^{FF}$. These symmetries for $\boldsymbol{c}$ imply a symmetry in the deviations $\boldsymbol{x}$. Figure 1 provides a schematic illustration for two sexes and $n = 5$ age groups: the $5 \times 5 \times 4 = 100$ elements correspond to the 100 nodes in this graph; symmetry in value for each element is guaranteed by forcing identical values in the lower and upper triangular parts of the matrix. For example, data records 2 and 6 provide information for node 2. Thus, 55 unique node values are inferred from the 100 data records.

Smoothing is achieved by imposing the condition that neighboring node values of $\boldsymbol{x}$ should be similar. The neighborhood structure is defined by the entries of the structure matrix. Figure 1 provides a schematic illustration: the non-zero elements of structure matrix $\boldsymbol{R}$ correspond to the edges in this graph. Only the nodes in the lower triangular part of the matrix need to be linked, the values for the nodes in the upper triangular part follow directly because of the imposed symmetry.

We use the second order random walk prior (RW2) (Rue and Held, 2005), which reflects the prior belief that the gradient of $\boldsymbol{x}$ varies smoothly and that sudden jumps between neighboring values of the gradient are unlikely, in other words, regularize the difference of the differences. Other options include the RW1 prior, which reflects the prior belief that sudden jumps between neighboring values are unlikely, or higher order RW priors. For the RW2 prior in one dimension a Normal prior is put on the second order differences:

$$\Delta^2 x_i = x_{i-1} - 2x_i + x_{i+1} \sim \text{Normal}(0, \tau). \tag{2.10}$$

Let us for the moment consider the FM contacts with $n = 5$ age groups, as in the lower right panel of Figure 1. Here $\boldsymbol{x}^{FM}$ is a 5×5 matrix. In two dimensions, smoothness is achieved by placing the RW2 prior on the rows and columns of $\boldsymbol{x}^{FM}$ simultaneously. We follow the approach by Currie, Durban and Eilers (2004), using Kronecker products to construct a two-dimensional prior.

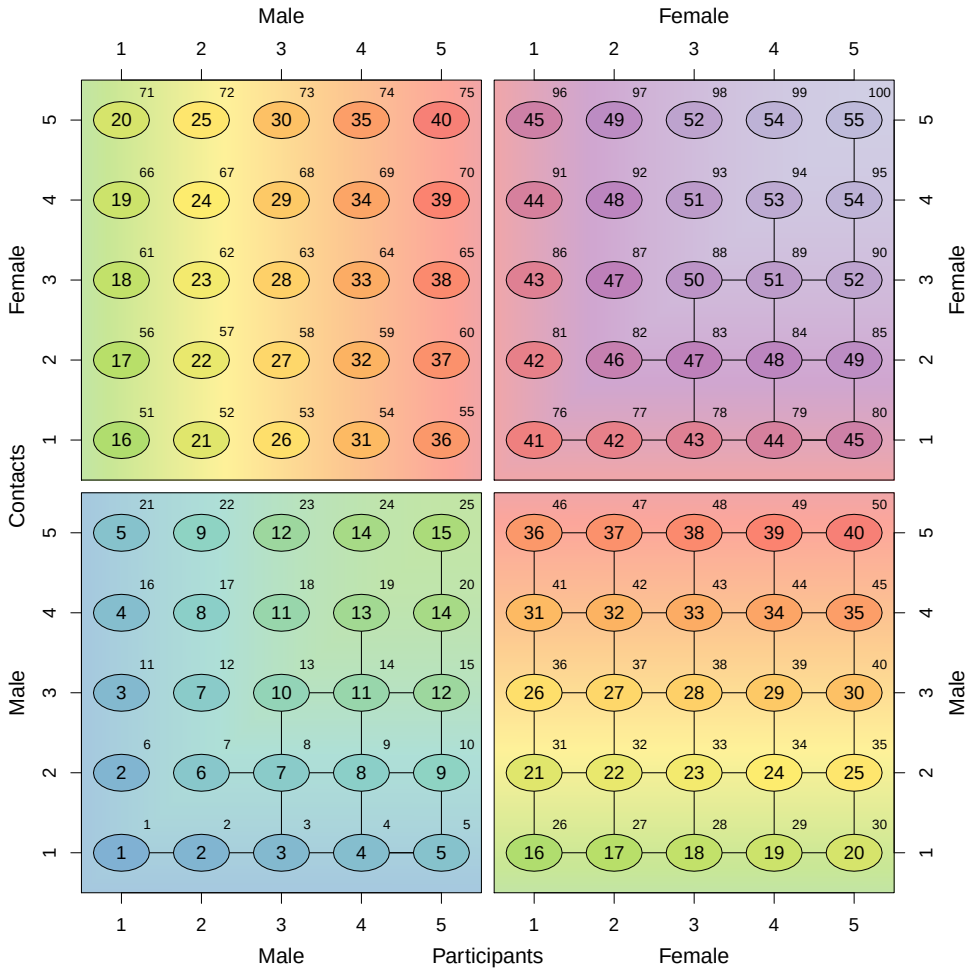

FIG 1. *Graphical representation of the construction of a smooth and symmetric contact matrix stratified by age ($n = 5$ age groups) and sex. Contacts are categorized as male-to-male (bottom left panel), female-to-male (bottom right panel), male-to-female (top left panel), and female-to-female (top right panel). The horizontal axis of each panel gives the age of the participants; the vertical axis gives the age of the contacts. The data records, indicated by superscripts, are numbered sequentially 1 to 100. The nodes, indicated by ellipses, are numbered sequentially 1 to 55 and ordered such that symmetry between ages and sexes is guaranteed. For illustrative purpose, identical nodes are indicated by identical colors. The edges denote the dependencies between triplets of nodes (RW2 prior).*

Let $\boldsymbol{D}_0$ be the matrix form of the one-dimensional difference operator in equation 2.10. For $n = 5$ it has size 3×5, because near the two boundaries it is impossible to take second order differences. The difference operator matrix operating on all rows and columns of $\boldsymbol{x}^{FM}$ simultaneously can then be written as $[\boldsymbol{D}_1 : \boldsymbol{D}_2]^{FM}$, with

$$\boldsymbol{D}_1 = \boldsymbol{I}_n \otimes \boldsymbol{D}_0 \text{ and } \boldsymbol{D}_2 = \boldsymbol{D}_0 \otimes \boldsymbol{I}_n. \tag{2.11}$$

$\boldsymbol{D}_1$ operates in horizontal direction along the age of the participants, while $\boldsymbol{D}_2$ operates in vertical direction along the age of the contacts. Both $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ have size 15×25, so $[\boldsymbol{D}_1 : \boldsymbol{D}_2]^{FM}$ has size 30×25 and operates on $\text{vec}(\boldsymbol{x}^{FM}) = (x_{1,1}^{FM}, x_{2,1}^{FM}, \ldots, x_{5,5}^{FM})$ of length 25.

Next, we consider the *MM* contacts with $n = 5$ age groups, as in the lower left panel of Figure 1. Using equation 2.11, we can construct a similar difference operator matrix $[\boldsymbol{D}_1 : \boldsymbol{D}_2]^{MM}$ for $\text{vec}(\boldsymbol{x}^{MM})$. However, because of the imposed symmetry, we only need to estimate the lower triangular part (including the diagonal), so $\text{vec}(\boldsymbol{x}^{MM})$ is a vector of length 15. As it is impossible to take second order differences near the boundaries of the triangle (the missing edges between nodes 2 and 6, and 13 and 14 in Figure 1), we can subsequently drop the corresponding rows and columns of $[\boldsymbol{D}_1 : \boldsymbol{D}_2]^{MM}$, which then reduces to a 12×15 matrix. The same applies to the *FF* contacts. No difference operator matrix is needed for the *MF* contacts because of the imposed symmetry.

The three resulting difference operator matrices are put together in one block diagonal difference operator matrix $\boldsymbol{D}$ of size 54×55. This matrix operates on all 55 elements in $\boldsymbol{x}$ simultaneously. The corresponding 55×55 structure matrix $\boldsymbol{R}$ of the GMRF is given by $\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}$ (Rue and Held, 2005). The resulting prior is improper because the precision matrix $\boldsymbol{Q}$ is not of full rank (rank deficiency is $3 \times 2 \times 2 = 12$). The posterior, however, is proper. A similar principle applies to n age groups and other RW priors. The implementation can be found in the Supplementary material.

2.4. *Model choice and validation.* We examine the effect of applying RW1, RW2 and RW3 priors. The models are compared in terms of the Watanabe-Akaike or widely applicable information criterion (WAIC) (Watanabe, 2013). We compute the probability integral transform (PIT) to check the validity of the models (Dawid, 1984). The sensitivity of the outcome to the particular choice of hyper priors is examined using the methodology described in Roos et al. (2015).

The WAIC closely approximates Bayesian cross-validation and can be viewed as an improvement on the deviance information criterion (DIC) for

Bayesian models (Gelman, Hwang and Vehtari, 2014). The PIT can be used as a Bayesian 'leave-one-out' predictive measure of fit or calibration check. If the observation is drawn from the predictive distribution, the PIT has a standard uniform distribution. The PIT is usually being visualized in histograms. We use the nonrandomized version of the PIT that is suitable for count data (Czado, Gneiting and Held, 2009). The sensitivity to the hyper priors is examined by comparing the local change in the posterior parameter distribution to the unmodified posterior, in case the prior distribution is modified in a standardized way. This can be done without rerunning the model (Roos et al., 2015). The sensitivity is summarized by a single number, here the worst-case sensitivity, expressed as a percentage. Percentages above 100% indicate super-sensitivity.

3. Application.

3.1. *Contact survey data.* We use data from a prospective survey of social contact patterns to illustrate our method. Data were collected within the POLYMOD multi-country study. The goal of this study was to quantify contact behavior relevant for the spread of infections by the respiratory or close-contact route. A detailed description of the study design has been provided in Mossong et al. (2008). Because of differences in data collection between countries, we only use data from the Netherlands, in which 825 participants recorded characteristics of 11,225 contacts with unique individuals during one day. As the multi-country dataset was published while data collection was still going on in the Netherlands, only 269 of these 825 Dutch participants were included in the study published by Mossong et al. (2008). The additional 556 participants have been included in our analysis. Data collection took place in 2006 and 2007.

Participants were asked to complete a diary on one assigned day on the individuals with whom they had contact, as defined in section 2.1. Participants were asked to record their own age and sex, as well as the age and sex of each contacted person. There were three different types of diaries: one for children (age 0 - 8), which was completed by their parents; one for teens (age 9 - 17); and one for adults (age 18+).

Some participants reported the age of contacts as a range. We multiple imputed (10 times) these records by uniformly sampling an age from that range. From age 20 onwards, the reported age of contacts showed a preference at ages that were multiples of five. To prevent spurious results, we corrected these ages by uniformly redistributing the peak in an age range between two years younger and two years older. The additional uncertainty associated with the multiple imputations are included in all results.

Because only 22 participants reported the maximum number of contacts of 45, we ignored possible right censoring of the number of contacts. We only analyzed the reported contacts and ignored the missed contacts (see Supplementary material). Four participants reported not having any contacts. We excluded records where there was no information on the age or sex of participants or contacts. We also excluded contacts older than 80 years, because only individuals of age 0 - 80 participated in the contact survey. For these reasons, in total 53 participants and 1,037 contacts were excluded.

The total observed number of contacts of any age and sex-specific combination, y_{ij}^{MM}, y_{ij}^{FM}, y_{ij}^{MF} and y_{ij}^{FF}, is found by cross tabulation of the participant ID's in the data set, stratified by participant age, contact age, participant sex and contact sex. The total number of participants of any age and sex-specific combination, t_i^M and t_i^F, is found by tabulation of the number of unique participant IDs, stratified by participant age and sex. The age and sex-specific population numbers w_j^M and w_j^F with reference date January 1, 2007, are obtained from Statistics Netherlands (StatLine, 2015).

The three observables $\boldsymbol{y}$, $\boldsymbol{t}$ and $\boldsymbol{w}$ are assembled into one large dataset. For each record the denominator U_{ij} is calculated. The value of the denominator, which is typically $\mathcal{O}(10^6)$, is scaled by dividing the value by one million contacts. If there are no participants of age i, the number of contacts y_{ij} is set to a missing value and the value of U_{ij} is set to 1. In that way, these records do not contribute to the likelihood in the estimation procedure. Details are found in the Supplementary material.

3.2. *Crude and smoothed contact rates.* The crude age and sex-specific contact rates $\boldsymbol{c}$ are shown in Figure 2. With "crude" we mean contact rates that are directy calculated from the data without applying any regularization. They are obtained by equation 2.3, where $\mathrm{E}(\boldsymbol{Y})$ is replaced by $\boldsymbol{y}$. There are no male participants of age 25 and 80, and no female participants of age 80, resulting in white vertical lines (not visible at age 80). The values of the crude contact rates fluctuate, are often equal to zero and are not necessarily symmetric.

Figure 3 shows the age and sex-specific contact rates after smoothing while accounting for reciprocity, using the approach as described in section 2, using the RW2 prior. Reciprocity of contact imposes symmetry on the four panels, with the axis of symmetry running diagonal from the bottom left to the top right. The figure reveals highly structured contact rates that strongly depend on age. Higher rates occur along the diagonals. This indicates that contacts are mostly assortative with respect to age: people contact other people of their own age. Children and adolescents have the highest rates. The

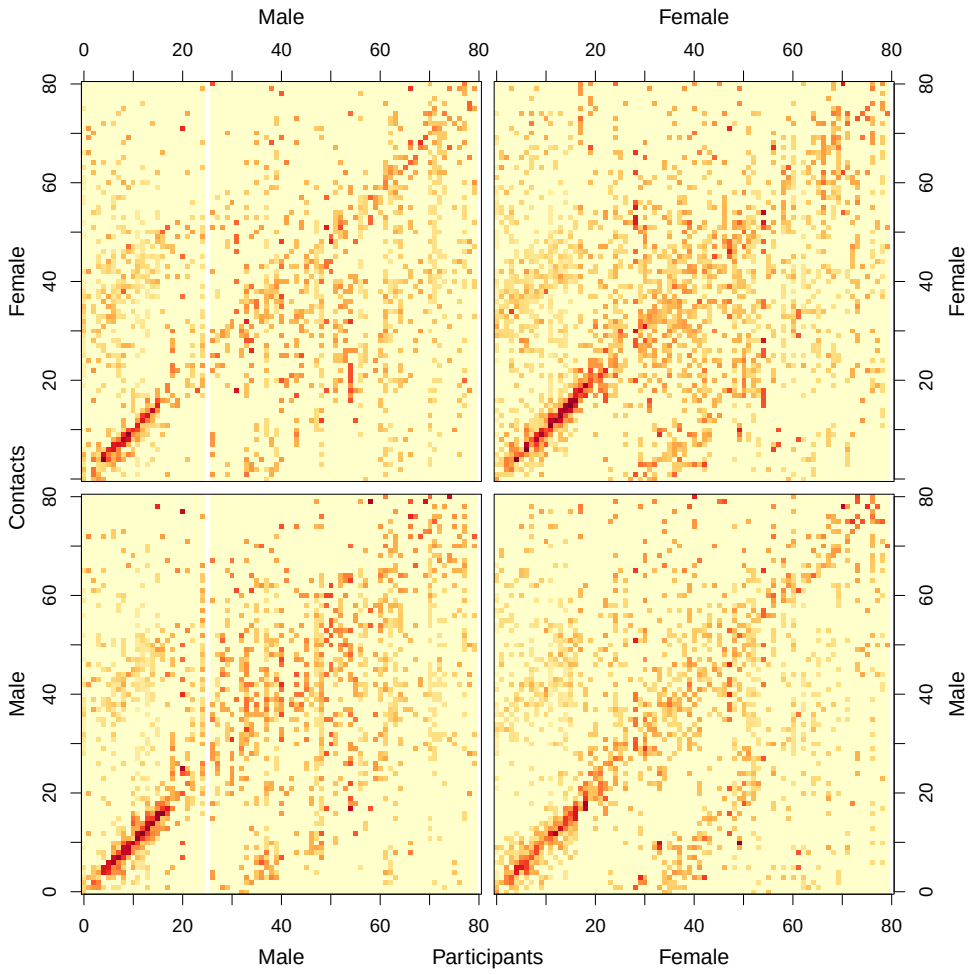

FIG 2. *Crude age and sex-specific contact rates. Similar to Figure 1, contacts are categorized as male-to-male (bottom left panel), female-to-male (bottom right panel), male-to-female (top left panel), and female-to-female (top right panel). The horizontal axis of each panel gives the age of the participants; the vertical axis gives the age of the contacts. The color scale indicates the relative values of the contact rates from low (yellow) to high (red).*

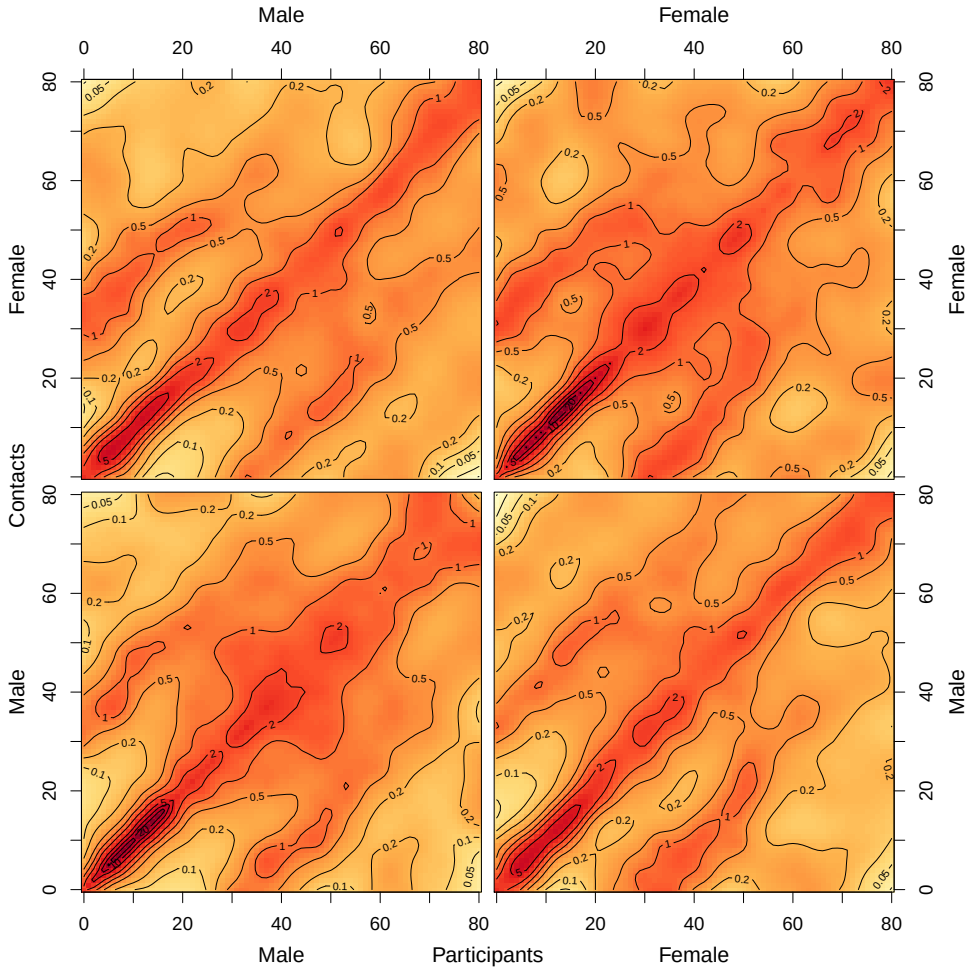

FIG 3. *Estimated smooth and symmetric age and sex-specific contact rates. The color scale indicates the relative values of the contact rates from low (yellow) to high (red); the contour lines are the absolute values of the contact rates per* 10^6.

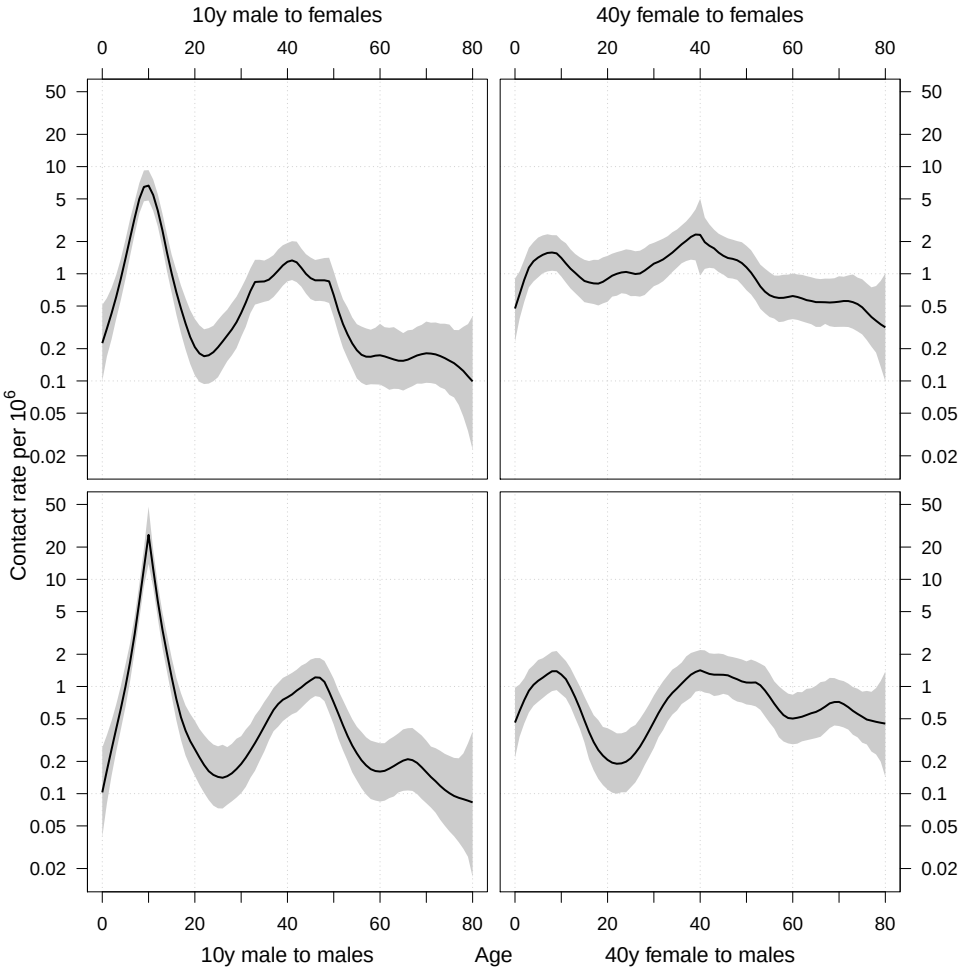

FIG 4. *Cross-sections of Figure 3 showing the contact rates per* 10^6 *for a male participant at age 10 having contact with males (bottom left panel) and females (top left panel), and a female participant at age 40 having contact with males (bottom right panel) and females (top right panel). Shaded areas indicate 95% credible intervals.*

assortative pattern with respect to age continues in adults, although contacts rates become lower and more diffuse with respect to age. For children and adolescents, contacts are also assortative with respect to sex: children and adolescents contact other children and adolescents of their own sex. For adults the assortative pattern with respect to sex disappears: adults contact other adults regardless of their sex. Disassortative patterns are also present. Children have more contact with adults who are approximately 30 years older than with adults. Women have higher contact rates with children than men.

Figure 4 shows cross-sections of Figure 3 for a male participant at age 10 and a female participant at age 40. The figure illustrates the uncertainties associated with the estimates. Contact rates for a 10 year old male with other males and females of different ages increase rapidly with age, having a sharp maximum with 10 year old males and females (classmates) and reaching another peak with 45 year old males (fathers) and 40 year old females (mothers). Contact rates for a 40 year old female with other males and females of different ages show a more gradual pattern. Note the discontinuity in the gradient at contact age 10 (10 year old males) and contact age 40 (40 year old females) as a result of the tailor made prior distribution.

3.3. *Model choice and validation.* Table 1 shows the WAIC and effective number of parameters for models with three different RW priors. The RW2 prior resulted in the lowest WAIC, and is therefore to be preferred. The RW1 prior implies the least regularization and highest effective number of parameters, the RW3 prior implies the most regularization, the smoothest surface and lowest effective number of parameters.

Figure 5 shows the PIT histograms for three models with different RW priors. In particular the histograms for the models with RW2 and RW3 prior are nearly uniform, indicating a good model fit, i.e. the observations are drawn from the predictive distribution.

The worst-case sensitivity estimates of the hyper parameters were 13% for the log-transformed precision parameter τ and 0.25% for the log-transformed dispersion parameter θ. These values indicate that the posterior distributions for the contact intensities are insensitive to the choice of the prior parameter values.

3.4. *Differences in projected age-specific risk of infection between the sexes.* The estimated contact rates could result in relevant differences between the sexes with respect to the age-specific risk of acquiring respiratory infections. To explore whether such differences occur, the posterior contact rates are converted into contact intensities by applying equation 2.3, e.g. for male-

TABLE 1

Comparison of three different priors for smoothing the contact rates, with the widely applicable information criterion $WAIC$ and effective number of parameters p_{eff}. The lowest value of the $WAIC$ indicates the most parsimonious model.

Prior	$WAIC$	p_{eff}
RW1	32505.5	1620.0
RW2	<u>32431.3</u>	479.3
RW3	32660.0	261.6

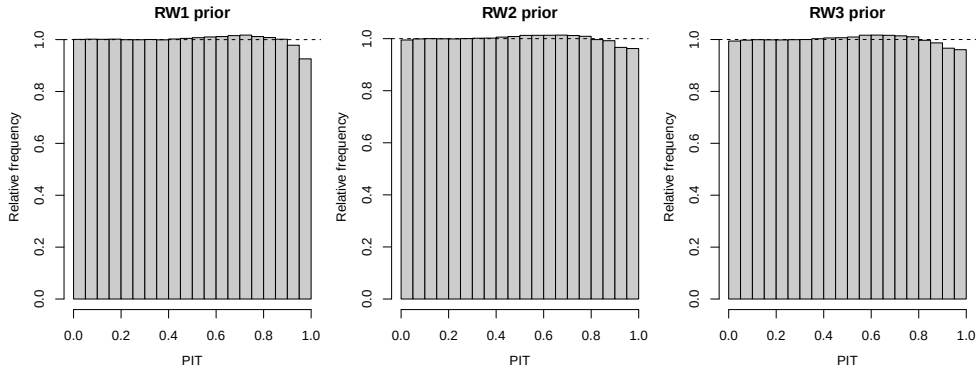

FIG 5. *PIT histograms for three models with different RW priors. The histograms show the probability integral transform of the observed total number of contacts relative to the model predictions.*

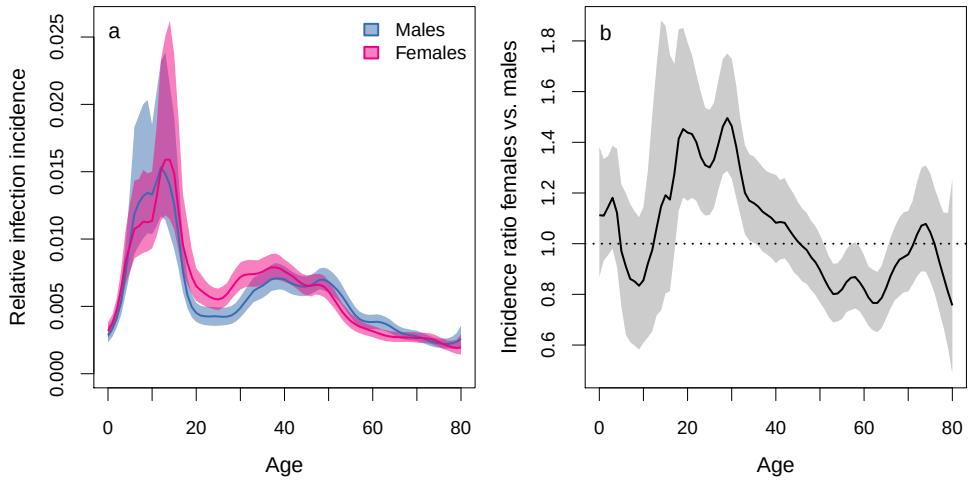

FIG 6. *Projected risk of infection, based on the estimated contact rates. a) Relative incidence of infection when a new emerging infection spreads in a completely susceptible population. b) Incidence ratio between females and males; above the dotted horizontal line females have higher incidence compared to males. Shaded areas indicate 95% credible intervals.*

to-female contacts $m_{ij}^{MF} = c_{ij}^{MF} w_j^F$, and assemble contact intensities into a block matrix. This matrix is called the next generation matrix. In our notation of this matrix, the index i refers to the characteristics of the infectors and the index j to the characteristics of the infected. The right dominant eigenvector of this matrix can be interpreted as the age and sex-specific risk of an infection that is transmitted via close contacts or respiratory droplets when a new emerging infection spreads in a completely susceptible population (Wallinga, Boven and Lipsitch, 2010). Here we use this right dominant eigenvector to quantify possible differences between the sexes to the age-specific risk of infection.

The resulting normalized age-specific risk of infection shows differences between the sexes (Figure 6a). The infection risk increases at a young age, from a relatively low risk for infants to a high risk for teenagers. There is a sharp decrease in infection risk for young adults, followed a modest rise in risk around the age of 40. For older ages the risk decreases again. The ratio between infection risk of females and males shows that between ages 18 - 38 women have a significantly higher infection risk than men, and between ages 50 - 65 women have a lower risk than men (Figure 6b).

We compared these contrasting age-specific infection risks for men and women to observations on influenza related mortality during the 1957 in-

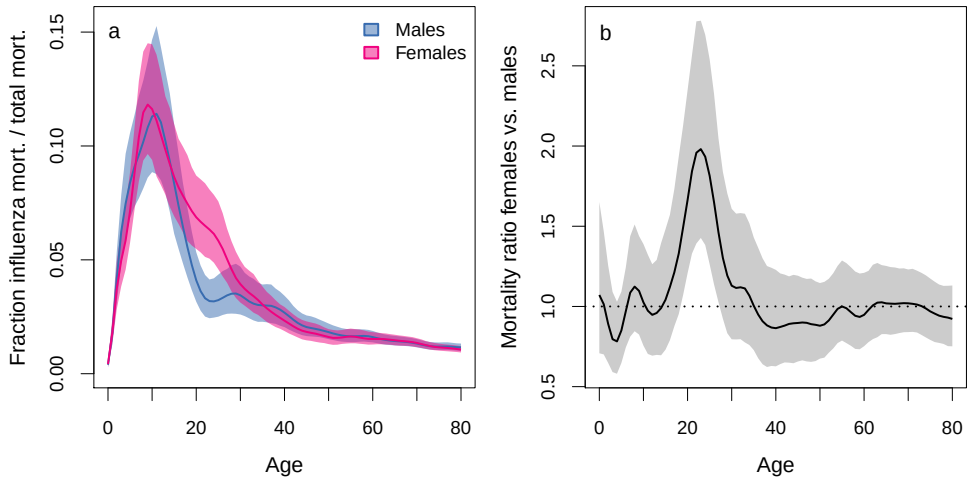

FIG 7. *Observed risk of death upon infection during the 1957 influenza pandemic in the Netherlands. a) Risk of death due to influenza relative to all-cause mortality by age and sex. b) Mortality ratio between females and males.*

fluenza pandemic in the Netherlands. We calculated the age and sex-specific fraction of deaths due to the 1957 influenza pandemic relative to total number of deaths in the Netherlands (Polak, 1959). Mortality rates were modeled by a Poisson generalized additive model using penalized splines for age. The mortality rates increase at a young age, from low rates for infants to high rates for teenagers. The broad pattern is similar to that of the projected infection risk: there is a decrease in infection risk for young adults, followed by a leveling off at the age of 40. For older ages the risk decreases further (Figure 7a). The ratio between mortality rates of females and males shows that between the ages 20 - 30 women have a significantly higher risk than men (Figure 7b). Figure 6 compared to Figure 7 shows that the differences in infection risk between the sexes, as expected from the observed age and sex-specific social contact patterns, could account to a large extend for the differences in mortality risk due to infection, as observed in mortality statistics.

4. Discussion and conclusions. We have estimated social contact rates at an unprecedented level of detail for men and women of all ages. The estimation of contact rates requires smoothing while at the same time enforcing consistency with reciprocal nature of contacts. We achieve smoothing by a Gaussian Markov Random Field approach where we impose a random walk prior in two dimensions directly on the logarithm of the age-specific

contact rates.

The proposed approach has important advantages over existing alternatives. First, the proposed approach makes it possible to increase the resolution up to 81-year age cohorts and stratification by sex. This is a significant improvement over existing methods for statistical analysis of social contact data that aggregate the entire age range in six crude age classes (Wallinga, Teunis and Kretzschmar, 2006) or 16 age classes (Mossong et al., 2008), and ignore sex. Second, we are able to explicitly specify the reciprocity constraints. This is a significant improvement over existing methods that do not guarantee reciprocity and risk inconsistent outcomes (Mossong et al., 2008). Third, the amount of smoothing is directly estimated from the amount of information in the data, and we can specify where the borders of smooth surfaces are located - at the boundaries and along the diagonal of the contact matrix - therefore preventing undesired artifacts such as smoothing across the diagonal. This is a significant improvement over existing methods that use a smoothing tensor spline with constraints (Goeyvaerts et al., 2010). Fourth, the proposed approach allows for fast and efficient estimation of contact rates. One model run takes a few minutes on a standard desktop computer.

The estimated social contact patterns require a critical checking with respect to model choice, validity and prior sensitivity. The patterns also require a critical checking with respect to plausibility to patterns of disease transmission. We will briefly look at each of these aspects in turn.

We have described the total number of contacts by a Negative Binomial distribution that is parameterized with a mean and a dispersion parameter. The mean is allowed to vary by age and sex. In this study we focus on the expected contact rates and therefore we treat the dispersion parameter θ and precision or smoothing parameter τ as nuisance parameters. If these parameters would be of direct interest they can be varied by age and sex, although this would be computationally challenging. We believe that it is a reasonable assumption that the dispersion and rate of change between ages for the contact process between men, between men and women, and between women are the same for all sexes.

Based on the WAIC, the RW2 prior is the preferred prior. In frequentist statistics it is common use to penalize the second derivative of the curve (Wood, 2006). Particularly in P-spline smoothing it is a common use to put a discrete second order penalty on the spline coefficients (Eilers and Marx, 1996). In a Bayesian context, this is exactly the same as putting a RW2 prior on the coefficients (Lang and Brezger, 2004). Our 2-dimensional RW2 prior is based on the work by Currie, Durban and Eilers (2004), where second

order differences in both the horizontal and vertical dimension are taken simultaneously. Our approach is a simple and intuitive way of constructing a smoother in two dimensions for triangular shaped grids. To develop some intuition for the proposed 2-dimensional RW2 prior: simulated realizations for a specified marginal prior variance and prior correlations for $n = 81$ and a fixed τ reveal that the prior variance is the largest near the boundaries and along the diagonal of the grid and that prior variance decreases towards the interior parts. The prior correlation between some given node and its surrounding nodes decreases with node distance. There are alternative approaches possible for a prior. A 2-dimensional RW2 prior could be constructed by applying the biharmonic operator, a discrete analogue of thin plate splines (Rue and Held, 2005). Adapting this prior to a triangular grid as required here is challenging.

The validity of the model was checked using PIT histograms (Czado, Gneiting and Held, 2009). The model accurately reflects the contact data as reported by the survey participants. We have examined prior sensitivity using the methodology described in Roos et al. (2015). The results indicated that the posterior distributions are insensitive to the choice of the prior hyper parameter values.

We checked the relevance of the estimated contact intensities as proxy measure for transmission of infection by comparing the project risk of infection to actual risk of mortality upon infection with influenza during the 1957 influenza pandemic. We chose the 1957 pandemic as it was the last large epidemic of a respiratory infection that hit an almost completely susceptible population in the Netherlands, and therefore provided the closest comparison with the model projections. As we cannot rule out any differences in age-dependent mortality rates between the sexes, and as the study population for the contact survey was not representative for the general population in 1957, this test does suggest that the estimated age- and sex-specific contact intensities sexes could provide a parsimonious explanation for much of the observed difference in mortality between the sexes. This contrasts with an implicit, but dominating assumption in the study of infectious diseases, that risk of respiratory infection is determined by age, and that sex-differences in risk of infection are negligible (Klein et al., 2010). Our findings suggest that, within each age group, much of the observed variation risk of infection within age groups might be due to sex-specific differences.

Additionally, we checked whether the findings can be repeated with similar data sets, for instance contact data collected in several different European countries in the POLYMOD study (Mossong et al., 2008), and with these data stratified even further by day of the week and by contact setting

(home, school, work, leisure, transport and other). We found a very similar overall contact patterns with respect to age and sex. Even though this additional analysis has some inconsistencies because the reciprocity of contacts, which holds true by definition over all settings, may become questionable for contacts within settings, its outcome strongly suggests that the estimated patterns are very plausible, and not specific to the particular contact data we have used here.

The observed differences in the contact pattern of adult men and women indicate that sex differences in infection risk only come into play when there is significant risk of infection for adults. This is the case for vaccine-preventable diseases where a high vaccination coverage increased the age at infection, such as measles, mumps, and rubella; for diseases that are poorly transmissible, such as human parvovirus B19 (Young and Brown, 2004), and diseases that cause repeated infections, such as influenza A, B (Klein et al., 2010) and cytomegalovirus (Pass et al., 2009). Furthermore, detailed social contact patterns help to improve the parameterization of mathematical models used to design and evaluate control strategies, allowing for models that are more realistic. In particular, the highly detailed age-specific contacts are relevant for evaluating vaccination schedules aimed at protecting infants against diseases such as pertussis or pneumococcal infections and sex-specific contacts are relevant for evaluating vaccination schedules aimed to protect against infections that complicate pregnancy or may lead to spontaneous abortion, such as rubella virus, human parvovirus B19, or cytomegalovirus.

Social contact intensities between men and women of all ages can be estimated using hierarchical Bayesian modeling approach with an underlying Gaussian Markov Random Field. A tailor-made construction of the precision matrix of the GMRF prior allows us to regularize the estimates by smoothing while at the same time enforces consistency with reciprocal nature of contacts. Estimation can be done efficiently using Integrated Nested Laplace Approximations. Application of this approach to social contact data from a large prospective survey revealed that social contact patterns, as a rule, are assortative: people, especially children and adolescents, mostly contact other people of their own age and sex. Relevant exceptions are that children have more contact with adults who are approximately 30 years older than with other adults, and women have more contact with children than men. These social contact patterns result in exposure patterns that significantly differ between sexes. For an emerging respiratory infection that spreads in a susceptible population, we would expect the risk of infection to vary for three broad age categories: for children, the risk of infection is similarly high for both sexes; for younger adults, women are at higher risk of infection; for

older adults, men are at higher risk of infection.

Acknowledgements. The authors thank Janneke Heijne for providing the contact data from the POLYMOD study in the Netherlands, and Hårvard Rue and Albert Wong for their useful suggestions concerning the construction of the structure matrix. Special thanks to all reviewers and editors who have provided critical and constructive comments that have helped to improve the manuscript considerably.

SUPPLEMENTARY MATERIAL

Supplementary material: Efficient estimation of age-specific social contact rates between men and women
(link will follow in final version).

References.

Borgdorff, M. W., Nagelkerke, N. J. D., Dye, C. and Nunn, P. (2000). Gender and tuberculosis: a comparison of prevalence surveys with notification data to explore sex differences in case detection. *The International Journal of Tuberculosis and Lung Disease* **4** 123–132.

Brown, A. C. and Moss, W. J. (2010). Sex, Pregnancy and Measles. In *Sex Hormones and Immunity to Infection* (S. L. Klein and C. Roberts, eds.) 281–302. Springer Berlin Heidelberg.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., Swerdlow, D., Sodha, S. V., Moll, M. E., Angulo, F. J., Palekar, R., Archer, W. R. and Finelli, L. (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences* **108** 2825–2830.

Conlan, A. J. K., Eames, K. T. D., Gage, J. A., Kirchbach, J. C. v., Ross, J. V., Saenz, R. A. and Gog, J. R. (2011). Measuring social networks in British primary schools through scientific engagement. *Proceedings of the Royal Society B: Biological Sciences* **278** 1467–1475.

Currie, I. D., Durban, M. and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* **4** 279–298.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics* **65** 1254–1261.

Danon, L., Read, J. M., House, T. A., Vernon, M. C. and Keeling, M. J. (2013). Social encounter networks: characterizing Great Britain. *Proceedings of the Royal Society of London B: Biological Sciences* **280** 20131037.

Davis, N. F., McGuire, B. B., Mahon, J. A., Smyth, A. E., OMalley, K. J. and Fitzpatrick, J. M. (2010). The increasing incidence of mumps orchitis: a comprehensive review. *BJU International* **105** 1060–1065.

Dawid, A. P. (1984). Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)* **147** 278–292.

Dodd, P. J., Looker, C., Plumb, I. D., Bond, V., Schaap, A., Shanaube, K., Muyoyeta, M., Vynnycky, E., Godfrey-Faussett, P., Corbett, E. L., Beyers, N.,

Ayles, H. and White, R. G. (2016). Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. *American Journal of Epidemiology* **183** 156–166.

Eames, K., Bansal, S., Frost, S. and Riley, S. (2015). Six challenges in measuring contact networks for use in modelling. *Epidemics* **10** 72–77.

Edmunds, W. J., O'Callaghan, C. J. and Nokes, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society B: Biological Sciences* **264** 949–957.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–121.

Falagas, M. E., Mourtzoukou, E. G. and Vardakas, K. Z. (2007). Sex differences in the incidence and severity of respiratory tract infections. *Respiratory Medicine* **101** 1845–1863.

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Damme, P. V. and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59** 255–277.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infectious Diseases* **9** 5.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data. Statistics for Biology and Health* **63**. Springer New York, New York, NY.

Holmes, C. B., Hausler, H. and Nunn, P. (1998). A review of sex differences in the epidemiology of tuberculosis. *The International Journal of Tuberculosis and Lung Disease* **2** 96–104.

Keeling, M. J. and White, P. J. (2011). Targeting vaccination against novel infections: risk, age and spatial structure for pandemic influenza in Great Britain. *Journal of the Royal Society Interface* **8** 661–670.

Klein, S. L., Passaretti, C., Anker, M., Olukoya, P. and Pekosz, A. (2010). The impact of sex, gender and pregnancy on 2009 H1N1 disease. *Biology of Sex Differences* **1** 1–12.

Kucharski, A. J., Kwok, K. O., Wei, V. W. I., Cowling, B. J., Read, J. M., Lessler, J., Cummings, D. A. and Riley, S. (2014). The Contribution of Social Behaviour to the Transmission of Influenza A in a Human Population. *PLoS Pathog* **10** e1004206.

Kwok, K. O., Cowling, B. J., Wei, V. W. I., Wu, K. M., Read, J. M., Lessler, J., Cummings, D. A., Peiris, J. S. M. and Riley, S. (2014). Social contacts and the locations in which they occur as risk factors for influenza infection. *Proceedings of the Royal Society of London B: Biological Sciences* **281** 20140709.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics* **13** 183–212.

Medlock, J. and Galvani, A. P. (2009). Optimizing Influenza Vaccine Distribution. *Science* **325** 1705–1708.

Mehl, M. R., Vazire, S., Ramrez-Esparza, N., Slatcher, R. B. and Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science* **317** 82–82.

Miller, E., Hoschler, K., Hardelid, P., Stanford, E., Andrews, N. and Zambon, M. (2010). Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *The Lancet* **375** 1100–1108.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, W. J. (2008). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Med* **5** e74.

Neyrolles, O. and Quintana-Murci, L. (2009). Sexual Inequality in Tuberculosis. *PLoS Med* **6** e1000199.

Pass, R. F., Zhang, C., Evans, A., Simpson, T., Andrews, W., Huang, M.-L., Corey, L., Hill, J., Davis, E., Flanigan, C. and Cloud, G. (2009). Vaccine Prevention of Maternal Cytomegalovirus Infection. *New England Journal of Medicine* **360** 1191–1199.

Polak, M. F. (1959). Influenzasterfte in de herfst van 1957. *Nederlands Tijdschrift voor Geneeskunde* **103** 1098–109.

Read, J. M., Edmunds, W. J., Riley, S., Lessler, J. and Cummings, D. a. T. (2012). Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiology & Infection* **140** 2117–2130.

Rohani, P., Zhong, X. and King, A. A. (2010). Contact Network Structure Explains the Changing Epidemiology of Pertussis. *Science* **330** 982–985.

Roos, M., Martins, T. G., Held, L. and Rue, H. (2015). Sensitivity Analysis for Bayesian Hierarchical Models. *Bayesian Analysis* **10** 321–349.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, 1 edition ed. Chapman and Hall/CRC.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 319–392.

Salath, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W. and Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* **107** 22020–22025.

StatLine (2015). Population on 1 January by age and sex. http://statline.cbs.nl/Statweb/search/?Q=population&LA=EN.

Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling*. Oxford University Press.

Wallinga, J., Boven, M. v. and Lipsitch, M. (2010). Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences* **107** 923–928.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006). Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents. *American Journal of Epidemiology* **164** 936–944.

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *J. Mach. Learn. Res.* **14** 867–897.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC.

Young, N. S. and Brown, K. E. (2004). Parvovirus B19. *New England Journal of Medicine* **350** 586–597.

Department of Statistics, Informatics
and Mathematical Modeling
National Institute for Public Health
and the Environment (RIVM)
PO Box 1, 3620BA
Bilthoven, the Netherlands
E-mail: jan.van.de.kassteele@rivm.nl
jan.van.eijkeren@rivm.nl

Epidemiology and Surveillance Unit
National Institute for Public Health
and the Environment (RIVM)
PO Box 1, 3620BA
Bilthoven, the Netherlands
E-mail: jacco.wallinga@rivm.nl