

Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases

Joël Mossong^{1,2*}, Niel Hens³, Mark Jit⁴, Philippe Beutels⁵, Kari Auranen⁶, Rafael Mikolajczyk⁷, Marco Massari⁸, Stefania Salmaso⁸, Gianpaolo Scalia Tomba⁹, Jacco Wallinga¹⁰, Janneke Heijne¹⁰, Malgorzata Sadkowska-Todys¹¹, Magdalena Rosinska¹¹, W. John Edmunds⁴

1 Microbiology Unit, Laboratoire National de Santé, Luxembourg, Luxembourg, **2** Centre de Recherche Public Santé, Luxembourg, Luxembourg, **3** Center for Statistics, Hasselt University, Diepenbeek, Belgium, **4** Modelling and Economics Unit, Health Protection Agency Centre for Infections, London, United Kingdom, **5** Unit Health Economic and Modeling Infectious Diseases, Center for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium, **6** Department of Vaccines, National Public Health Institute KTL, Helsinki, Finland, **7** School of Public Health, University of Bielefeld, Bielefeld, Germany, **8** Istituto Superiore di Sanità, Rome, Italy, **9** Department of Mathematics, University of Rome Tor Vergata, Rome, Italy, **10** Centre for Infectious Disease Control Netherlands, National Institute for Public Health and the Environment, Bilthoven, The Netherlands, **11** National Institute of Hygiene, Warsaw, Poland

Funding: This study formed part of POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract number: SSP22-CT-2004-502084. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Academic Editor: Steven Riley, Hong Kong University, Hong Kong

Citation: Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5(3) e74. doi:10.1371/journal.pmed.0050074

Received: August 8, 2007

Accepted: February 15, 2008

Published: March 25, 2008

Copyright: © 2008 Mossong et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: BE, Belgium; DE, Germany; FI, Finland; GB, Great Britain; IT, Italy; LU, Luxembourg; NL, The Netherlands; PL, Poland

* To whom correspondence should be addressed. E-mail: joel.mossong@lns.etat.lu

ABSTRACT

Background

Mathematical modelling of infectious diseases transmitted by the respiratory or close-contact route (e.g., pandemic influenza) is increasingly being used to determine the impact of possible interventions. Although mixing patterns are known to be crucial determinants for model outcome, researchers often rely on a priori contact assumptions with little or no empirical basis. We conducted a population-based prospective survey of mixing patterns in eight European countries using a common paper-diary methodology.

Methods and Findings

7,290 participants recorded characteristics of 97,904 contacts with different individuals during one day, including age, sex, location, duration, frequency, and occurrence of physical contact. We found that mixing patterns and contact characteristics were remarkably similar across different European countries. Contact patterns were highly assortative with age: schoolchildren and young adults in particular tended to mix with people of the same age. Contacts lasting at least one hour or occurring on a daily basis mostly involved physical contact, while short duration and infrequent contacts tended to be nonphysical. Contacts at home, school, or leisure were more likely to be physical than contacts at the workplace or while travelling. Preliminary modelling indicates that 5- to 19-year-olds are expected to suffer the highest incidence during the initial epidemic phase of an emerging infection transmitted through social contacts measured here when the population is completely susceptible.

Conclusions

To our knowledge, our study provides the first large-scale quantitative approach to contact patterns relevant for infections transmitted by the respiratory or close-contact route, and the results should lead to improved parameterisation of mathematical models used to design control strategies.

The Editors' Summary of this article follows the references.



Introduction

Preparing for outbreaks of directly transmitted pathogens such as pandemic influenza [1–3] and SARS [4–9], and controlling endemic diseases such as tuberculosis and meningococcal diseases, are major public health priorities. Both can be achieved by nonpharmaceutical interventions such as school closure, travel restrictions, and contact tracing, or by health-care interventions such as vaccination and use of antiviral or antibiotic agents [2,10–13]. Mathematical models of infectious disease transmission within and between population groups can help to predict the impact of such interventions and inform planning and decision making. Contact rates between individuals are often critical determinants of model outcomes [14]. However, few empirical studies have been conducted to determine the patterns of contact between and within groups and in different social settings.

In comparison to HIV and sexually transmitted diseases [15–17] and drug/needle sharing networks [18], where a number of large-scale empirical studies have been conducted on contact patterns, relatively little effort has been devoted to infections spread by respiratory droplets or close contact. Instead, the contact structure for these infections has been assumed to follow a predetermined pattern governed by a small number of parameters that are then estimated using seroepidemiological data [19,20]. A small number of studies have attempted to directly quantify such contact patterns, but they were conducted in small or nonrepresentative populations [14,21–25]. Hence, it is unclear to what extent the results can be generalized to an overall population and across different geographical areas. To address this lack of empirical knowledge, we present here results from, to our knowledge the first, large-scale, prospectively collected, population-based survey of epidemiologically relevant social contact patterns. The study was conducted in eight different European countries using a common paper diary approach and covering all age groups. We use these data to assess how an emerging infection could spread in a wholly susceptible population if it were transmitted by the social contacts measured here.

Methods

Survey Methodology

Information on social contacts was obtained using cross-sectional surveys conducted by different commercial companies or public health institutes in Belgium (BE), Germany (DE), Finland (FI), Great Britain (GB), Italy (IT), Luxembourg (LU), The Netherlands (NL), and Poland (PL). The recruitment and data collection were organised at the country level according to a common agreed quota sampling methodology and diary design. The surveys were conducted between May 2005 and September 2006 with the oral informed consent of participants and approval of national institutional review boards following a small pilot study to test feasibility of the diary design and recruitment [26].

Survey participants were recruited in such a way as to be broadly representative of the whole population in terms of geographical spread, age, and sex. In BE, IT, and LU, survey participants were recruited by random digit dialling using land line telephones; in GB, DE, and PL survey participants were recruited through a face-to-face interview; survey

participants in NL and FI were recruited via population registers and linked to a larger national sero-epidemiology survey in NL. Children and adolescents were deliberately oversampled, because of their important role in the spread of infectious agents. For more details on the survey methodology in the various countries, see Table S1.

Briefly, only one person in each household was asked to participate in the study. Paper diaries were either sent by mail or given face to face to participants. Participants were coached by telephone or in person on how to fill in the diary.

Diaries recorded basic sociodemographic information about the participant, including employment status, level of completed education, household composition, age, and sex. Participants were assigned a random day of the week to record every person they had contact with between 5 A.M. and 5 A.M. the following morning. Participants were instructed to record contacted individuals only once in the diary. A contact was defined as either skin-to-skin contact such as a kiss or handshake (a physical contact), or a two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact (a nonphysical contact). Participants were also asked to provide information about the age and sex of each contact person. If the age of a contact person was not known precisely, participants were asked to provide an estimate of the age range (the midpoint was used for data analysis). For each contact, participants were asked to record location (home, work, school, leisure, transport, or other), the total duration of time spent together (less than 5 min, 5–15 min, 15 min to 1 h, 1–4 h, or 4 h or more) as well as the frequency of usual contacts with this individual (daily or almost daily, about once or twice a week, about once or twice a month, less than once a month, or for the first time).

Diaries were translated into local languages (see Text S1 for the diary used in GB) and are available on request in the following languages: Dutch, English, French, Finnish, German, Italian, Polish, Portuguese, and Swedish. Diaries for young children were filled in by a parent or guardian on their behalf. Older children who obtained parental consent were given diaries with simplified language to fill in on their own (see Table S1 for more details).

Data Analysis

Main effects of covariates (age, sex, household size, and country) on numbers of contacts were assessed using multiple censored negative binomial regression [27]. The data were right censored at 29 contacts for all countries because of a limited number of possible diary entries in some countries. Additionally, a sensitivity analysis was performed to assess the effects of different handling of professional contacts between the countries.

The log-likelihood function ll for the censored negative binomial was

$$ll = \sum_{i=1}^n w_i (\delta_i \log(P(Y = y_i | \mathbf{X}_i)) + (1 - \delta_i) \log(1 - \sum_{j=0}^{28} P(Y = j | \mathbf{X}_i))),$$

where w_i is the weight of observation i , $\delta_i = \begin{cases} 1 & \text{if } y_i < 29 \\ 0 & \text{if } y_i \geq 29 \end{cases}$ is an indicator variable for censoring, y_i is the number of observed

contacts, \mathbf{X}_i is the vector of explanatory variables, and P is the probability function of the negative binomial distribution:

$$P(Y = y_i | \mathbf{X}_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu} \right)^{1/\alpha} \left(\frac{\alpha\mu}{1 + \alpha\mu} \right)^{y_i},$$

where $\mu = \exp(\mathbf{X}_i\beta)$; β is the vector of coefficients and α is the overdispersion parameter.

Sampling weights—the inverse of the probability that an observation is included because of the sampling design—were calculated for each country separately, based on official age and household size data of the year 2000 census round data published by Eurostat (<http://epp.eurostat.ec.europa.eu/>) (see Table S2) and used to correctly estimate population-related quantities. Overall statistics should be considered indicative of general trends and levels, but specific statistical representativity for the whole of Europe is not claimed, since participating countries, although geographically and socially diverse, are not a representative or random selection at the European level.

Association Rule Analysis

Mining association rules is a tool for discovering patterns between variables in large databases [28]. Let X, Y denote disjoint nonempty items in the contact survey, such as daily frequency, duration of more than 4 h, and physical contact. Association rules are rules of the form $X \rightarrow Y$ that measure how likely the event Y is, given X . In this context X is called antecedent while Y is called consequent. Rules are typically extended to include more items in the antecedent but are restricted to include only one item in the consequent. The length of the rule is defined as the total number of items in both antecedent and consequent.

Selecting interesting rules from the set of all possible rules is based on various measures of significance and interest. The best-known are support, confidence, and lift. The support of an association rule $X \rightarrow Y$ is defined as the relative frequency of $X \cap Y$. Finding rules with high support can be seen as a simplification of the learning problem called “mode finding” or “bump hunting.” The confidence of a rule is the conditional probability $P(Y|X)$ indicating what percentage of times the rule holds and thus measuring the association between $\{X, X^c\}$ and $\{Y, Y^c\}$. Using both constraints, the set of rules can further be filtered by the lift, which is defined as the ratio of the relative frequency of $X \cap Y$ and the product of relative frequencies of X and Y . The lift can be interpreted as the ratio of the rule’s observed support to the support expected under independence. Greater lift values indicate stronger associations. Additionally, a Chi-square test for the rule-corresponding two-by-two table consisting of cells $X \cap Y, X^c \cap Y, X \cap Y^c, X^c \cap Y^c$, where c refers to the complementing set of items, can be used to test statistical significance of the association. Whenever the Chi-square distribution seemed inappropriate due to small sample size, a Fisher exact test was used. For a more extensive overview of applying association rules on contact data see [29].

Contact Surface Smoothing

Contact surface smoothing was performed by applying a negative binomial model on the aggregated number of contacts (both physical and nonphysical) over 5 y age bands for both responders and contacts using a tensor product spline as a smooth interaction term [30,31].

Epidemiological Modelling: Simulating the Initial Phase of an Epidemic

We explore the age-specific incidence of infection during the initial phase of an epidemic of an emerging infectious disease agent that spreads in a completely susceptible population. We focus on the generic features of epidemic spread along the transmission route that is specified by physical and nonphysical contacts as defined here. We partition the population into 5 y age bands, and we group all individuals aged 70 y and older together. This process results in 15 age classes. We denote the number of at-risk contacts of an individual in age class j with individuals in age class i by k_{ij} . We take k_{ij} as proportional to the observed number of contacts (both physical and nonphysical) that a respondent in age band j makes with other individuals in age band i . The matrix with elements k_{ij} is known in infectious disease epidemiology as the next generation matrix \mathbf{K} [32]. The next generation matrix can be used to calculate the distribution of numbers of new cases in each generation of infection from any arbitrary initial number of introduced infections. For example, when infection is introduced by one single 65-y-old infected individual into a completely susceptible population, we can denote the number of initial cases in generation 0 by the vector $\mathbf{x}_0 = (0,0,0,0,0,0,0,0,0,0,0,0,1,0)^T$. The expected numbers of new cases in the i th generation are denoted by the vector \mathbf{x}_i , and this vector is calculated by applying the next generation matrix \mathbf{K} i times to the initial numbers of individuals \mathbf{x}_0 , that is, $\mathbf{x}_i = \mathbf{K}^i \mathbf{x}_0$. For large i , the vector \mathbf{x}_i will be proportional to the leading eigenvector of \mathbf{K} . We find that, in practice, the distribution of new cases is stable after five generations; that is, the distribution no longer depends on the precise age of the initial case. The incidence of new infections per age band is obtained by dividing the expected number of new cases per age class by the number of individuals in each age class. To facilitate comparison among countries, we normalized the distribution of incidence over age classes such that for each country the age-specific incidences sum to one.

Results

Description of Sample

A total of 7,290 diaries covering all contacts made by respondents during a full day were collected in eight countries ranging from 267 in NL to 1,328 in DE (see Table 1). 37.6% of participants in our survey were under 20 y of age, 12.4% of participants were over 60 y of age, and the medians were 28 y in BE (the lowest) to 33 y in DE (the highest). Returns of diaries by female participants showed a slight excess in all countries (ranging from 50.8% in FI to 55.7% in DE). In all countries except DE, single-person households were underrepresented in our sample (Table S2). This can be partially explained by the fact that children and adolescents were deliberately oversampled, and they are more likely to live in larger households.

Overall, 35.3% of the participants were in full-time education, 32.6% employed, 11% retired, 6.1% homemakers, 3.6% unemployed or seeking employment, whereas 8.6% recorded "other," and 2.8% failed to record their occupation. The proportion employed or in full-time education was fairly consistent across the eight countries; the other categories differed somewhat between countries.

Table 1. Number of Recorded Contacts per Participant per Day by Different Characteristics and Relative Number of Contacts from the Weighted Multiple Censored Negative Binomial Regression Model

Category	Covariate	Number of Participants	Mean (Standard Deviation) of Number of Reported Contacts	Relative Number of Reported Contacts (95% Confidence Interval) ^a
Age of participant, y	0–4	660	10.21 (7.65)	1.00
	5–9	661	14.81 (10.09)	1.42 (1.28–1.55)
	10–14	713	18.22 (12.27)	1.73 (1.57–1.90)
	15–19	685	17.58 (12.03)	1.68 (1.52–1.84)
	20–29	879	13.57 (10.60)	1.45 (1.33–1.57)
	30–39	815	14.14 (10.15)	1.45 (1.34–1.57)
	40–49	908	13.83 (10.86)	1.38 (1.27–1.50)
	50–59	906	12.30 (10.23)	1.31 (1.20–1.42)
	60–69	728	9.21 (7.96)	1.06 (0.96–1.16)
	70+	270	6.89 (5.83)	0.81 (0.73–0.88)
	Missing value	65	9.63 (9.05)	0.91 (0.66–1.17)
Sex of participant	Female	3,808	13.39 (10.57)	1.00
	Male	3,429	13.51 (10.67)	0.99 (0.96–1.02)
	Missing value	53	10.92 (8.60)	1.57 (1.09–2.05)
Household size	1	749	8.87 (8.27)	1.00
	2	1,645	10.65 (9.14)	1.17 (1.11–1.24)
	3	1,683	12.87 (10.26)	1.20 (1.13–1.27)
	4	2,041	15.84 (11.17)	1.36 (1.28–1.44)
	5	814	16.47 (11.21)	1.46 (1.35–1.56)
	6+	358	17.69 (10.98)	1.56 (1.43–1.70)
Day of the week	Sunday	862	10.10 (8.76)	1.00
	Monday	1,032	13.32 (10.31)	1.33 (1.24–1.41)
	Tuesday	1,116	14.17 (10.83)	1.39 (1.31–1.48)
	Wednesday	1,017	14.58 (11.14)	1.38 (1.29–1.47)
	Thursday	1,069	14.70 (11.23)	1.41 (1.32–1.50)
	Friday	1,122	14.72 (11.25)	1.43 (1.34–1.52)
	Saturday	936	11.63 (9.11)	1.20 (1.12–1.28)
	Missing value	136	12.48 (10.66)	1.24 (1.08–1.40)
Country ^b	BE	750	11.84 (9.85)	1.00
	DE	1,341	7.95 (6.26)	0.70 (0.65–0.74)
	FI	1,006	11.06 (7.89)	0.94 (0.88–1.00)
	GB	1,012	11.74 (7.67)	0.99 (0.92–1.05)
	IT	849	19.77 (12.27)	1.66 (1.55–1.78)
	LU	1,051	17.46 (12.81)	1.42 (1.33–1.51)
	NL	269	13.85 (10.54)	1.34 (1.20–1.47)
	PL	1,012	16.31 (11.45)	1.37 (1.28–1.47)

^aDispersion parameter $\alpha = 0.36$ (95% CI 0.34–0.37); $\alpha = 0$ would correspond to no overdispersion, i.e., a censored Poisson distribution.

^bDirect comparisons between countries are difficult because of different approaches to recording frequent professional contacts. In BE, DE, FI, and NL, participants were instructed not to record professional contacts in the diary if they had more than 20 (BE) or 10 (DE, FI, NL) of them per day.
doi:10.1371/journal.pmed.0050074.t001

Number of Contacts

A total of 97,904 contacts with different persons were recorded (mean = 13.4 per participant per day) in the diaries. On average, German participants reported the fewest daily number of contacts (mean = 7.95, standard deviation [SD] = 6.26) and Italians the highest number (mean = 19.77, SD = 12.27). The contact distributions in all countries are slightly skewed, the skewness statistics ranging from 0.62 in IT to 2.96 in DE (Figure S1). Analysis of the total number of reported contacts with a multiple regression model shows a consistent pattern of contact frequency by age, with a gradual rise in the number of contacts in children, a peak among 10- to 19-y-olds, followed by a fall to a lower plateau in adults until the age of 50 and a sharp decrease after that age (Table 1). Living in a larger household size was associated with higher number of reported contacts. Weekdays were associated with 30%–40% more contacts than Sundays. The influence of the country in which the survey was performed was also apparent (Table 1), even when adjusting for the main different

recording formats we used in different countries (diary sizes and estimates of professional contacts) (see Table S3). The overdispersion parameter in the model was significantly different from zero, indicating the necessity to use a negative binomial model as opposed to a Poisson model.

Frequency, Intensity, and Location of Contacts

The intensity of contacts was measured in a number of ways, all of which were found to be highly correlated with each other (see Figure 1 for pooled data from all countries, Figure S2 for country-specific data). Contacts of long duration or of daily frequency were much more likely to involve physical contact. Approximately 70% of contacts made on a daily basis last in excess of an hour, whereas approximately 75% of contacts made with individuals who have never been contacted before lasted for less than 15 min. Approximately 75% of contacts at home and 50% of school and leisure contacts were physical, whereas only a third of contacts recorded in other settings were physical; approximately two-thirds of the persons contacted in multiple

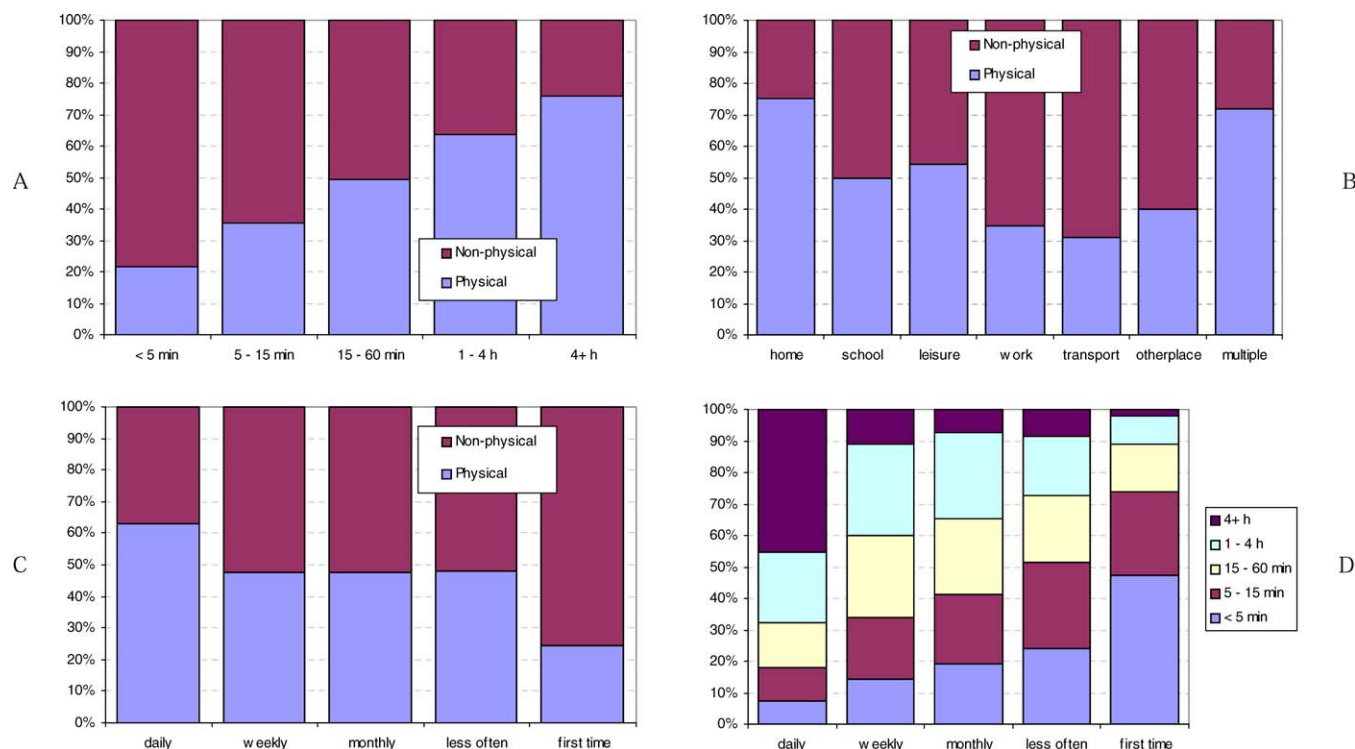


Figure 1. The Mean Proportion of Contacts That Involved Physical Contact, by Duration, Frequency, and Location of Contact in All Countries. Graphs show data by (A) duration, (B) location, and (C) frequency of contact; the correlation between duration and frequency of contact is shown in (D). All correlations are highly significant ($p < 0.001$, χ^2 -test). The figures are based on pooled contact data from all eight countries and weighted according to sampling weights as explained in the Methods (based on household size and age). doi:10.1371/journal.pmed.0050074.g001

settings involved a contact at home, and so a high proportion were physical.

Mining the contact data for frequency, duration, and type of contact based on association rules of maximum length 3 using thresholds of 0.5% (about 500 contacts) on the occurrence, positive dependence, and a 5% significance level on the Chi-square test of dependence resulted in a total of 99 rules of which 46 were of length 2 (see Table S4). 75% of the contacts lasting 4 h or more involved physical contact and occurred on a daily basis (83%), while 83% of the first-time contacts lasting less than 5 min were nonphysical. First time and occasional contacts mostly lasted less than 15 min (lift values 3.3 and 1.8, respectively) and, when nonphysical, this association was intensified (lift values 3.6 and 2.6, respectively). Whether contacts were physical or not did not influence the association between contacts lasting at least four hours and occurring on a daily basis nor did it influence the association between contacts lasting from five minutes up to one hour and occurring on a weekly or monthly basis. Physical contacts and contacts lasting 1–4 h were the only characteristics that were symmetric—that is, they had the same level of confidence in both directions (66% and 64%, respectively). Overall, 67% of all physical contacts lasted for at least 1, while 56% of all physical contacts occurred on a daily basis. All previously reported rules had high lift-values and were significant at the 1% significance level. Due to the high degree of correlation between physical contact and other measures of intimate contact, in the remainder of the paper we use physical contacts as a proxy measure for high-intensity contacts.

Of all pooled reported contacts, 23%, 21%, 14%, 3%, and 16% are made at home, at work, at school, while travelling, and during leisure activities, respectively (Figure 2A). More than half of all reported contacts occur at home, at work, or at school. It is interesting to note, however, that on a population level the overall number of reported contacts made during leisure activities is very close to the number of reported contacts made at school. A higher proportion of physical contacts are made at home, and leisure settings are the second most frequently reported location for such high intensity contacts (Figure 2B).

Age-Related Mixing Patterns

Figure 3 shows the average number of contacts reported per participant with individuals of different age groups for each of the eight countries for all reported (Figure 3A) and physical contacts (Figure 3B) only (full contact matrix data can be found in Table S5). Apart from the remarkable similarity of the general contact pattern structure in the different countries, three main features are apparent from the data. First, the dominant feature is the strong diagonal element: individuals in all age groups tend to mix assortatively (i.e., preferentially with others of similar age). This pattern is most pronounced in those aged 5–24 years, and least pronounced in those aged 55–69.

Second, two parallel secondary diagonals starting at roughly 30–35 years for both contacts and participants are offset from the central diagonal. This pattern represents children mixing with adults in the 30–39 age range (mainly at home, see Figure S3) and vice versa. Older children mix with

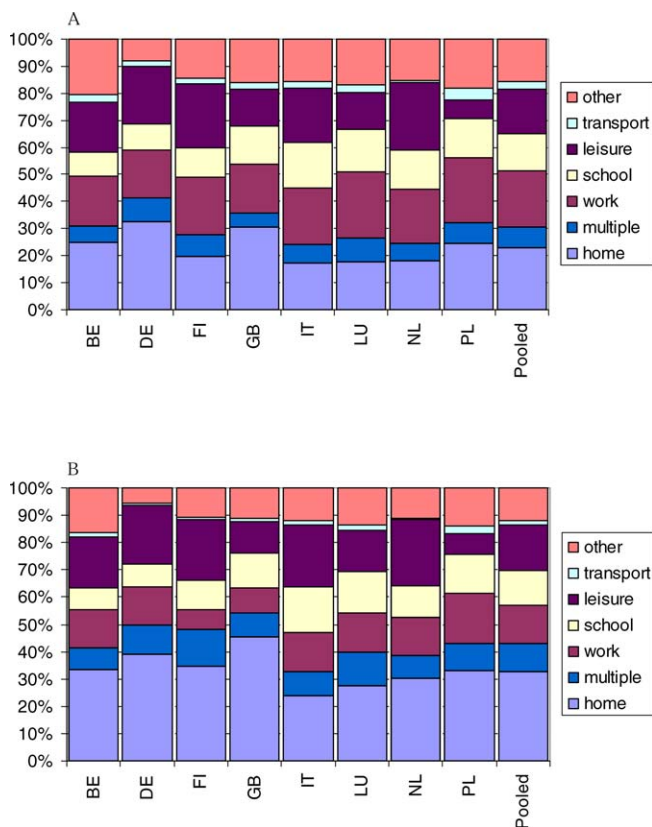


Figure 2. The Distribution by Location and by Country of (A) All Reported Contacts and (B) Physical Contacts Only. Sampling weights were used for each country. “Other” refers to contacts made at locations other than home, work, school, travel, or leisure. “Multiple” refers to the fact that the person was contacted during the day in multiple locations, not just a single location. doi:10.1371/journal.pmed.0050074.g002

middle-aged adults. Note, though, that the contact rates of the secondary diagonals at 30–35 years offset are an order of magnitude lower than the main assortative diagonal. Mixing between middle-aged adults and the elderly (above 60 y) was also apparent (see Figure S3).

The third feature is more apparent in the data for all reported contacts (Figure 3A) than for physical contacts only: a wider contact “plateau” of adults with other adults primarily due to low-intensity contacts, with many of these contacts occurring at work (see also Figure S4).

Simulated Initial Phase of an Epidemic

According to our mathematical model, the age distribution of cases during the initial phase of an epidemic of a new, emerging infection that spreads according to the reported social contacts in a completely susceptible population reveals a typical pattern that is similar across countries (see Figure 4). The highest incidence occurs among schoolchildren (ranging from 5- to 9-y-olds in NL to 5- to 19-y-olds in IT), and a less pronounced second peak in incidence occurs among adults (ranging from 30- to 34-y-olds in PL to 40- to 44-y-olds in FI). The high incidence among school-aged children results from their high number of contacts relative to other groups, and their tendency to make contacts within their own age group. The tendency to contact others within the same age group could potentially lead to a slow dispersion of infection across

age groups. However, the contacts outside age groups are often with others about 30–35 years older or younger, and this tendency results in fairly rapid dispersion of infection across all age groups. Therefore, the observed contact patterns reveal that schoolchildren drive the epidemic in all age groups during the initial phase of spread for infections transmitted by droplets and through close contacts.

Discussion

Mathematical models are increasingly used to evaluate and inform infectious disease prevention and control policy. At their heart all models must make assumptions about how individuals contact each other and transmit the infectious agent. Until now, modellers have relied on proxy measures of contacts and calibration to epidemiological data. For instance, household size, class size, transport statistics, and workplace size distribution have been used in recent models to define the contact structure [2,3,33,34]. Our study complements those relying on proxy measures by using direct estimates of the number, age, intimacy levels, and distribution of actual contacts within various settings. The analysis of population-based contact patterns can help inform the structure and parameterisation of mathematical models of close-contact infectious diseases.

One of the most important findings of our study is that the age and intensity patterns of contact are remarkably similar across different European countries even though the average number of contacts recorded differed. This similarity implies that the results may well be applicable to other European countries, and that the initial phase of spread of newly emerging infections in susceptible populations, such as SARS was in 2003, is likely to be very similar across Europe and in countries with similar social structures.

Another major insight gained from our study comes from the observation that the contacts made by children and adolescents are more assortative than contacts made by other age groups. That is, most of the individuals contacted by children and teenagers are of very similar age, and these contacts tend to be of long duration. This pattern is likely to be the main reason why children and teenagers are and have been an important conduit for the initial spread of close-contact infections in general and for influenza in particular [11,14] and our preliminary modelling work confirms this.

Our study allows us to assess and quantify the risk of transmission in different settings. We took a number of different measures of “closeness of contact,” including duration and frequency of contact and whether skin-to-skin contact occurred. These measures correlated highly with each other, such that the longer-duration contacts tended to be frequent and to involve physical contact (and vice versa). More-intimate contacts are likely to carry a greater risk of transmission. Furthermore, these types of contact tend to occur in distinct social settings: the most-intimate contacts occur at home or in leisure settings, whereas the least-intimate tend to occur while travelling. Thus, the risk of infection in these settings can be inferred to vary. This variation has important implications for contact tracing during outbreaks of a new infection. Our results suggest that if efforts concentrate on locating contacts in the home, school, workplace, and leisure settings, on average more than 80% of all contacts would be found.

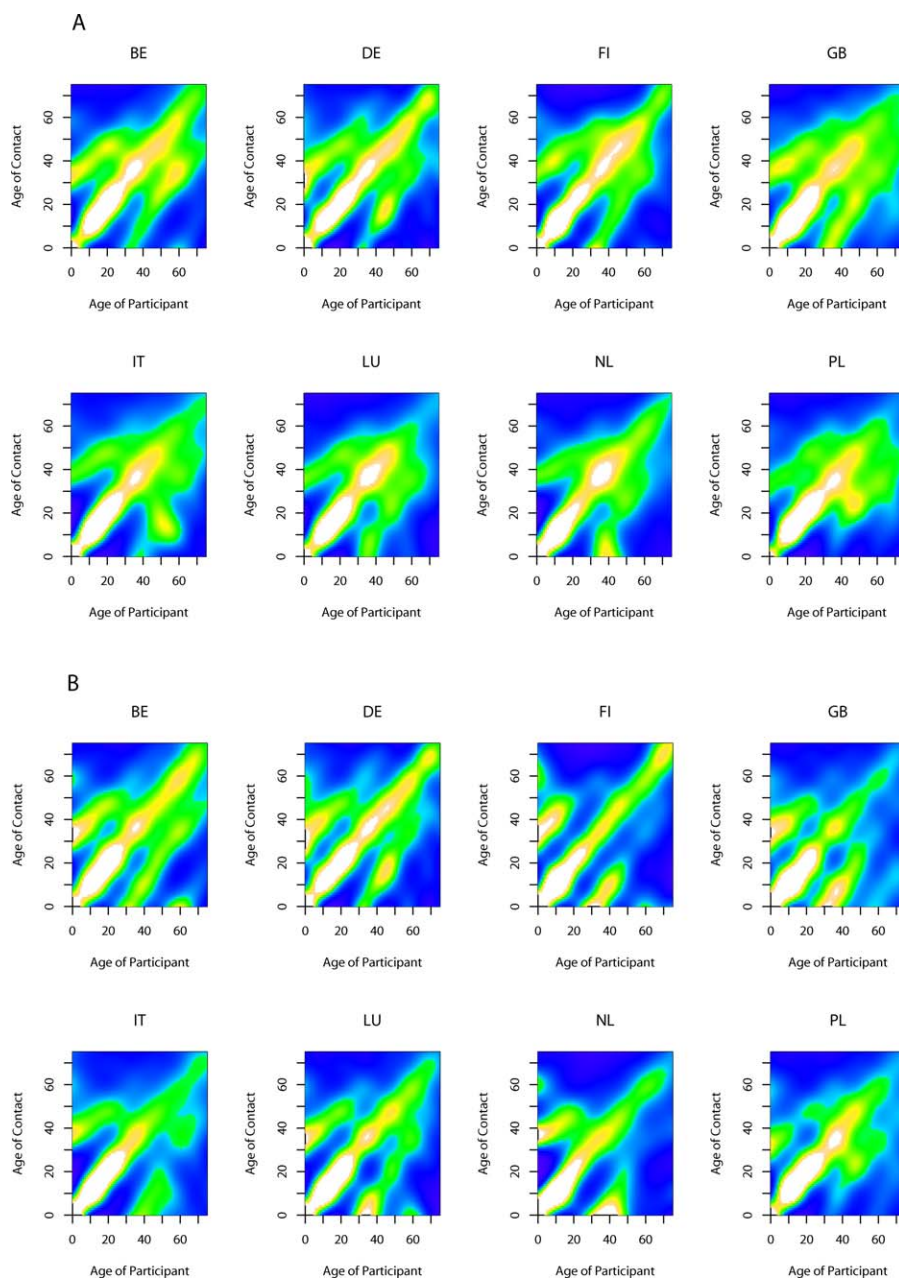


Figure 3. Smoothed Contact Matrices for Each Country Based on (A) All Reported Contacts and (B) Physical Contacts Weighted by Sampling Weights. White indicates high contact rates, green intermediate contact rates, and blue low contact rates, relative to the country-specific contact intensity. Fitting is based on a tensor-product spline to contact matrix data using a negative binomial distribution to account for overdispersion.
doi:10.1371/journal.pmed.0050074.g003

We have used simulations to expand on two particular types of contacts (physical and nonphysical) and to sketch the consequences of the observed contact patterns on the age distribution of incidence in the initial phase of an epidemic, when a new infectious disease is introduced into a completely susceptible population. As shown clearly by our simulations, the highest incidence of infection will occur among the younger age classes (5–19 y) for all countries. It is tempting to link such contact patterns to the observation during the 1957 Asian influenza A H2N2 pandemic that the first few generations of infection primarily affected those aged 11–18 y [35]. However, we note that our survey did not address the clustering of contacts; such clustering of contacts might

result in less-pronounced differences in age-specific incidence than suggested by our calculations. Addressing the frequency of clustered contacts, duration and type of contact, differential impact of pathogen on different age groups, time correlation of contacts, and assortative mixing by demographic factors other than age should be key priorities for future research.

One of the major assumptions behind our approach is that talking with or touching another person constitutes the main at-risk events for transmitting infectious diseases. There may be other at-risk events that our methodology does not capture, such as being in a confined space or in close physical proximity with other individuals and not talking to them [23].

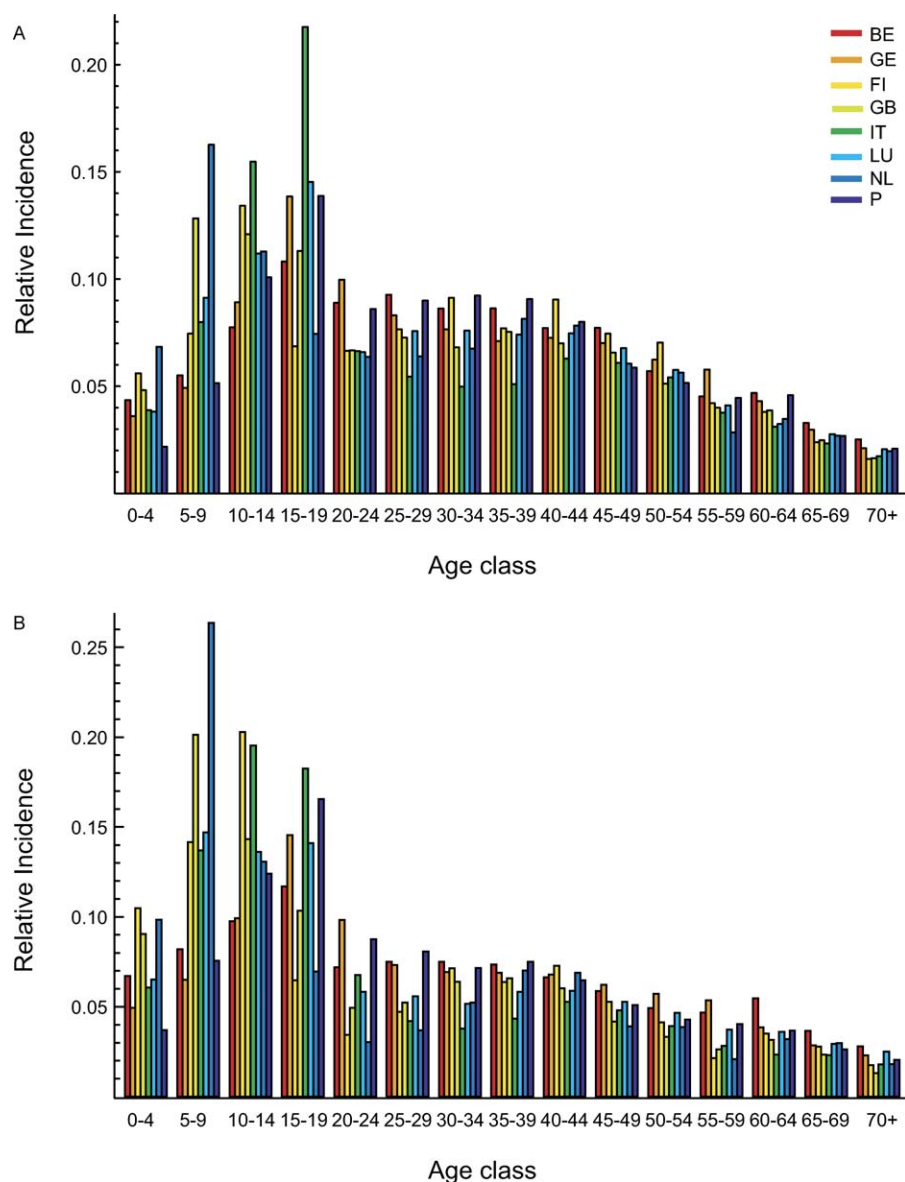


Figure 4. Relative Incidence of a New Emerging Infection in a Completely Susceptible Population, When the Infection Is Spread between and within Age Groups by the Contacts as Observed in Figure 3

For each country, we monitored incidence five generations of infection after the introduction of a single infected individual in the 65–70 age group; the incidence is normalized such that height of all bars sums to one for each country. (A) Results for all reported contacts; (B) for physical contacts only. doi:10.1371/journal.pmed.0050074.g004

Such events are difficult to record or to measure without using intrusive and expensive surveillance methods, and are probably of lower risk than the communication events captured by our approach. Similarly, our framework does not apply to pathogens that, in addition to the respiratory route, can be also spread by other means, for example, the sewage contamination events for SARS [8]. Although we believe that it is plausible that the contact patterns observed in our study are predictive of disease transmission, further work is clearly needed to establish the types of contacts that represent transmission risks for different diseases and to determine the circumstances under which lower-intensity contacts could be epidemiologically relevant. The data reported in this study should not be considered a substitute for epidemiological studies that quantify, for instance, the

intensity of transmission of influenza in households, schools, or other settings. However, this study does provide invaluable data on the relative importance of “leisure” and “other” contacts, which are very difficult to assess in other ways, and it highlights the relatively small contribution of personal contacts during travel based on our approach of defining a contact.

Using contact diaries in the general population was a feasible method for our specific study objectives, but as with all self-reported data, future research should validate our findings with different approaches, including interviews or direct observation. The latter might be particularly useful in assessing contacts of young children who spend time in day-care centres and kindergartens, because parental proxy reporting for young children is likely to be problematic.

Despite the limitations of self-reported egocentric data [36], contact diaries can provide extensive details regarding contact structures and have been used successfully for social network analysis [37]. Our contact diaries yielded detailed information about intimacy, frequency, and epidemiological relevance of contacts with an acceptable burden on respondents. In five countries, participants were given the opportunity to report whether they had any problems filling in the diary. The low proportion reporting problems (4% in adults, 4.9% in older children self-reporting contacts, and 4.9% in parents as proxy for children) suggest that the contact diary was readily accepted and understood by responding participants.

A further limitation of our study is that the comparison of contact patterns between countries is complicated by the variations of diary design (see Table S1), recruitment, and follow-up methodology (see Table S1). Our surveys were conducted in each country by different commercial companies with different recruitment and follow-up methods. Conducting surveys on contact behaviour and networks that entail a certain burden on participants and follow identical methodology in different countries is a challenging task, given that cultural factors in response also play a role. Further research is definitely warranted to determine optimal survey methodologies in different international settings, including developing countries, to improve comparability of contact data. Diaries used in BE, DE, FI, and NL instructed respondents not to record all of their professional contacts, but to provide an estimate if they had a lot of them. The reason for this instruction was to try to capture information from those people who make very large numbers of contacts (shop assistants and bus drivers, for instance), given that it might be very difficult or impossible for such people to fill out the full contact diary. This instruction may have led to some underreporting of contact frequencies and thus have affected the distribution of age and circumstance of contacts for these four countries, although we have taken account of this possibility to some extent using a censored model. Additional analyses for these countries that combine and compare the estimated frequency of professional contacts with the diary data will provide additional insights about the number of contacts for all countries. The differences between diaries do not, however, affect the age-specific pattern, nor the similarity in age-specific patterns found across countries.

Our survey is, to our knowledge, the first population-based prospective survey of mixing patterns pertinent to the spread of airborne and close-contact infectious diseases performed in several European countries using a similar diary methodology. The quantification of these mixing patterns shows a remarkable similarity in degree of assortativeness, which likely results in similar patterns of spread in different populations. This finding represents a significant advance in our understanding of the spread of these infectious diseases and should help to improve the parameterisation of mathematical models used to design control strategies.

Supporting Information

Figure S1. Histogram of Number of Reported Contacts by Country
Found at doi:10.1371/journal.pmed.0050074.sg001 (7.7 KB PDF).

Figure S2. The Proportion of Contacts That Involved Physical

Contact, by (a) Duration, (b) Frequency, (c) Location of Contact; and (d) Correlation between Duration and Frequency of Contact

Contacts were weighted by country-specific sampling weights in BE (A), DE (B), FI (C), GB (D), IT (E), LU (F), NL (G), and PL (H).

Found at doi:10.1371/journal.pmed.0050074.sg002 (84 KB DOC).

Figure S3. Smoothed Weighted Contact Matrices for Each Country Based on Reported Contacts Occurring in the Home Setting

White indicates high contact rates, green intermediate contact rates, and blue low contact rates. Fitting is based on a tensor-product spline to contact matrix data using a negative binomial distribution to account for overdispersion.

Found at doi:10.1371/journal.pmed.0050074.sg003 (282 KB PDF).

Figure S4. Smoothed Weighted Contact Matrices for Each Country Based on Reported Contacts Occurring in the Work Setting

White indicates high contact rates, green intermediate contact rates, and blue low contact rates. Fitting is based on a tensor-product spline to contact matrix data using a negative binomial distribution to account for overdispersion.

Found at doi:10.1371/journal.pmed.0050074.sg004 (254 KB PDF).

Table S1. Details of Survey Methodology in Each Country

Found at doi:10.1371/journal.pmed.0050074.st001 (52 KB DOC).

Table S2. Comparison of Household Size and Age Distribution of Census Data (2000) and Sample in BE, DE, FI, GB, IT, LU, NL, and PLRatio C/S (census versus sample), corresponds to the sampling weights used in the statistical analysis.

Found at doi:10.1371/journal.pmed.0050074.st002 (715 KB DOC).

Table S3. Relative Number of Reported Contacts Estimated by Different Negative Binomial Models (95% Confidence Interval in Brackets)

The results of this model comparison show that neither the censored nature of the data, nor the differences in how professional contacts were handled, substantially changes the model outcome. Note that all covariates have overlapping confidence intervals for models A and B, which are directly comparable, although censoring does improve model fit.

Found at doi:10.1371/journal.pmed.0050074.st003 (282 KB PDF).

Table S4. Association Rules of Length 2 for Type, Duration, and Location of Contacts with Minimal Support of 0.5%, Significant Positive Dependence (0.01 Significance Level)

Support, confidence, lift, and χ^2 values are given.

Found at doi:10.1371/journal.pmed.0050074.st004 (254 KB PDF).

Table S5. Contact Matrices of All Reported and Physical Contacts Consisting of the Average Number of Contact Persons Recorded per Day per Survey Participant Separately for Each Country

Found at doi:10.1371/journal.pmed.0050074.st005 (714 KB DOC).

Text S1. Example of the Diary Used in Great Britain

Found at doi:10.1371/journal.pmed.0050074.sd001 (49 KB PDF).

Acknowledgments

This study has benefited greatly by the input of the following participants of the POLYMOD project: Marc Aerts, Oualid Akakzia, Francesco Billari, Paddy Farrington, Vera Friedrichs, Nigel Gay, Marc Hoffmann, Mirjam Kretzschmar, Piero Manfredi, Marja Snellman, Ziv Shkedy, Pierre van Damme, and Heather Whitaker.

Author contributions. JM, NH, MJ, JW, and WJE drafted the manuscript in consultation with all the other authors; the original idea and contact diary were conceived by WJE. JM conducted a pilot study on an adapted diary, and coordinated overall survey design and data collection. JM, MJ, PB, KA, RM, MM, GST, JW, JH, MST, and MR conducted the surveys in their respective countries. NH conducted the data mining and surface smoothing. JW and JH conducted the epidemic modelling. All authors approved the final manuscript.

References

- Longini IM Jr., Nizam A, Xu S, Ungchusak K, Hanshaworakul W, et al. (2005) Containing pandemic influenza at the source. *Science* 309: 1083–1087.

2. Germann TC, Kadau K, Longini IM Jr., Macken CA (2006) Mitigation strategies for pandemic influenza in the United States. *Proc Natl Acad Sci U S A* 103: 5935–5940.
3. Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, et al. (2006) Strategies for mitigating an influenza pandemic. *Nature* 442: 448–452.
4. Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* 160: 509–516.
5. Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, et al. (2006) Real-time estimates in early detection of SARS. *Emerg Infect Dis* 12: 110–113.
6. Pitman RJ, Cooper BS, Trotter CL, Gay NJ, Edmunds WJ (2005) Entry screening for severe acute respiratory syndrome (SARS) or influenza: policy evaluation. *BMJ* 331: 1242–1243.
7. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, et al. (2003) Transmission dynamics and control of severe acute respiratory syndrome. *Science* 300: 1966–1970.
8. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, et al. (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* 300: 1961–1966.
9. Dye C, Gay N (2003) Epidemiology. Modeling the SARS epidemic. *Science* 300: 1884–1885.
10. Wu JT, Riley S, Fraser C, Leung GM (2006) Reducing the impact of the next influenza pandemic using household-based public health interventions. *PLoS Med* 3: e361. doi:10.1371/journal.pmed.0030361
11. Vynnycky E, Edmunds WJ (2007) Analyses of the 1957 (Asian) influenza pandemic in the United Kingdom and the impact of school closures. *Epidemiol Infect*: 1–14.
12. Hatchett RJ, Mecher CE, Lipsitch M (2007) Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc Natl Acad Sci U S A* 104: 7582–7587.
13. Carrat F, Luong J, Lao H, Salle AV, Lajaunie C, et al. (2006) A 'small-world-like' model for comparing interventions aimed at preventing and controlling influenza pandemics. *BMC Med* 4: 26.
14. Wallinga J, Teunis P, Kretzschmar M (2006) Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol* 164: 936–944.
15. Fenton KA, Korolessis C, Johnson AM, McCadden A, McManus S, et al. (2001) Sexual behaviour in Britain: reported sexually transmitted infections and prevalent genital *Chlamydia trachomatis* infection. *Lancet* 358: 1851–1854.
16. Garnett GP, Hughes JP, Anderson RM, Stoner BP, Aral SO, et al. (1996) Sexual mixing patterns of patients attending sexually transmitted diseases clinics. *Sex Transm Dis* 23: 248–257.
17. Gregson S, Nyamukapa CA, Garnett GP, Mason PR, Zhuwau T, et al. (2002) Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *Lancet* 359: 1896–1903.
18. Brewer DD, Garrett SB (2001) Evaluation of interviewing techniques to enhance recall of sexual and drug injection partners. *Sex Transm Dis* 28: 666–677.
19. Farrington CP, Kanaan MN, Gay N (2001) Estimation of the basic reproduction number for infectious diseases from age-stratified serological data. *Applied Statistics* 50: 251–292.
20. Farrington CP, Whitaker HJ (2005) Contact surface models for infectious diseases: estimation from serologic survey data. *J Am Stat Assoc* 100: 370–379.
21. Edmunds WJ, O'Callaghan CJ, Nokes DJ (1997) Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proc Bio Sci* 264: 949–957.
22. Edmunds W, Kafatos G, Wallinga J, Mossong J (2006) Mixing patterns and the spread of close-contact infectious diseases. *Emerg Themes Epidemiol* 3: 10.
23. Beutels P, Shkedy Z, Aerts M, Van Damme P (2006) Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiol Infect* 134: 1158–1166.
24. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M (2007) Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiol Infect*. Epub 18 July 2007. doi:10.1017/S0950268807009181
25. Mikolajczyk RT, Kretzschmar M (2008) Collecting social contact data in the context of disease transmission: prospective and retrospective study designs. *Soc Networks*. In press.
26. Akakzia O, Friedrichs V, Edmunds WJ, Mossong J (2007) Comparison of paper diary vs computer assisted telephone interview for collecting social contact data relevant to the spread of airborne infectious diseases. *Eur J Public Health* 17: 189.
27. Hilbe JM (2007) Negative binomial regression. Cambridge (UK): Cambridge University Press.
28. Hahsler M, Grün B, Hornik KU (2006) arules: Mining Association Rules and Frequent Itemsets. R package version 0.4–3. ed.
29. Hahsler M, Hornik K (2007) New probabilistic interest measures for association rules. *Intell Data Anal* 11: 437–455.
30. Wood SN (2006) Generalized Additive Models: An Introduction with R. Boca Raton (Florida): Chapman & Hall.
31. Hens N, Aerts M, Shkedy Z, Kung'u Kimani P, Kojouharova M, et al. (2007) Estimating the impact of vaccination using age-time-dependent incidence rates of hepatitis B. *Epidemiol Infect* 136: 341–351.
32. Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious diseases. Model building, analysis and interpretation. Chichester (UK): Wiley.
33. Bansal S, Pourbohloul B, Meyers LA (2006) A comparative analysis of influenza vaccination programs. *PLoS Med* 3: e387. doi:10.1371/journal.pmed.0030387
34. Eubank S, Guclu H, Kumar VS, Marathe MV, Srinivasan A, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429: 180–184.
35. Fry J (1958) Influenza A (Asian) 1957: clinical and epidemiological features in a general practice. *Br Med J* 1: 259–261.
36. Marsden PV (2005) Recent developments in network measurement. In: Carrington P, Scott J, Wassermann S, editors. Cambridge (UK): Cambridge University Press. pp. 8–30.
37. Fu Y-C (2007) Contact diaries: building archives of actual and comprehensive personal networks. *Field Methods* 19: 194–217.

Editors' Summary

Background To understand and predict the impact of infectious disease, researchers often develop mathematical models. These computer simulations of hypothetical scenarios help policymakers and others to anticipate possible patterns and consequences of the emergence of diseases, and to develop interventions to curb disease spread. Whether to prepare for an outbreak of infectious disease or to control an existing outbreak, models can help researchers and policy makers decide how to intervene. For example, they may decide to develop or stockpile vaccines or antibiotics, fund vaccination or screening programs, or mount health promotion campaigns to help citizens minimize their exposure to the infectious agent (e.g., handwashing, travel restrictions, or school closures).

Respiratory infections, including the common cold, flu, and pneumonia, are some of the most prevalent infections in the world. Much work has gone into modeling how many people would be affected by respiratory diseases under various conditions and what can be done to limit the consequences.

Why Was This Study Done? Mathematical models have tended to use contact rates (the number of other people that a person encounters per day) as one of their main elements in predicting the outcomes of epidemics. In the past, contact rates were not based on direct observations, but were assumed to follow a certain pattern and calibrated against other indirect data sources such as serological or case notification data. This study aimed to estimate contact rates directly by asking people who they have met during the course of one day. This allowed the researchers to study in more detail different *patterns* of contacts, such as those between different groups of people (such as age groups) and in different social settings. This is particularly important for respiratory diseases, which are spread through the air and by close contact with an infected individual or surface.

What Did the Researchers Do and Find? The researchers wanted to examine the social contacts that people have in order to better understand how respiratory infections might spread. They recruited 7,290 people from eight European countries (Belgium, Germany, Finland, Great Britain, Italy, Luxembourg, The Netherlands, and Poland) to participate in their study. They asked the participants to fill out a diary that documented their physical and nonphysical contacts for a single day. Physical contacts included interactions such as a kiss or a handshake. Nonphysical contacts were situations such as a two-way conversation without skin-to-skin contact. Participants detailed the location and duration of each contact. Diaries also contained basic demographic information about the participant and the contact.

They found that these 7,290 participants had 97,904 contacts during the study, which averaged to 13.4 contacts per day per person. There was a great deal of diversity among the contacts, which challenges the idea that contact rates alone provide a complete picture of transmission dynamics. The researchers identified varied types of contacts, duration of contacts, and mixing patterns. For example, children had more contacts than adults, and those living in larger households had more contacts. Weekdays resulted in more daily contacts than Sundays. More intense contacts (of longer duration or more frequent) tended to be physical. Approximately 70% of contacts made on a daily basis lasted longer than an hour, whereas three-quarters of contacts with people who were not previously known lasted less than 15 minutes. While mixing patterns were very similar across the eight countries, people of the same age tended to mix with each other.

Analyzing these contact patterns and applying mathematical and statistical techniques, the researchers created a model of the initial phase of a hypothetical respiratory infection epidemic. This model suggests that 5- to 19-year-olds will suffer the highest burden of respiratory infection during an initial spread. The high incidence of infection among school-aged children in the model results from these children having a large number of contacts compared to other groups and tending to make contacts within their own age group.

What Do These Findings Mean? This work provides insight about contacts that can be supplemental to traditional measurements such as contact rates, which are usually generated from household or workplace size and transportation statistics. Incorporating contact patterns into the model allowed for a deeper understanding of the transmission patterns of a hypothetical respiratory epidemic among a susceptible population. Understanding the patterning of social contacts—between and within groups, and in different social settings—shows how diverse contacts and mixing between individuals really are. Physical exposure to an infectious agent, the authors conclude, is best modeled by taking into account the social network of close contacts and its patterning.

Additional Information. Please access these Web sites via the online version of this summary at doi:10.1371/journal.pmed.0050074.

- Wikipedia has technical discussions on the assumptions used in mathematical models of epidemiology (note that Wikipedia is a free online encyclopedia that anyone can edit; available in several languages)
- Plans for pandemic influenza are explained for the Government of Canada, the United Kingdom's Health Protection Agency, and the United States Department of Health and Human Services