Networks of infection: Online respondent-driven detection for studying infectious disease transmission and case finding
PhD thesis, Utrecht University, the Netherlands

ISBN:	978-90-393-6664-6
Design and layout:	Wasamon Sabaiwan-Stein
Cover illustration:	Wasamon Sabaiwan-Stein
Printing:	GVO drukkers & vormgevers B.V.

Financial support for the publication of this thesis by the Julius Center for Health Sciences and Primary Care and the National Institute for Public Health and the Environment (RIVM) is gratefully acknowledged.

Networks of infection:

Online respondent-driven detection for studying infectious disease transmission and case finding

Infectienetwerken:
Online respondentgestuurde detectie voor de bestudering van de verspreiding van infectieziekten en de opsporing van geïnfecteerden
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, Prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

dinsdag 13 december 2016 des ochtends te 10.30 uur

door

Mart Lambertus Stein

geboren op 26 maart 1987
te Houten

Promotoren:
Prof. dr. M.E.E. Kretzschmar
Prof. dr. P.G.M. van der Heijden
Prof. dr. ir. V. Buskens

Copromotor:
Dr. J.E. van Steenbergen

Beoordelingscommissie:

Prof. dr. ir. H.A. Smit
Prof. dr. T.J.M. Verheij
Prof. dr. J.B.F. de Wit
Prof. dr. C.J.P.A. Hoebe
Dr. S.D.W. Frost

แด่ภรรยาและลูกชายสุดที่รักของผม

CONTENTS

Chapter 1

GENERAL INTRODUCTION

Mart L. Stein

Although the global disease burden due to infectious diseases decreased in the last decades[1], the number of introductions of new pathogens and efforts needed for their control increased[2]. Factors related to human contact behaviour and increased mobility in modern society (e.g., increased air travel) influence the transmission and importation of pathogens[3]. Outbreak management has become an essential element in infectious disease control. Key public health priorities are to identify cases (i.e., a case is an individual with an infection, either with or without symptoms and either being or not being infectious), treat them, and control outbreaks by interrupting pathogen transmission[4]. An effective management of epidemic or endemic situations requires comprehensive disease surveillance, i.e., the continuous and systematic collection, analysis and interpretation of data on cases[5, 6]. However, case finding and outbreak management is often challenging and causes an intensive work load for public health professionals[7, 8]. This is illustrated by the management of the severe acute respiratory syndrome (SARS), Ebola and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) outbreaks[9-13].

Example of outbreak management

In 2012, MERS-CoV was for the first time isolated from a 60-year-old man in Saudi Arabia, who suffered from acute pneumonia and subsequent renal failure with a fatal outcome[14]. The virus spread to South Korea through one case who infected locally many other persons, which suggests a super-spreading event (i.e., a case who infects more than ten persons[15]). Super-spreading events were also observed during an outbreak of severe acute respiratory syndrome (SARS) in 2003[16]. Only a few travel-related MERS cases were reported in Europe and the United States[9]. Nevertheless, all contact persons of confirmed cases required comprehensive monitoring, causing a heavy work load for public health professionals[17, 18]. Surveillance data was likely incomplete due to a selection bias towards more severe cases. The majority of MERS-CoV infections seem to have presented only mild symptoms and did not visit a health care facility, and therefore went undetected[19]. This makes it difficult to control, and to estimate the extent and impact of an outbreak.

Networks of infection

The exact mode of human-to-human transmission is to be determined, but it is believed that MERS-CoV, like other coronaviruses and upper respiratory tract infections, is transmitted via respiratory secretions, such as coughing and sneezing[19]. Therefore, close person-to-person contact, such as touching, or sitting or standing within arm's length of each other, is assumed to be a good indicator for potential transmission events[4, 20]. Many infectious diseases, such as MERS-CoV, measles and sexually transmitted infections (STIs), therefore do not spread at random in a population, but follow the structure of contact networks. A

network is a representation of connections between individuals who are depicted by points joined together by lines, where points are referred to as nodes and lines as edges[21]. In this thesis, we are interested in networks of individuals, and in particular, in social and contact (or 'physical') networks. With social networks, edges represent specific relations between individuals such as friendships and co-worker relationships. With contact networks, edges are assumed to represent a contact that is relevant for the transmission of pathogens. In reality, social and contact networks presumably overlap considerably. Also, household members, colleagues, school children and other socially connected individuals, tend to have prolonged and repeated contacts that increase the likelihood of pathogen transmission[22, 23]. Empirical data on the (potential) transmission of pathogens in these contact networks are needed to inform mathematical models, which can provide evidence-based support for public health policy decisions[24]. However, so far, the majority of studies used an egocentric design to collect data on contact patterns, and collected no information about the network beyond the contact persons reported by participants[20].

Online respondent-driven detection

The hypothesis underlying the studies in this thesis is that by using the contact network of cases, identification of cases of emerging and re-emerging pathogens can be enhanced, and knowledge on the transmission of pathogens within these networks can be increased. In this thesis, the use of respondent-driven detection (RDD) via the Internet was piloted for sampling contact networks of individuals in the general population. RDD is based on the ideas of respondent-driven sampling, a variant of snowball sampling, which was initially introduced to estimate the prevalence of disease or risk factors in hard-to-reach populations (such as drug-users)[25]. With RDD, participants are asked to recruit contact persons whom they have met face-to-face during a specified time period. These contact persons are then asked to do the same, creating chains of contact persons connected through recruitment ('recruitment trees'). Unlike with snowball sampling, peer recruitment is tracked by means of personal codes. Such an approach may provide novel insights in contact network structures compared to studies that sampled individuals randomly and independently of one another[20]. As the purpose of RDD is to detect cases or clusters of disease within contact networks, rather than to estimate population proportions from a sample, RDD is in this thesis distinguished from respondent-driven sampling.

Objectives

The topic of this thesis falls within a new and fast expanding research field called 'digital epidemiology'[26]. This multidisciplinary research field includes tools from epidemiology and social sciences, and the use of the Internet, mobile phones and social media. It describes pioneering work that is highly innovative in multiple aspects.

The objectives of this thesis were to investigate the feasibility of using the contact network of cases to enhance case finding during outbreaks of emerging or re-emerging pathogens, and to analyse transmission dynamics within these networks. Studies included in this thesis pilot the use of online RDD to collect information on contact networks of cases and non-cases in the general population in the Netherlands and Thailand. Various statistical methods in the field of infectious disease epidemiology and social sciences were combined to compare characteristics of sampled individuals and examine mixing patterns. Factors important for the success of the online respondent-driven method were analysed by mathematical modelling.

Studies in this thesis focus on gaining information on transmission patterns of influenza-like-illness (ILI) symptoms. ILI is a medical diagnosis based on a set of common clinical symptoms (e.g., fever and cough or sore throat), possibly caused by an infection with influenza virus or one of many other ILI-causing pathogens (e.g., coronavirus, adenovirus, rhinovirus, etc.)[27]. ILI was chosen as it is easily captured with an online questionnaire and frequently occurring, especially between November and April in the Northern Hemisphere[28]. However, new insights in contact networks that are provided in this thesis are also relevant for other pathogens transmitted via similar or other close (non-sexual) contact routes.

Collaboration and software used in this thesis

This thesis is the result of work performed at Julius Center for Health Sciences and Primary Care of the University Medical Center Utrecht in close collaboration with the Centre for Infectious Disease Control of the National Institute for Public Health and the Environment (RIVM), the Faculty of Social and Behavioural Sciences of Utrecht University, the Netherlands, the Department of Public Health Sciences-Global Health of the Karolinska Institutet, Sweden, and the Faculty of Public Health of the Mahidol University in Bangkok, Thailand.

The first pilot study was conducted between May and June 2011 in the Netherlands, during a mumps outbreak among university students[29, 30]. A short online questionnaire was distributed via Facebook among a small group of close, seemingly healthy, friends of the researchers to test the feasibility of online peer recruitment, and whether symptomatic cases could be identified. Fifty-eight persons completed the survey, which reached up to four waves of

contact persons and included two cases of self-reported mumps. The survey was conducted using the freely available survey software "Google Forms". This service does not contain any functions for online respondent-driven sampling, such as an option for participants to send invitations to their contact persons or an automated system to track who invites whom. Therefore, recruitment could not be done anonymously. Participants were asked to provide their Facebook name and the ones of their recruiter, to be able to link contact persons. This was criticized by some participants and led in the analysis to difficulties with missing links.

During the second half of 2011, software suitable for online respondent-driven sampling became available through establishment of an international collaboration with the Karolinska Institutet. The Karolinska Institutet had developed and successfully implemented software for an online respondent-driven sampling study among men-who-have-sex-with-men in Vietnam[31]. Their software, based on the open-source survey tool "LimeSurvey", was used in our project as a starting point for a generic software system that could easily be used by researchers for all kinds of respondent-driven surveys, with only little support from programmers. The subsequent development was performed by Chinese programmers. Considerable time was invested to implement, among others, easy survey access for new participants, and new email invitation templates and recruitment options (such as sending email invitations directly to contact persons and sending invitations via Facebook private messenger). The software was subsequently used for data collections in the Netherlands and Thailand. The ultimate goal of this collaboration is the development of a platform for online respondent-driven research, where other researchers can use the survey software system, share knowledge and develop new statistical tools for the analysis of data collected with online respondent-driven methods.

The collaboration with the Mahidol University came forth from an earlier research project conducted in Thailand and five other Southeast Asian countries, and a mutual interest in conducting online respondent-driven sampling research[32, 33]. Southeast Asia is a hotspot for emerging infectious diseases, including those with pandemic potential[34]. Thailand has around 66 million inhabitants, of which 8.3 million are registered in the densely populated capital Bangkok. There is limited empirical data available on contact networks in the Thai population, nor on differences between contact patterns in Thailand and populations that differ considerably in socio-cultural characteristics. In addition, few studies tested an online sampling method in Thailand. The Internet-use in urban areas in Thailand was thought to be sufficient (51.5% in 2012[35]) to pilot an online RDD survey in Bangkok.

Outline of this thesis

In part I of this thesis, we piloted the use of online RDD to study contact network patterns, which are relevant for pathogens that are transmitted by the respiratory or close-contact route, and transmission of pathogens within these networks. Such data can inform mathematical models that are used to understand transmission dynamics and to estimate the effectiveness of control measures during infectious disease outbreaks. In **Chapter 2**, the feasibility of online RDD was tested for studying contact networks of students of two Bangkok universities in Thailand. A combination of statistical methods from epidemiology and social sciences was used to analyse structure and correlations of recruitment trees. In **Chapter 3**, the respondent-driven survey conducted in Thailand was repeated among students of two universities in the Netherlands. A cross-cultural comparison was made for online peer recruitment, characteristics of participants and contact networks, using data collected in both countries. In **Chapter 4**, we tested the feasibility of using RDD to collect information on the transmission of pathogens within networks. To do so, we collaborated with the founders of two large participatory surveillance panels in the Netherlands and Dutch speaking Flanders (Belgium). These Internet-based systems capture voluntarily submitted data on ILI symptoms during the winter season from the general public[36, 37]. Volunteers of these panels were invited to our respondent-driven survey and were asked to report numbers of contact persons at different locations and self-reported ILI symptoms. A large number of recruiter-recruit pairs were sampled, consisting of individuals of different ages and backgrounds. This enabled us to quantify, among others, with more certainty mixing patterns that allow the transmission of diseases that spread via close contact.

Part II focuses on identifying new cases through a known case using online RDD. In **Chapter 5**, we used the same dataset that we collected with the Dutch participatory surveillance system and assessed whether symptomatic cases recruited other symptomatic cases via their contact network. In **Chapter 6**, we discuss the use of social media (such as Facebook) by public health professionals for contact tracing, and in particular, the ethical concerns this may raise. In part III, we investigate the drivers of the peer recruitment process. In **Chapter 7**, we use our empirically collected data and a mathematical model to analyse how the success of sampling is influenced by various model parameters and which factors are most important for the success of RDD. Such information can inform future online RDD surveys in different target populations, by providing insight into which factors researchers should focus to increase the sample size or to influence the sample composition.

Finally, in **Chapter 8**, an overall discussion is provided on the use of online RDD for case finding and public health interventions. This chapter provides an overall perspective on all empirical studies and discusses implications, remaining challenges and future research.

REFERENCES

1. Murray CJ, Ezzati M, Flaxman AD, Lim S, Lozano R, Michaud C, et al. GBD 2010: a multi-investigator collaboration for global comparative descriptive epidemiology. Lancet. 2012;380(9859):2055-8.

2. Smith KF, Goldberg M, Rosenthal S, Carlson L, Chen J, Chen C, et al. Global rise in human infectious disease outbreaks. J R Soc Interface. 2014;11(101):20140950.

3. Morse SS. Factors in the emergence of infectious diseases. Emerg Infect Dis. 1995;1(1):7-15.

4. Heymann DL. Control of Communicable Diseases Manual. 19 ed: American Public Health Association; 2008.

5. Porta M. A dictionary of epidemiology. 6th ed. Greenland S, Hernán M, dos Santos Silva I, Last JM, editors. Oxford: Oxford University Press; 2014.

6. Reintjes R, Krickeberg K. Epidemiologic Surveillance. In: Krämer A, Kretzschmar M, Krickeberg K, editors. Modern Infectious Disease Epidemiology Concepts, Methods, Mathematical Models, and Public Health. LLC: Springer Science+Business Media; 2010. p. 143-58.

7. World Health Organization, Center for Disease Control and Prevention. Emergency guideline for the implementation and management of contact tracing for Ebola virus disease. September, 2015.

8. Macera CA, Shaffer R, Shaffer PM. Outbreak investigations. Introduction to Epidemiology: Distribution and Determinants of Disease. first ed2013.

9. European Centre for Disease Prevention and Control (ECDC). Severe respiratory disease associated with Middle East respiratory syndrome coronavirus (MERS-CoV). Stockholm: ECDC, 21 October 2015.

10. Chan SS, Leung GM, Tiwari AF, Salili F, Leung SS, Wong DC, et al. The impact of work-related risk on nurses during the SARS outbreak in Hong Kong. Fam Community Health. 2005;28(3):274-87.

11. Koh D, Lim MK, Chia SE, Ko SM, Qian F, Ng V, et al. Risk perception and impact of Severe Acute Respiratory Syndrome (SARS) on work and personal lives of healthcare workers in Singapore: what can we learn? Med Care. 2005;43(7):676-82.

12. Farquharson C, Baguley K. Responding to the severe acute respiratory syndrome (SARS) outbreak: lessons learned in a Toronto emergency department. J Emerg Nurs. 2003;29(3):222-8.

13. Frieden TR, Damon I, Bell BP, Kenyon T, Nichol S. Ebola 2014--new challenges, new global response and responsibility. The New England journal of medicine. 2014;371(13):1177-80.

14. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. The New England journal of medicine. 2012;367(19):1814-20.

15. Kucharski AJ, Althaus CL. The role of superspreading in Middle East Respiratory Syndrome Coronavirus (MERS-COV) transmission. Eurosurveillance. 2015;20(25):pii=21167.

16. Severe acute respiratory syndrome--Singapore, 2003. MMWR Morbidity and mortality weekly report. 2003;52(18):405-11.

17. Mollers M, Jonges M, Pas SD, van der Eijk AA, Dirksen K, al. e. Follow-up of Contacts of Middle East Respiratory Syndrome Coronavirus–Infected Returning Travelers, the Netherlands, 2014. Emerging Infectious Diseases. 2015;21(9).

18. Seroepidemiological investigation of contacts of Middle East respiratory syndrome coronavirus (MERS-CoV) patients [Internet]. 2013. Available from: http://www.who.int/csr/disease/coronavirus_infections/WHO_Contact_Protocol_MERSCoV_19_November_2013.pdf.

19. Cauchemez S, Van Kerkhove MD, Riley S, Donnelly CA, Fraser C, Ferguson NM. Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and how to tell them apart. Euro surveillance. 2013;18(24).

20. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiology and infection. 2012;140(12):2117-30.

21. Newman MEJ. Networks: an introduction. New York: Oxford University Press; 2010.

22. De Cao E, Zagheni E, Manfredi P, Melegaro A. The relative importance of frequency of contacts and duration of exposure for the spread of directly transmitted infections. Biostatistics. 2014;15(3):470-83.

23. Smieszek T. A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. Theor Biol Med Model. 2009;6:25. doi: 10.1186/1742-4682-6-25.

24. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, et al. Modeling infectious disease dynamics in the complex landscape of global health. Science. 2015;347(6227):aaa4339.

25. Heckathorn D. Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems. 1997;44:174-99.

26. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS computational biology. 2012;8(7):e1002616.

27. Cate TR. Clinical manifestations and consequences of influenza. Am J Med. 1987;82(6A):15-9.

28. Barr IG, McCauley J, Cox N, Daniels R, Engelhardt OG, Fukuda K, et al. Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009-2010 northern hemisphere season. Vaccine. 2010;28(5):1156-67.

29. Whelan J, van Binnendijk R, Greenland K, Fanoy E, Khargi M, Yap K, et al. Ongoing mumps outbreak in a student population with high vaccination coverage, Netherlands, 2010. Euro surveillance. 2010;15(17).

30. Sane J, Gouma S, Koopmans M, de Melker H, Swaan C, van Binnendijk R, et al. Epidemic of mumps among vaccinated persons, The Netherlands, 2009-2012. Emerg Infect Dis. 2014;20(4):643-8.

31. Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, Nguyen TA, et al. Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. PloS one. 2012;7(11):e49417.

32. Rudge JW, Hanvoravongchai P, Krumkamp R, Chavez I, Adisasmito W, Chau PN, et al. Health system resource gaps and associated mortality from pandemic influenza across six Asian territories. PloS one. 2012;7(2):e31800.

33. Stein ML, Rudge JW, Coker R, van der Weijden C, Krumkamp R, Hanvoravongchai P, et al. Development of a resource modelling tool to support decision makers in pandemic influenza preparedness: The AsiaFluCap Simulator. BMC Public Health. 2012;12:870.

34. Coker RJ, Hunter BM, Rudge JW, Liverani M, Hanvoravongchai P. Emerging infectious diseases in southeast Asia: regional challenges to control. Lancet. 2011;377(9765):599-609.

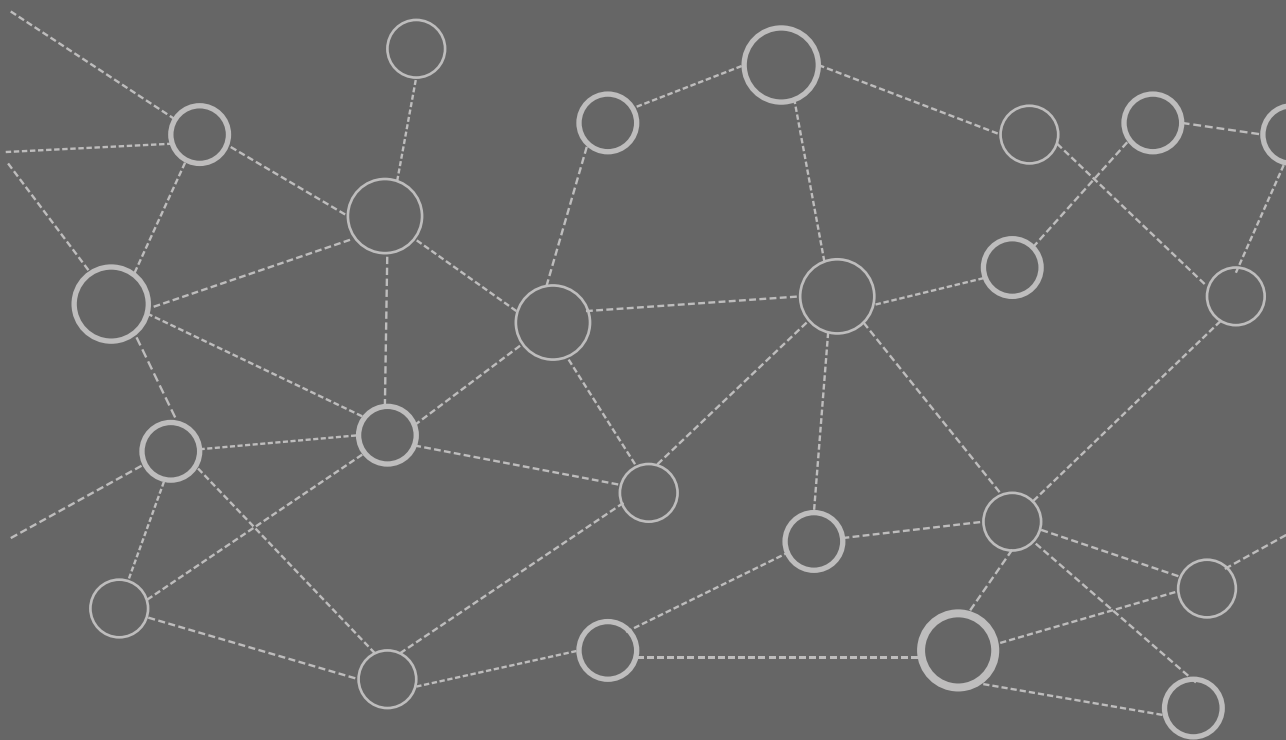
35. National Statistical Office. Ministry of Information and Communication Technology Thailand. The Information and Communication Technology Survey in Household. 2012.

36. Wojcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol. 2014;11:7.

37. Friesema IH, Koppeschaar CE, Donker GA, Dijkstra F, van Noort SP, Smallenburg R, et al. Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. Vaccine. 2009;27(45):6353-7.

Part I

Studying contact networks using an online respondent-driven method



Chapter 2

Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand.

Mart L. Stein[1,2*], Jim E. van Steenbergen[2,3], Charnchudhi Chanyasanha[4], Mathuros Tipayamongkholgul[4], Vincent Buskens[6], Peter G.M. van der Heijden[6,7], Wasamon Sabaiwan[8], Linus Bengtsson[9], Xin Lu[9,10], Anna E. Thorson[9], Mirjam E.E. Kretzschmar[1,2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[3] Centre for Infectious Diseases, Leiden University Medical Centre, Leiden, The Netherlands

[4] Department of Microbiology, Faculty of Public Health, Mahidol University, Bangkok, Thailand

[5] Department of Epidemiology, Faculty of Public Health, Mahidol University, Bangkok, Thailand

[6] Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, The Netherlands

[7] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom

[8] Faculty of Communication Arts, Chulalongkorn University, Bangkok, Thailand

[9] Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

[10] College of Information System and Management, National University of Defense Technology, Changsha, China

PLoS ONE 2014, 9(1):e85256

ABSTRACT

Background

Information on social interactions is needed to understand the spread of airborne infections through a population. Previous studies mostly collected egocentric information of independent respondents with self-reported information about contacts. Respondent-driven sampling (RDS) is a sampling technique allowing respondents to recruit contacts from their social network. We explored the feasibility of webRDS for studying contact patterns relevant for the spread of respiratory pathogens.

Materials and Methods

We developed a webRDS system for facilitating and tracking recruitment by Facebook and email. One-day diary surveys were conducted by applying webRDS among a convenience sample of Thai students. Students were asked to record numbers of contacts at different settings and self-reported influenza-like-illness symptoms, and to recruit four contacts whom they had met in the previous week. Contacts were asked to do the same to create a network tree of socially connected individuals. Correlations between linked individuals were analysed to investigate assortativity within networks.

Results

We reached up to 6 waves of contacts of initial respondents, using only non-material incentives. Forty-four (23.0%) of the initially approached students recruited one or more contacts. In total 257 persons participated, of which 168 (65.4%) were recruited by others. Facebook was the most popular recruitment option (45.1%). Strong assortative mixing was seen by age, gender and education, indicating a tendency of respondents to connect to contacts with similar characteristics. Random mixing was seen by reported number of daily contacts.

Conclusions

Despite methodological challenges (e.g., clustering among respondents and their contacts), applying RDS provides new insights in mixing patterns relevant for close-contact infections in real-world networks. Such information increases our knowledge of the transmission of respiratory infections within populations and can be used to improve existing modelling approaches. It is worthwhile to further develop and explore webRDS for the detection of clusters of respiratory symptoms in social networks.

INTRODUCTION

For important respiratory pathogens, like influenza, SARS and tuberculosis, spatial proximity between social contacts is a major determinant in the transmission process[1]. To understand the dynamics of transmission of pathogens through a population, information on contact patterns is needed. First explorations to quantify contact patterns relevant for respiratory infections using contact diary questionnaires were conducted by Edmunds et al. in 1997[2]. Subsequently, many other studies were performed in which the distribution of contacts was analysed among different age-groups[3], settings and across countries[4,5]. Various study designs were applied and tested for performance and validity[6-9]. Up till now, most of these studies focused on recording and analysing contact patterns of randomly sampled individuals. Surveys were 'egocentric' and contained only self-reported information about characteristics of contacts of respondents and the links between them. Other studies investigated participants wearing digital devices that sense their proximity to others including their spatial movements [10,11]. However, such studies can only be performed in specific settings and with small numbers of participants.

The rate at which infections spread across a community depends, among others, on the topology of the contact network[12]. Theoretically, it has been shown that network properties like the clustering of contacts and heterogeneity in 'degree' (i.e., the total number of contacts per individual) influence transmission dynamics[13-16]. For designing optimal control strategies, it is important to have knowledge on the network properties that allow or enhance the spread of infections along links in a network of individuals. More specifically, it might be advantageous to identify individuals who act as a bridge between communities or as hubs spreading to many other individuals.

Here we report on a pilot study using respondent-driven sampling (RDS) to recruit respondent as well as their close social contacts into a survey to study transmission related contact patterns. RDS is a variant of chain referral sampling, which includes ascertainment of degree distributions[17]. Different from snowball sampling, researchers keep track of who recruited whom and their numbers of social contacts. RDS is predominantly used for making prevalence estimations of characteristics in otherwise hard-to-reach populations such as injecting drugs users[18-20]. In this study we explored the technical feasibility and implementation of webRDS in a general population. We employed RDS with a different aim, namely as a sampling tool to study contact patterns relevant for the transmission of respiratory pathogens.

Increasing access to the internet, especially in low- and middle-income countries, offers new opportunities for epidemiological research[21-24]. Earlier studies found that social mixing patterns can be measured through simple internet-based surveys[6,25]. Until now, only few

studies used the internet for RDS (webRDS)[26]. A recent study performed in VietNam among internet-using MSM ('men who have sex with men') showed the potential of webRDS for sampling hidden and stigmatized populations[27]. The applicability of (web-)RDS for recording contacts relevant for respiratory-transmitted infections has, to our knowledge, not yet been investigated. We analysed the applicability of online social networks (i.e., Facebook) for inviting contacts, and the use of non-material incentives to stimulate recruitment. Secondly, we studied the correlations between individuals linked by recruitment chains and the distribution of connected components of the recruitment trees. Finally, we investigated whether webRDS can be used to detect clusters of influenza-like-illness symptoms in social networks

MATERIALS AND METHODS

Respondent Driven Sampling

RDS begins with the selection of initial respondents, called "seeds". The seed is asked to complete a survey and afterwards provided with a limited number of coupons (usually three or four) to invite contacts who are then asked to do the same. Limiting the number of coupons for each participant forces the sample recruitment chain to penetrate into the social network of seeds. This process continues in recruitment 'waves', either until the desired sampled size is reached or until the distribution of participants' characteristics has stabilised between waves or until chains go extinct, ensuring that the final sample is not biased by the choice of the seeds[17].

Web based RDS survey system

We developed a web based RDS survey system for facilitating and tracking online recruitment, based on a system that was implemented earlier by Bengtsson and colleagues[27]. We designed templates for invitation and reminder emails, which contained a link that provided direct access to the survey. These unique links were based on personal codes and automatically generated by the system for each participant. Each link could only be used once for filling in the survey to prevent repeated participation and participants seeing each others answers, and to control the number of friends that could be recruited. The questionnaire was divided over multiple pages (maximum of four questions per page). All text was provided in both Thai and English.
After submission of answers, respondents were redirected to a page for inviting contacts. We provided three options for inviting contacts: (1) sending an invitation email directly by the system to contact persons for whom email addresses were provided or (2) receive four separate invitation emails that could be forwarded to contact persons or (3) connect to Facebook to invite Facebook friends with a private Facebook message. For the first two email options the name of the recruiter was required, which appeared in the subject line of the email to personalise the invitation. As with the emails for seeds, the template of the invitation emails used for recruitment were standardised and contained background information about the

project, a (new) unique link to the questionnaire, and a personal code. All emails and the first page of the questionnaire contained a link to unsubscribe for the survey, which led to another page on which reason(s) for not participating could be provided.

For inviting friends on Facebook, participants needed to login to their personal account, followed by accepting our app. The app automatically created Facebook private messages, each containing a unique link to the survey that could be forwarded one by one. As with option 2, there was a possibility for the recruiter to add personal text to each message, next to the invitation text provided by us.

Illustration of network trees

As a non-material incentive, the progression of each network tree could be followed by the respondent on the accompanying institute website. Each network tree started with the code of the seed, which was linked to the corresponding codes of the recruited contacts. These contacts were each linked to contacts of contacts, and so on. The trees were anonymous and did not contain information provided by participants in the survey. All participants were referred to the institute website after sending out invitations. The network trees were updated daily and participants could return to the institute website at any moment.

Informed consent and privacy

Apart from recruitment, the survey was anonymous. The questionnaire pages were preceded by an informed consent page containing the research purposes, information on data security, subjects' privacy, confidentiality (e.g., data was not shared with any third parties) and non-material incentive, and the contact details of the researcher. Participation was voluntary and individuals could withdraw at any moment during the survey by closing the browser. Only after accepting the form, individuals could proceed to the questionnaire.

More detailed information about the research project was provided in a separate link, which could be accessed at any time during the survey. Furthermore, the logos of the participating research institutes were displayed on every page of the survey (as well as in every invitation email), and were linked to the associated websites. The system converted IP addresses to a unique anonymous code using a one-way encryption algorithm; the original IP addresses were deleted. All communication between the user and the server was encrypted. The online database was also encrypted and password protected.

We obtained ethical approval from the Medical Ethical Committee from both the Faculty of Public Health Mahidol University (Thailand) and University Medical Centre Utrecht (The Netherlands).

Study design, seeds, and setting

The developed webRDS system was applied for conducting contact diary questionnaires in Thailand between November 2012 and February 2013. As seeds we approached students from two Bangkok universities with an invitation email. Students, in physical group meetings varying between 6 and 30 persons, were first informed with a short oral presentation about the project, and afterwards contacted with an invitation email. In addition, respondents could become friends with the researcher on a Facebook page that was developed for this project, where updates on the network trees were presented and where participants could suggest friends as seeds (i.e., other than those personally approached).

The seeds were selected from a convenience sample of students, given sufficient spread of different curricula and academic years, in collaboration with lecturers and faculty deans of student affairs to prevent too much immediate overlap among contacts recruited by the seeds. Each seed was asked to recruit four close contacts (e.g. friends, family members and/or colleagues) whom they had met (according to the contact definition described below) in real-life in the past seven days. The time span of seven days was decided as it was a reasonable period for remembering close contacts and a seven days period includes the generation time of influenza [28,29]. Each participant was provided with an invitation for each of their four contacts. In principle, there were no exclusion criteria; however, a contact had to have access to the internet. We sent out reminder emails to seeds and contacts who did not respond within two weeks after sending the invitation email.

Thailand has around 65.9 million inhabitants (20.5 million households, with an average size of 3.2 persons), of which 8.3 million are registered in the densely populated capital Bangkok. In 2012, 27.5% of the Thai population in rural areas (aged six years and older) was using a computer, 20.5% the internet and 66.2% a mobile phone. Bangkok has higher proportions of users: 44.4%, 51.5% and 84.0% respectively. The age groups of 6-14, 15-24 and 25-34 use the internet most frequently (46.5%, 54.8%, and 29.7% respectively within each age group). Internet-use is much lower for ages of 35 years and above[30]. In May 2013, Thailand counted around 18.5 million Facebook accounts of which most users were between the ages of 18-34 years[31].

Online questionnaire

We asked participants to record the number of contacts they had during one full day (namely 'yesterday'). A contact was defined as a person standing or sitting close – defined as within reach of an arm's length[7,32] – to the participant for 30 seconds or longer. This space within arm's length was denoted as 'YourSpace'. The definition was illustrated with pictures to clearly indicate which contacts should be recorded.

To limit the burden for each participant, the online questionnaire was kept short; in total it consisted of eleven questions. Participants were asked to record the number of contacts while travelling (e.g., train, metro, bus, shuttle boot, minibus, car, tuk-tuk) and at different locations (e.g., home, work, school/university, restaurant, coffee shop, sport/leisure, concert or other places). For the different locations, participants were asked to specify for each contact whether this person was younger than, the same age or older than the participant.

In Thailand, it is custom to share food with friends, family and/or colleagues. Therefore we asked for the number of contacts (within arm's length) with whom participants had breakfast, lunch, dinner and/or a snack break. In addition, we included a question on the number and age (specified in age groups) of contacts that lived in the household in the past seven days. To facilitate participation, participants were instructed to leave answer options empty when these were not applicable to them (instead of having to fill in a zero). Empty cells were treated as a zero during analyses for all participants who reached the last page of the questionnaire. The following basic demographic information was also collected for each participant: gender, age, educational level and postal code. We also asked participants to report any influenza-like-illness (ILI) symptoms (provided in a list) that they and/or their household contacts experienced in the past seven days

Analyses

Degree was defined as the sum of the numbers of contacts while travelling and at different locations reported by each respondent for one day. We censored degree and number of contacts while eating to a maximum of respectively 500 and 75 contacts per day for each respondent, which were considered as highest reasonable values. We fitted a negative binomial distribution to the observed degree distribution using maximum likelihood estimation (see Text S1).

The tendency of individuals in a network to be linked to similar individuals ('assortative' mixing, and vice versa 'disassortative' mixing if links are made between dissimilar individuals) can be measured by correlation coefficients between pairs of individuals[33]. To investigate mixing patterns within our sample, we calculated correlations between the recruiter and his/her recruited contacts. We used Pearson's *r* for integer variables (e.g., age, degree, household size, contacts while eating, and number of self-reported symptoms). We used phi coefficient (r_{φ}) for binary variables (e.g., gender and two-or-more self-reported symptoms), and Spearman rank-order (r_{rank}) for ordinal variables (e.g., educational level). For conducting null hypothesis tests for Pearson's *r*, a bivariate normal distribution is assumed[34]. Count variables were therefore log transformed, and bivariate normality was visually assessed using joint probability distribution plots (see Text S2).

Besides studying the correlation between characteristics of neighbouring nodes, RDS allows for comparison of individuals' characteristics across more than one link in the network. Investigating such correlations shows whether correlations seen for pairs of directly linked nodes persist beyond the first link. We calculated correlations for all possible link distances between respondents in the same component.

In RDS, respondents recruit contacts they know, with the result that respondents have similar characteristics. Thus, characteristics of respondents in the same component are correlated, and this affects the standard errors of survey estimates. A measure for this correlation is the intraclass correlation (ICC), derived in multilevel analysis[35]. The ICC can be interpreted as the expected correlation between two randomly chosen individuals within the same component. The ICC for a two-level model is defined as:

$$\mathrm{ICC} = \frac{\text{population variance at component level}}{\text{total variance}} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2}$$

where the numerator is the variance at the component level $\boldsymbol{\sigma_{u0}^2}$, and the denominator represents the total variation in the model, which includes the variance at the component level and the variance at the respondent level $\boldsymbol{\sigma_{e0}^2}$. The ICC can be calculated for each variable separately and varies between 0-1. When the ICC is zero, observations can be considered independent ($\boldsymbol{\sigma_{u0}^2}$ = 0). An ICC of 1 indicates that respondents in the same component respond identically ($\boldsymbol{\sigma_{e0}^2}$ = 0). The higher the ICC, the less representative the sample is for the population given the variable considered and the sample size. Variance estimates were derived from an intercept-only multilevel model with restricted maximum likelihood estimation[35].

In the supplementary materials we analysed which variables are important in the recruitment process using logistic regression analyses (see Text S3). Furthermore, we explored whether correlations for age, gender and education between recruiters and their contact persons were only dependent on the direct recruiter (i.e., one step away, a first-order Markov assumption), by using a Monte Carlo technique to simulate a first-order autoregressive process (see Text S4). We also visually assessed equilibrium for all variables and applied the Volz-Heckathorn estimator[36] to estimate population proportions from our sample (see Text S5).

Analyses were performed with R (version 2.5.3); Figure 1 was created with the package Rgraphviz. RDS data file is available online, doi: 10.6084/m9.figshare.860458. We are currently improving the user interface of the webRDS system, researchers interested in the survey system are welcome to contact the authors.

RESULTS

Study participants and recruitment waves

In total, we invited 191 students. Of those students, 89 (46.6%) entered the online survey using the personal link provided in the invitation email and 80 (41.9%) completed the online questionnaire, out of which 44 (23.0%) students invited contacts. Thirty-nine (20.4%) of these led to completion of at least one new questionnaire. The maximum number of recruitment waves was six. There were 15 (34.1%) network trees with two or more waves (Table S1). The two largest network trees consisted in total of 30 and 34 participants (Figure 1).

In total 257 individuals entered the online survey, of which 220 (85.6%) completed the survey. The age range of respondents was 14-52 years (mean age 26.7 years; standard deviation [SD] 6.73). There were 157 (61.6%) female and 98 (38.4%) male participants. Regarding education, 160 (63%) respondents had a bachelor degree, 53 (21.0%) had a master degree, and 41 (16.0%) had a junior high school, high school or higher diploma. When plotting the sample proportion by increasing sample size, we observed that the sample composition regarding age, gender and education did not stabilise (see Text S5). Most respondents (122, 47.5%) provided a postal code of a district outside Bangkok, 95 (37.0%) indicated to live in Bangkok, and 40 (15.5%) did not provide their postal code (see Table 1).

Two invited contacts indicated that they did not participate in our survey as they had participated earlier as a seed. Repeated recruitment among contacts occurred for an additional seven times (e.g., individual first invited by seed and later also invited by friend from another tree), based on the number of duplicates in email addresses used for recruitment. One individual indicated not to be interested in the subject of the survey. Two individuals entered the questionnaire but did not provide any information.

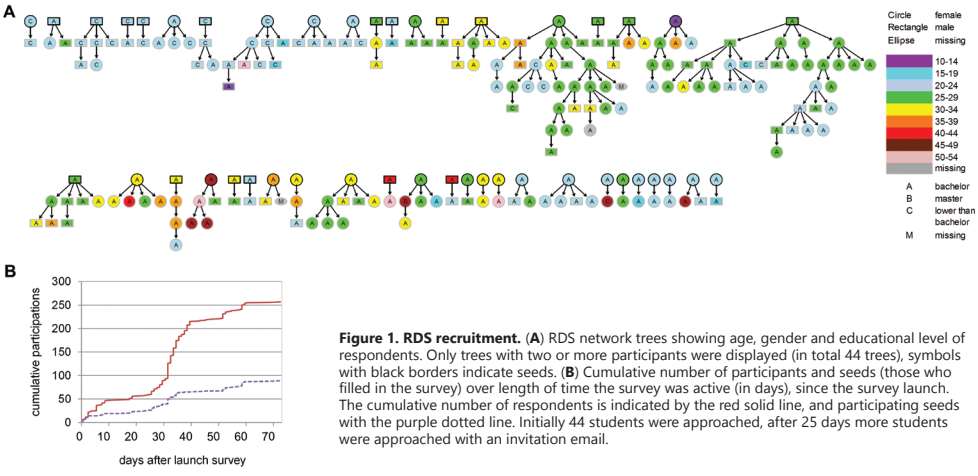


Figure 1. RDS recruitment. (**A**) RDS network trees showing age, gender and educational level of respondents. Only trees with two or more participants were displayed (in total 44 trees), symbols with black borders indicate seeds. (**B**) Cumulative number of participants and seeds (those who filled in the survey) over length of time the survey was active (in days), since the survey launch. The cumulative number of respondents is indicated by the red solid line, and participating seeds with the purple dotted line. Initially 44 students were approached, after 25 days more students were approached with an invitation email.

Table 1. Number of recorded contacts by different characteristics.

Category	Covariate	Number of participants	Mean (median;SD) number of reported contacts while travelling	Mean (median;SD) number of reported contacts at different locations
Age of participant	14-19	10 (3.9%)	61.4 (5.0; 99.8)	212.7 (144.0; 200.6)
	20-29	180 (70.0%)	17.3 (8.0; 33.0)	36.8 (17.0; 58.3)
	30-39	50 (19.5%)	16.4 (5.0; 8.3)	50.2 (15.5; 97.1)
	40+	13 (5.2%)	24.8 (2.0; 68.3)	42.3 (15.0; 69.0)
	Missing value	4 (1.6%)		
Gender participant	Female	157 (61.1%)	19.4 (7.0; 41.9)	42.0 (16.5; 75.8)
	Male	98 (38.1%)	18.3 (5.0; 38.5)	51.1 (18.0; 88.2)
	Missing value	2 (0.8%)		
Educational level participant	Master degree or higher	53 (20.6%)	16.7 (5.0; 42.3)	37.0 (15.5; 53.9)
	Bachelor degree	160 (62.3%)	20.7 (5.0; 44.8)	48.3 (15.5; 92.2)
	Lower than bachelor degree	41 (16.0%)	14.8 (11.0; 13.3)	45.2 (33.0; 63.1)
	Missing value	3 (1.2%)		
Household size	1	45 (17.5%)	12.6 (7.0; 22.3)	27.9 (14.0; 44.0)
	2	21 (8.2%)	13.9 (9.0; 16.6)	20.2 (14.0; 16.9)
	3	36 (14.0%)	13.6 (7.5; 23.4)	28.3 (14.5; 37.7)
	4	45 (17.5%)	13.9 (5.0; 47.7)	34.5 (14.0; 66.9)
	5	32 (12.5%)	20.4 (7.0; 33.5)	36.5 (21.0; 36.8)
	6	15 (5.8%)	11.9 (9.0; 17.2)	47.2 (27.0; 52.4)
	7+	27 (10.5%)	46.6 (9.0; 77.2)	95.9 (28.5; 115.8)
	Missing value	36 (14.0%)		
Participant living in Bangkok[a]	Yes	95 (37.0%)	19.4 (8.0; 40.1)	38.6 (16.0; 66.5)
	No	122 (47.5%)	18.1 (5.0; 40.8)	38.9 (17.5; 58.5)
	Missing value	40 (15.5%)		

[a]Most seeds provided a postal code from a district far away from Bangkok, however we assume that most of these students stayed in a student dorm in Bangkok during the study week.

Recruitment options used

Facebook was the most frequently used option for recruiting contacts. Facebook was used by 116 (45.1%) respondents compared to 29 (11.3%) who used one of the two email options. A total number of 580 coupons were handed out. Of those, 168 (29.0%) actually entered the survey and 140 (24.1%) completed the questionnaire. This means that we obtained 140 pairs of linked individuals who both completed the survey. Of the successfully recruited contacts, 117 (83.6%) were invited by 63 respondents who used Facebook, compared to 23 (16.4%) contacts who received an email from their recruiter (Table 2). Seventy-five (29.1%) respondents did not use a recruitment option after finishing the questionnaire. See Table S1 for a detailed overview of the used recruitment options.

Table 2. Number of successful recruitments by recruitment option.

Recruitment option used	0 ('no')	1	2	3	4	Total successful recruitments (n = 140)
Facebook	53	35	10	10	8	117 (83.6%)
Indirectly email	8	5	1	0	0	7 (5.0%)
Direct email	7	3	3	1	1	16 (11.4%)

Successful recruitment (of 1 to max. 4 contacts) counts when the invited contact also completed the survey; 0 ('no') indicates that recruiter invited his/her contacts but these contacts did not complete the survey. 75 respondents did not invite anyone after filling the survey.

Technical issues

Although the invitation emails were evaluated for spam content, a number of email providers blocked our emails. This severely affected participation rates in the first phase of the pilot. Furthermore, as personal codes could only be used once for completing the survey, participants could not return to the survey after closing the browser. Therefore, respondents who postponed recruitment were unable to recruit contacts at a later stage.

Facebook made it easier for respondents to recruit contacts, as they did not have to provide their email addresses. However, the Facebook 'Send Dialog' application (for sending private messages to friends) that was used is not supported for mobile devices. Therefore, the Facebook invitation option was either invisible when using a mobile device, or in some operating systems, the option was provided but gave an error when selected. With the Send Dialog the length of the standard invitation text that can be provided by the researcher is restricted, and in some occasions only partly displayed depending on the settings of the recipient. In addition, we were unable to send out reminders to contacts that had been invited via the online social network. The latter was possible for the two email recruitment options, but errors were sometimes made with filling in email addresses.

Number of reported contacts

A total of 19501 contacts were recorded, ranging from 0 to 4456 contacts per respondent per day. Three respondents reported more than 500 contacts. Figure 2 displays the degree distribution and the best fitting negative binomial distribution (mean=88.2; dispersion k=0.57). Less contacts per respondent were reported while travelling (the sum of contacts while travelling with different transport vehicles) than for locations (the sum of contacts for all locations together), respectively 19.2 (median=6, SD=40.8) and 45.1 (median=17, SD=80.8) mean contacts per day. Of all recorded contacts, 7574 (55.0%) were made during weekdays, compared to 6205 (45.0%) contacts that were made during the weekend (degree was censored to a maximum of 500 contacts per respondent). On average a degree of 62.3 (median=25, SD=100.7) was reported per participant per day, with Friday and Saturday having the highest degree per person (median 46.5 and 34.0 respectively) and Sunday the lowest (median of 20.5)

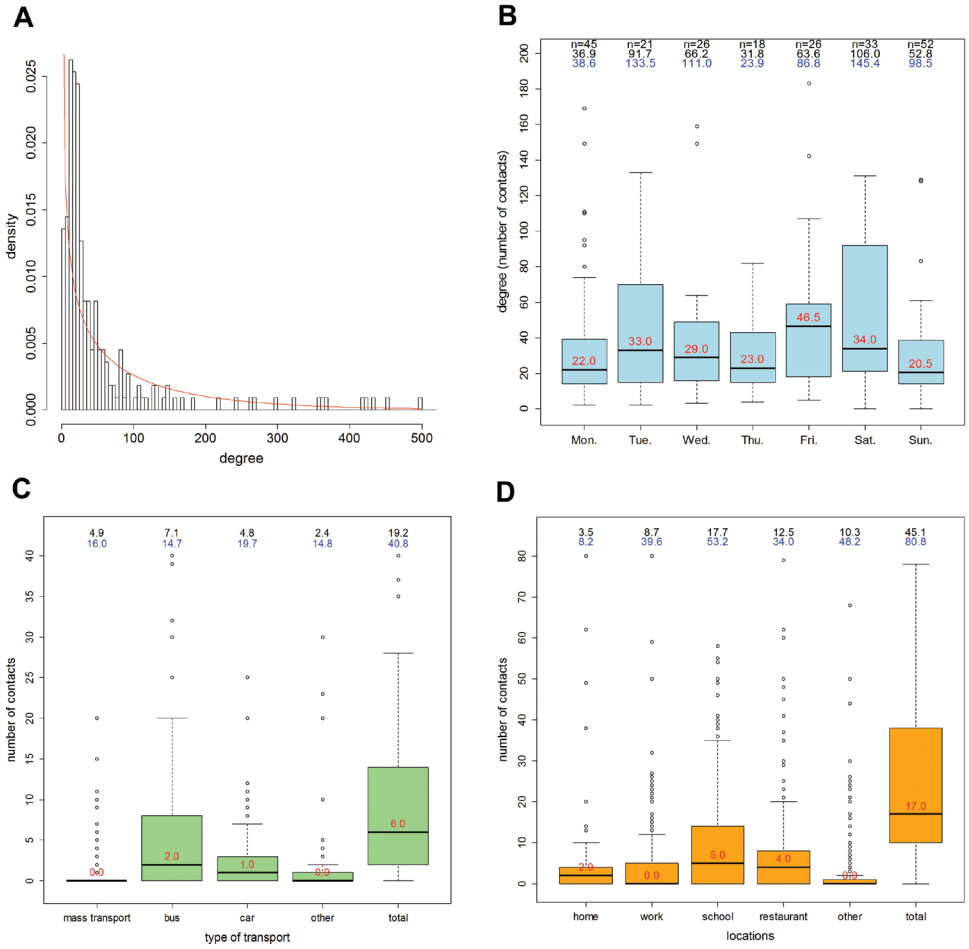


Figure 2. Recorded contacts. (**A**) Distribution of reported degree, the line indicates the fitted negative binomial distribution; (**B**) degree by day of the week (outliers >200 are not shown); (**C**) contacts while travelling with mass transport (sky train, subway and/or airplane), bus/minibus/shuttle boot, car/taxi and/or motorbike/tuk-tuk (outliers >40 are not shown); (**D**) numbers of contacts at different locations (outliers >80 are not shown). School was defined as 'school/university', 'restaurant' includes contacts at coffee shop, and 'other' is the sum of contacts encountered at sport/ leisure, concert and 'other places'. Above each plot in B, C and D the mean and SD (in blue) are displayed and within each plot the median (in red); B also contains the number of observations. Plots in C and D are based on an equal number of observations (n=221).

Contacts while travelling and at different locations

While travelling (Figure 2), on average most contacts were made in the bus, mini bus or shuttle boat (mean of 7.1 contacts per person). Remarkably, the mean number of contacts per person for car or taxi was not much lower (4.8). For mass transport (e.g., sky train, subway or airplane) the number of reported contacts per person was over-dispersed. Comparable patterns were seen for motorbike/tuk-tuk, which is probably due to the larger versions of tuk-tuks (which can carry >3 persons). Twenty-six (11.8%) respondents indicated that they did not use any transport vehicles on the specific recording day.

Figure 2d shows the average number of contacts reported per participant for different locations. Participants in the age class 14-19 reported the highest numbers of contacts for each location, except for work. Most contacts in this age class were reported for school/ university and 'other' (e.g., sports/leisure, concert and other places) locations. For the oldest age class (40+), most contacts were made at work. Table 1 shows the number of contacts while travelling and at different locations for different respondents' characteristics. Regarding contact numbers while eating, most contacts were reported during lunch and dinner. Again, the age class 14-19 reported overall the highest averages (which is an average of the sum of contacts during breakfast, lunch, dinner and/or snack break), and the age classes 20-29 and 30-39 had more varied numbers of contacts (see Table S2).

Table 3. Correlations between directly linked individuals in one component.

	Correlation	df	p value
Age (r)	0.555 [0.439 – 0.652]	163	<0.001
Gender (r_{φ})	0.205 [0.054 – 0.346]	164	0.008
Education (r_{rank})	0.520 [0.383 – 0.653]	164	<0.001
Degree log (r)	0.010 [-0.156 – 0.176]	138	0.907
Household size log (r)	0.058 [-0.110 – 0.222]	137[a]	0.499
Food with log (r)	0.215 [0.051 – 0.368]	138	0.011
Number of reported symptoms[b] log (r)	-0.095 [-0.257 – 0.072]	138	0.266
Two or more reported symptoms[b] (r_{φ})	-0.060 [-0.223 – 0.108]	138	0.488

[a]One case who reported >500 household contacts was removed. [b]Based on total number of self-reported symptoms by each respondent.

Link recruiter and recruited

We examined correlations for characteristics and numbers of contacts between linked respondents (recruiter versus recruited contact) (Table 3). Strong correlations were found for age (r=0.555, p<0.001), gender (r_{φ}=0.205, p=0.008) and education (r_{rank}=0.520, p<0.001). These positive correlations indicate that recruitment is assortative by age (Figure 3a), gender and educational level.

No assortativeness was observed by degree between linked nodes. Figure 3b shows the degree of the recruiter versus the degree of the recruited contact. The distribution corresponds with a random mixing in a population with negative binomially distributed degree (Figure 2a). The summary statistic between pairs of linked nodes indicated random mixing by degree (r=0.010, p=0.907). Comparable correlations were seen for household sizes of pairs of respondents (r=0.058, p=0.499). For numbers of contacts while eating a weak assortative tendency was seen (r=0.215, p=0.011). See Text S2 in supplementary materials for scatterplots.

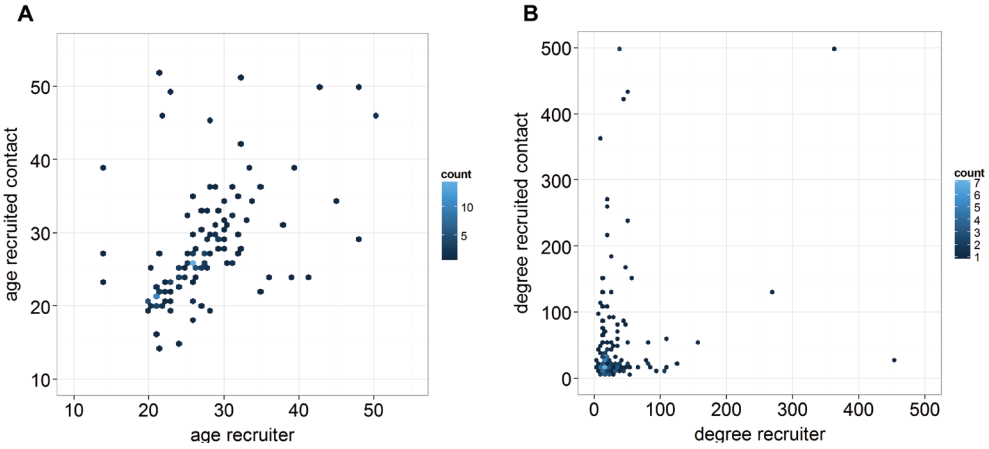


Figure 3. Correlations between recruiter and recruited contact. Graphs display correlations for (**A**) age, (**B**) degree (untransformed). Overlapping points were made visible with a colour scale.

Successive links between contact persons in the same component

Figure 4 shows the correlations between individuals across several steps in the recruitment chain. For age, gender and education the positive correlations decrease after the first link and disappear after a distance of three or more steps between any two individuals in the same chain (Figure 4a). For degree, numbers of contacts while having food, and household size, no correlations were observed over all distances (Figure 4b). Comparable analyses can be performed based on recruitment waves (i.e., only forward steps, see Text S4).

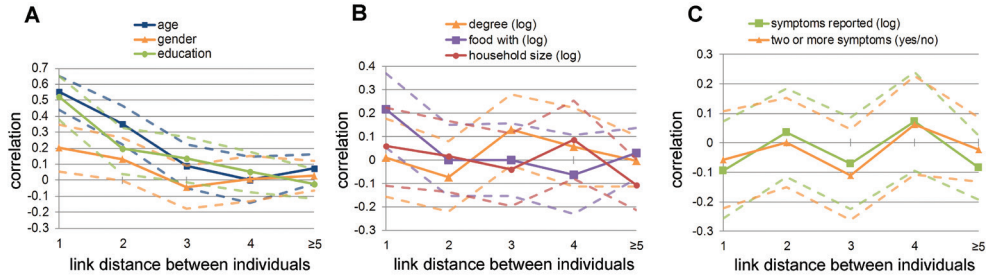


Figure 4. Correlations between any two individuals with different link distance. With graph **A** showing the correlations for age, gender, and education. Graph **B** displays the correlations in degree, number of contacts while having food, and household size (all after log transformation). Graph **C** shows the correlations for total number of self-reported symptoms (after log transformation), and for two or more self-reported symptoms (yes/no). Distances of five or more links were lumped together. The dotted lines show the confidence intervals.

Intraclass correlations

The ICC values for age, gender and education (respectively 0.470, 0.203 and 0.435) suggest a relatively large homogeneity for these characteristics within network components (Table 4). By contrast, low ICC values were found for log degree (0.064), numbers of contact while eating

(0.142), household size (0), numbers of self-reported symptoms (0), and two or more reported symptoms (0). This confirms the results obtained through the correlations reported above in which we did not explicitly take interdependence of observations in components into account.

Table 4. Intraclass correlation coefficients.

	ICC	Variance between components (σ^2_{u0})	Variance within components (σ^2_{e0})
Age	0.470	22.848	25.812
Gender	0.203	0.049	0.203
Education	0.435	0.391	0.509
Degree (log)	0.064	0.087	1.281
Food with (log)	0.142	0.150	0.910
Household size (log)	0	0.000	0.511
Number of reported symptoms (log)	0	0.000	0.659
Two or more reported symptoms (yes/no)	0	0.000	0.210

Self-reported symptoms

The mean number of self-reported symptoms was 1.8 (median=2, SD=1.5; varying from 0 to 8 symptoms per person). Of the respondents who completed the survey, 66 (30.0%) reported two-or-more symptoms, of which 7 (3.2%) persons reported a set that indicates flu-like symptoms (e.g., a combination of the symptoms fever, headache and muscle pain), and 13 (5.9%) reported common cold-like symptoms (e.g., a combination of the symptoms runny nose, sore throat and cough). Seventy-four (33.6%) respondents reported one-or-more household contacts with symptoms in the same period (Table S3). Random mixing was seen by total number of self-reported symptoms and by the variable two-or-more self-reported symptoms at all distances (Table 3 and Figure 4c).

DISCUSSION

We presented the results of a pilot study demonstrating the applicability of webRDS for studying contact patterns, especially those relevant to the transmission of airborne infections, of socially connected individuals. To our knowledge, this is the first online study in which RDS was applied to collect data specifically on human contact patterns relevant for close-contact pathogens, and the first in Thailand in which contact data was collected via the internet. By studying the correlations in characteristics between recruiter and recruited contacts, we found assortative mixing by age, gender and education, and random mixing by numbers of contacts. Building on earlier work in Vietnam[27], we have shown that webRDS can be used to elicit information about social contact networks and to collect empirical data on mixing patterns

that are relevant for communicable disease transmission.

In contrast to most previous contact studies that focused on egocentric data[5,8-11,37], we used RDS to include contacts and contacts of contacts of the initial respondents into the study. This sampling method allows researchers to collect contact data in connected components of respondents, i.e., within the social structures where transmission actually occurs. In particular, mixing patterns and heterogeneity in numbers of contacts in social networks can be studied directly thereby providing information on possible transmission routes of communicable diseases. Assortative mixing by, for instance, age affects the spread of infections through a community. When individuals of a similar age class primarily have contact with each other, infections are likely to spread more within those subgroups. On a similar note, data on the distribution of degree within a community provides information on individuals who have many contacts and are therefore more likely to become infected and to infect others than individuals with fewer contacts. Theoretically it has been shown that highly over-dispersed degree distributions strongly affects the basic reproduction number, which is an important indicator of how fast an infection spreads and what fraction of the population will be infected[38]. Such RDS collected information on the contact network can be used as input for mathematical models to better describe transmission dynamics and impact of public health interventions, such as vaccination or isolation of certain groups within the population.

In contrast to previous webRDS studies[26,27,39], we have demonstrated that sampling can be performed without material incentives. Our non-material incentive was considered a fun motivator for recruiting contacts. Although the use of a monetary reward or a combination of incentives as was applied by Bengtsson et al. 2012[27], could have increased the number of waves in our study, applying material incentives generally increases the risk of attracting 'cheaters' (e.g., respondents that participate multiple times by recruiting themselves to receive multiple rewards) that can severely affect study validity. In general, surveys in Thailand are performed without incentives as contributing to research is culturally considered an activity that does not require a monetary reward.

Our results underpin the importance of recruiting motivated and well-informed seeds, as was also seen in previous RDS studies. The use of a Facebook option for recruitment of friends has shown to be of critical importance for recruitment of contacts in the Thai study population. By providing recruiters with four private Facebook messages or invitation emails, we facilitated recruitment and gained more control over the sampling process, despite some technical challenges such as that the Facebook Send Dialog is not supported on mobile devices. Recruitment was more personal and directed, compared to the sharing of survey links in public areas as was done by Bauermeister et al. 2012[39].

However, our results are based on a rather small sample and mainly represent students and their contacts of similar age (e.g., the age class of 20-35, see also Text S5). Several factors might have restricted recruitment and consequently the penetration of our survey into different layers of the Thai population, for example limited internet access, unfamiliarity with online recruitment of friends, and perceiving an invitation mail as spam. In addition, the criterion of recruiting only contacts that were seen in the past seven days may have contributed to limit sample size and composition. In case students and their contacts did not travel back to their parental home during the weekend, they were not likely to meet many other contacts outside the university and / or work.

In contact diary studies, differences in contact definition cause heterogeneities in numbers of recorded contacts and influence the assessment of the importance of settings in disease transmission risk[32]. Although the applied contact definition in our study was fairly simple and not limited to physical contact, for events like travelling by metro or bus during rush hours it is difficult to estimate the number of persons within arm's length. For these events reported numbers of contacts may be less reliable. For contacts that occur repeatedly within stable relationships such as within households, schools and workplaces, repeated measures on different days and asking recruited contacts about the contact they had with their recruiters (and vice versa) will provide insight in reporting bias and the validity of the data[8]. Moreover, repeated measures would aid participants in recalling daily contacts.

RDS is by nature subjected to clustering due to its sampling process. Clustering refers to how many of an individual's contacts, and subsequently how many of the contacts reported by individuals in the same referral chain, also had contact among each other. Clustering can have a profound effect on disease spread[16,40-42]. For example, high clustering of contacts means more local spread and thus a rapid local depletion of susceptible individuals. In future surveys more information should be collected on repetitive recruitment among participants. Such information can be used to make estimates of clustering. Repeated recruitment could possibly be measured by asking respondents to report their personal code, with which they were initially invited to the survey, the moment they receive a second invitation (or more).

While RDS population estimates were not the aim of this study, we were interested in obtaining samples of respondent-contact pairs. Typically with RDS, well-connected individuals (i.e., high degree individuals) tend to be over-sampled because many recruitment paths lead to them. The bias that is introduced, with respect to number of contacts, is corrected for by the RDS estimators when making inferences to the population[43]. Increasing the number of coupons could provide a better view on an individual's entire contact structure, although it would increase the participation burden (e.g., more contacts have to be approached) and increase the probability of multiple recruitment.

In addition, with RDS respondents tend to recruit contacts who they think will participate, making the peer recruitment anything but random. For studying the representativeness of network links, more information is required about the (non-randomly) chosen contacts. For example, during future research recruited contacts can be asked to specify the relation he/she has with the recruiter, and to indicate the type, frequency and duration of contact, to learn more about the contact persons that are included in the sample. In addition, by defining more specific recruitment criteria link sampling could become more controlled in order to counter bias.

In future research, recruitment of seeds could be organised through online communities (e.g., panels[44] or online social networks), to capture a variety of seeds from all levels of the population. In general, participants from online panels are used to fill in web surveys, which will help researchers in the search for generative seeds. In addition, selecting seeds through the web might result in the inclusion of seeds from different geographical locations, which decreases the clustering among contacts. However, this requires that seeds can be motivated for peer-recruitment without researchers having to physically meet with these individuals. If possible, it will then also be interesting to explore the feasibility of using a probability based sample of seeds, i.e. selecting seeds randomly, thereby providing every individual in the population a chance of being selected as a seed. Representatively selected seeds for webRDS would retain the benefits of a random sample, such as collected in earlier egocentric studies, but are also likely to reach a more representative sample of the contact networks in a population. In theory, longer recruitment chains ensure that the sociometric distance between the seeds and the bulk of the sample will be large, hereby enhancing the diversity and representativeness of the sample[45]. Furthermore, the use of mobile devices for recruitment of contacts should also be further explored. Communication through smartphones or other mobile devices will continue to grow in Thailand and elsewhere providing new opportunities for webRDS research.

In principle, with webRDS it is possible to recruit a sample relatively fast compared to offline RDS[46] or traditional sampling techniques. In our pilot study, recruitment of additional seeds after day 25 led to a three times higher number of recruited participants (from around 50 participants to a sample size of over 200, see Figure 1b) within only 15 days. Although the pilot sample size was too small to include a large number of (linked) individuals with influenza-like-illness symptoms, the sampling speed of webRDS and the assortative recruitment by age, gender and education is potentially useful for reaching contacts at risk for infection, and for detecting clusters and studying the spread of respiratory agents in social networks at the level at which it is actually occurring. For example, webRDS could be applied for case-contact tracing where reported cases act as seeds, who are then asked to recruit contacts that they have physically seen during their infectious period. The benefit is that respondents are in control of recruitment, instead of health authorities, which is efficient as individuals know with

whom they had contact and they can approach their contacts directly. In addition, with the use of the internet the tracing of contacts will be accelerated (compared to traditional methods) that will save time, human resources, and possibly provides the health authorities with options to intervene earlier during an outbreak (e.g., applying of interventions to control the spread) thereby preventing new cases.

Despite the methodological challenges, RDS allowed us to study connected components of individuals and obtain information about links within the network. Such information increases our understanding of contact networks relevant for the transmission of respiratory infections and can be used to improve existing modelling approaches. The application of webRDS for the purpose of studying contact patterns within real-life network structures is promising and will be explored further in future studies.

Acknowledgements

The study was conducted within the Utrecht Center for Infection Dynamics. We thank the deputy deans of student affairs of the Faculty of Public Health of the Mahidol University in Thailand for their assistance with contacting students. Furthermore, we thank Albert Wong from the Department of Statistics of the National Institute for Public Health and the Environment (RIVM) in the Netherlands for his help with the programming in R, and Aura Timen of the Centre for Infectious Disease Control (RIVM) for fruitful discussions on the study design. We also thank Titan Tang and Adam Ju from Qianyuan Software Co. China for programming the RDS survey system. Finally, we are grateful to Richard Coker and his London School of Hygiene and Tropical Medicine team in Bangkok for their input on the study design and providing working facilities during data collection.

REFERENCES

1. 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.

2. 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 Biol Sci 264: 949-957.

3. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M (2008) Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol Infect 136: 813-822.

4. Kretzschmar M, Mikolajczyk RT (2009) Contact profiles in eight European countries and implications for modelling the spread of airborne infectious diseases. PLoS One 4: e5931.

5. 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: e74.

6. 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.

7. McCaw JM, Forbes K, Nathan PM, Pattison PE, Robins GL, et al. (2010) Comparison of three methods for ascertainment of contact information relevant to respiratory pathogen transmission in encounter networks. BMC Infect Dis 10: 166.

8. Smieszek T, Burri EU, Scherzinger R, Scholz RW (2012) Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiol Infect 140: 744-752.

9. Read JM, Eames KT, Edmunds WJ (2008) Dynamic social networks and the implications for the spread of infectious disease. J R Soc Interface 5: 1001-1007.

10. Stehle J, Voirin N, Barrat A, Cattuto C, Colizza V, et al. (2011) Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. BMC Med 9: 87.

11. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 107: 22020-22025.

12. Salathe M, Jones JH (2010) Dynamics and control of diseases in networks with community structure. PLoS Comput Biol 6: e1000736.

13. Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: homogeneous and network models in epidemiology. J R Soc Interface 4: 879-891.

14. Miller JC (2009) Spread of infectious disease through clustered populations. J R Soc Interface 6: 1121-1134.

15. Ball F, Britton T, Sirl D (2013) A network with tunable clustering, degree correlation and degree distribution, and an epidemic thereon. J Math Biol 66: 979-1019.

16. Volz EM, Miller JC, Galvani A, Ancel Meyers L (2011) Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol 7: e1002042.

17. Heckathorn D (1997) Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems 44: 174-199.

18. McCreesh N, Frost SD, Seeley J, Katongole J, Tarsh MN, et al. (2012) Evaluation of respondent-driven sampling. Epidemiology 23: 138-147.

19. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, et al. (2011) Evaluation of the role of location and distance in recruitment in respondent-driven sampling. Int J Health Geogr 10: 56.

20. Wejnert C (2009) An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and out-of-Equilibrium Data. Sociol Methodol 39: 73-116.

21. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, et al. (2012) Digital epidemiology. PLoS Comput Biol 8: e1002616.

22. Brooks-Pollock E, Tilston N, Edmunds WJ, Eames KT (2011) Using an online survey of healthcare-seeking behaviour to estimate the magnitude and severity of the 2009 H1N1v influenza epidemic in England. BMC Infect Dis 11: 68.

23. Dugas AF, Hsieh YH, Levin SR, Pines JM, Mareiniss DP, et al. (2012) Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. Clin Infect Dis 54: 463-469.

24. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014.

25. Eames KT, Tilston NL, Brooks-Pollock E, Edmunds WJ (2012) Measured dynamic social contact patterns explain the spread of H1N1v influenza. PLoS Comput Biol 8: e1002425.

26. Wejnert C, Heckathorn DD (2008) Web-Based Network Sampling: Efficiency and Efficancy of Respondent-Driven Sampling for Online Research. Sociological Methods and Research 37: 105–134.

27. Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, et al. (2012) Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. PLoS One 7: e49417.

28. van der Weijden CP, Stein ML, Jacobi AJ, Kretzschmar ME, Reintjes R, et al. (2013) Choosing pandemic parameters for pandemic preparedness planning: a comparison of pandemic scenarios prior to and following the influenza A(H1N1) 2009 pandemic. Health Policy 109: 52-62.

29. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, et al. (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437: 209-214.

30. National Statistical Office, Ministry of Information and Communication Technology Thailand (2012) The Information and Communication Technology Survey in Household.

31. Socialbakers (2013) Checkfacebook: Facebook statistics of Thailand.

32. Bolton KJ, McCaw JM, Forbes K, Nathan P, Robins G, et al. (2012) Influence of contact definitions in assessment of the relative importance of social settings in disease transmission risk. PLoS One 7: e30893.

33. Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89: 208701.

34. Chen CY, Popovich PM (2002) Correlation: Parametric and Nonparametric Measures. Thousand Oaks, CA: Sage: Sage University Paper Series on Quantitative Applications in the Social Sciences.

35. Hox JJ (2010) Multilevel Analysis: Techniques and Applications, Second Edition: Routledge.

36. Volz E, Heckathorn DD (2008) Probability based estimation theory for respondent driven sampling. Journal of Official Statistics 24: 79-97.

37. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA (2012) Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect 140: 2117-2130.

38. May RM (2006) Network structure and the biology of populations. Trends Ecol Evol 21: 394-399.

39. Bauermeister JA, Zimmerman MA, Johns MM, Glowacki P, Stoddard S, et al. (2012) Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, Respondent-Driven Sampling (webRDS) strategy. J Stud Alcohol Drugs 73: 834-838.

40. Eames KT (2008) Modelling disease spread through random and regular contacts in clustered populations. Theor Popul Biol 73: 104-111.

41. Keeling MJ (1999) The effects of local spatial structure on epidemiological invasions. Proc Biol Sci 266: 859-867.

42. Szendroi B, Csanyi G (2004) Polynomial epidemics and clustering in contact networks. Proc Biol Sci 271 Suppl 5: S364-366.

43. Salganik MJ, Heckathorn D (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34: 193-239.

44. Marquet RL, Bartelds AI, van Noort SP, Koppeschaar CE, Paget J, et al. (2006) Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season. BMC Public Health 6: 242.

45. Heckathorn DD (2002) Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. Social Problems 49: 11-34.

46. Johnston LG, Whitehead S, Simic-Lawson M, Kendall C (2010) Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: lessons learned from four case studies. AIDS Care 22: 784–792.

SUPPLEMENTARY MATERIALS

Text S1. Fitting of a negative binomial distribution to degree.

Degree was defined as the sum of the numbers of contacts while travelling and at different locations reported by each respondent for one day (Figure Ia). We observed no assortativeness (i.e., random mixing) by degree between linked nodes (Figure Ib). We investigated whether the observed distribution agreed with a random mixing in a population with a negative binomially distributed degree, by fitting a theoretical distribution to the empirical degree distribution.

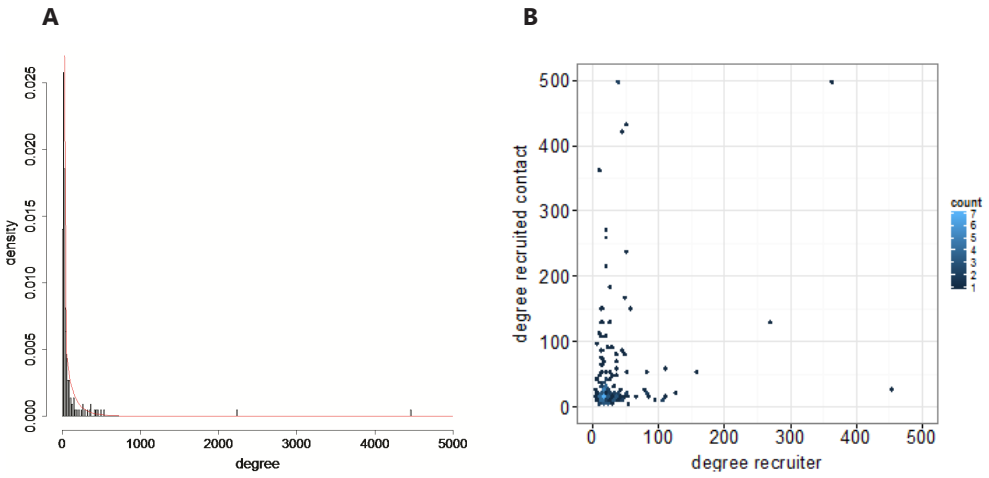


Figure I. Distribution of degree. A. Distribution of reported degree, the red line indicates the fitted negative binomial distribution. **B.** Correlations between degree recruiter and recruited contact. Overlapping points were made visible with a colour scale.

The count variable degree exhibited extreme overdispersion, i.e., the sample variance (117028.2) highly exceeded the mean (88.24), making the negative binomial (NB) distribution an appropriate model for fitting[1]. We fitted an NB distribution to the observed degree distribution using maximum likelihood (ML) estimation in R (version 2.5.3). In R, the mean number of counts (μ) in a sample is defined as μ, and the overdispersion parameter (k) is defined as size. The overdispersion parameter measures the amount of heterogeneity (or clustering) in the data. A larger k means more heterogeneity (i.e., as k becomes small the variance approaches the mean and the distribution approaches the Poisson distribution)[2].

ML estimates for degree: μ = 88.2; k = 0.57. We did not assess statistically the fit of the NB distribution to the observed degree distribution. As was shown earlier by Lloyd-Smith (2007), small-sample estimates of k can be biased towards underestimating k and

consequently to underestimation of the level of overdispersion in the data, when using maximum likelihood. Smaller samples are less likely to include values from the right-hand tail of the NB distribution, and without these outliers a dataset appears to be more homogeneous[3].

Literature

1. Bliss CI, Fisher RA (1953) Fitting the negative binomial distribution to biological data - note on the efficient fitting of the negative binomial. Biometrics 9: 176–200.
2. Bolker B (2007) Probability Distributions: Negative binomial. Ecological Models and Data in R. Princeton and Oxford: Princeton University Press.
3. Lloyd-Smith JO (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS One 2: e180.

Text S2. Links between recruiter and recruited contact person: descriptive statistics.
We studied the correlations between recruiter and his/her recruited contact for different characteristics. This supplementary text file provides an overview of the descriptive statistics from which the correlation coefficients shown in Tables 3 and 4 (see main article) were estimated. The scatterplots for the variables age and degree are displayed in the manuscript. Count data was log transformed for conducting null hypothesis tests.

Gender: Table I shows the absolute number of females/males who recruited other females/males. In total, 166 respondent-contact person pairs provided their gender. 57.8% was female and 42.2% male.

Table I. Who recruited whom? Gender

Gender recruiter	Gender recruited contact	
	female	male
female	71 (42.8%)	38 (22.9%)
male	25 (15.1%)	32 (19.3%)

Degree: Degree was calculated as the sum of the number contacts while travelling and at different locations (Figure I). Both the log distribution (Figure Ib) and the joint probability distribution of log degree visually approximated a normal distribution. There were three respondents who reported more than 500 contacts for the particular recording day (namely 533, 2233, 4456 contacts) that were censored to a maximum of 500 contacts.

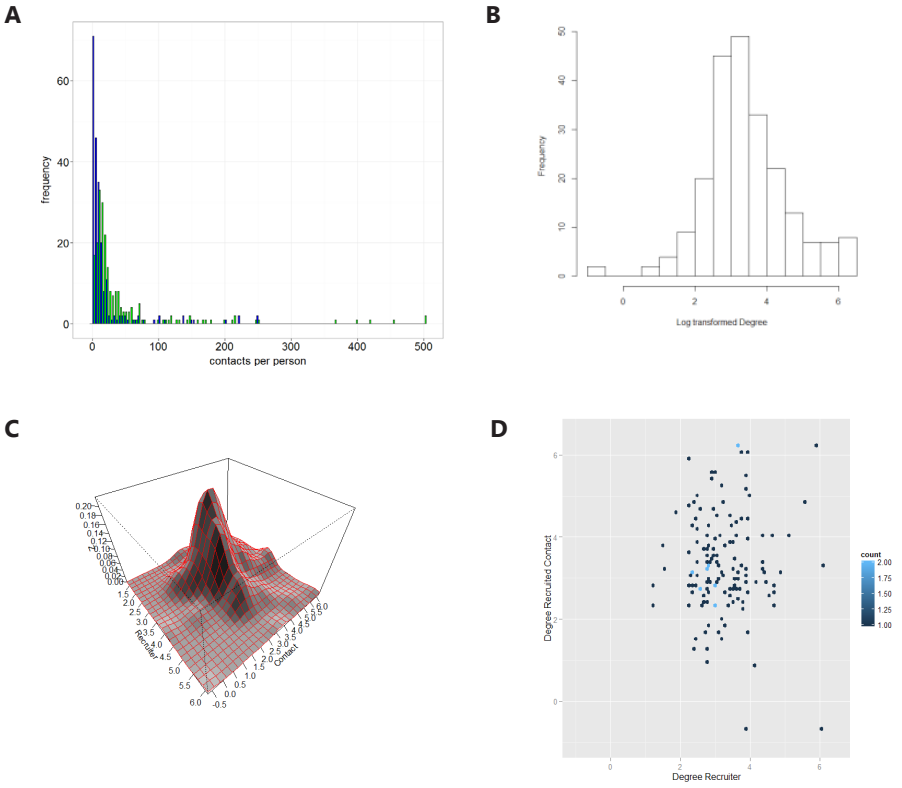


Figure I. Distributions of number of contacts in YourSpace. A. Reported numbers of contacts reported while travelling (in blue) and at different locations (in green), untransformed. **B.** Distribution of log degree. Degree is the sum of the two distributions displayed in graph A. **C.** The joint probability distribution of log degree (recruiter versus contact) approximated bivariate normality. **D.** Scatterplot of log degree recruiter versus log degree of his/her recruited contact, which shows random mixing by degree.

Household size: Household size was defined the number of persons living the household of the respondent in the seven days prior to the recording day. The log distribution of household size did not approximate a normal distribution (as shown in Figures IIa and IIb) due to a large number of respondents (>40) who reported only a one-person household. Figures IIc and IId showed random mixing by household size.

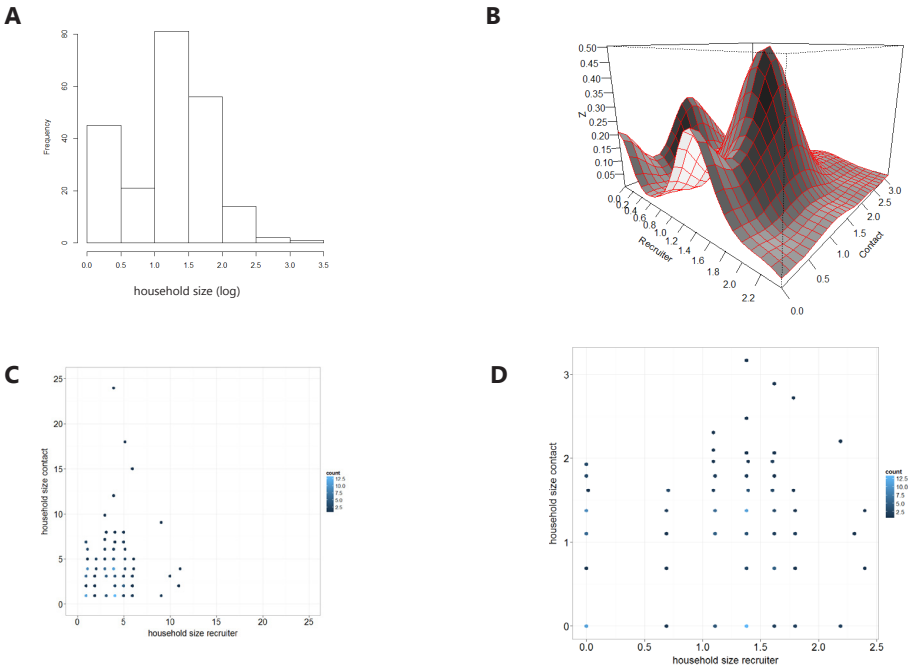


Figure II. Distribution of household size. A. Distribution of household size, log transformed. **B.** The joint probability distribution of log household size (recruiter versus contact). The distribution did not approximate bivariate normality due to >40 respondents who reported only one-person households. **C.** Scatterplot of household size recruiter versus household size of his/her recruited contact, untransformed. **D.** Scatterplot of log household size versus log household size of his/her recruited contact.

Contacts while eating: We asked participant to report the number of contacts (according to the contact definition) while having breakfast, lunch, dinner and snack break. There were four respondents who reported in total more than 75 contacts for the particular recording day (namely 76, 125, 376 and 1086 contacts). These were censored to a total maximum of 75 contacts while eating. The log distribution of contacts while eating approximated a normal distribution (see Figures IIIa and IIIb).

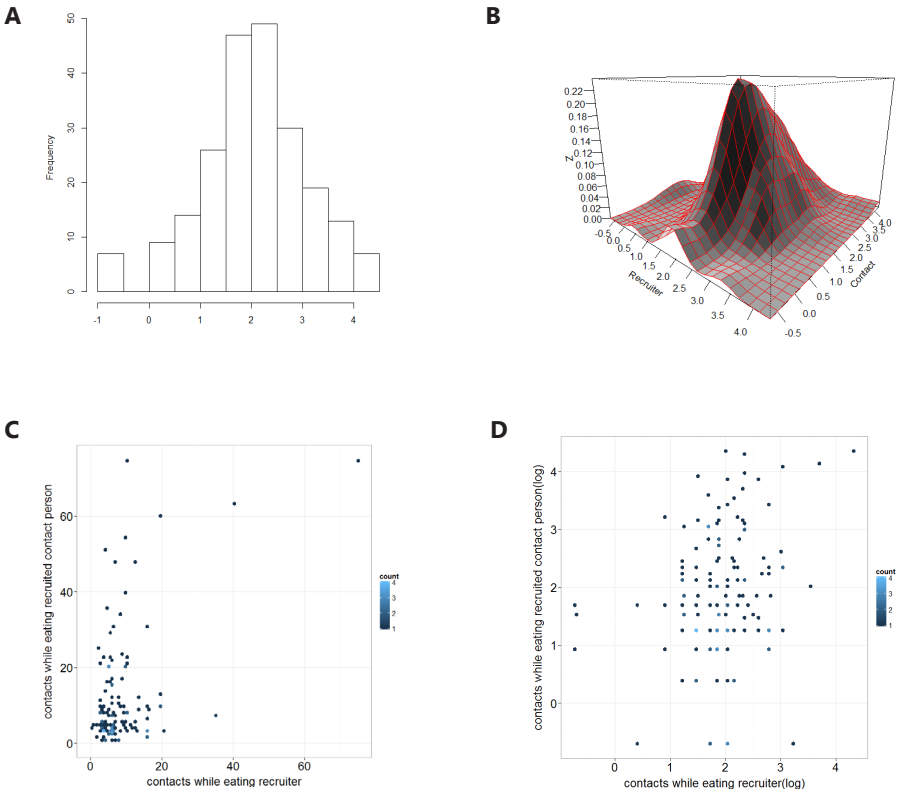


Figure III. Distribution of contacts while eating. A. Distribution of numbers of contacts while eating, log transformed. **B.** The joint probability distribution of log contacts while eating (recruiter versus contact). This distribution approximated bivariate normality. **C.** Scatterplot of numbers of contacts while eating of recruiter versus his/her recruited contact person, untransformed. **D.** Scatterplot of log numbers of contacts while eating of recruiter versus his/her recruited contact person.

Number of self-reported symptoms: Self-reported influenza-like-illness (ILI) symptoms were collected by providing respondents with a list of nine symptoms. For each respondent the number of symptoms reported was summed. The log distribution did not approximate a normal distribution due to the large proportion of respondents who reported no symptoms (Figures IVa and IVb). Table S3 contains the reported numbers for each symptom separately.

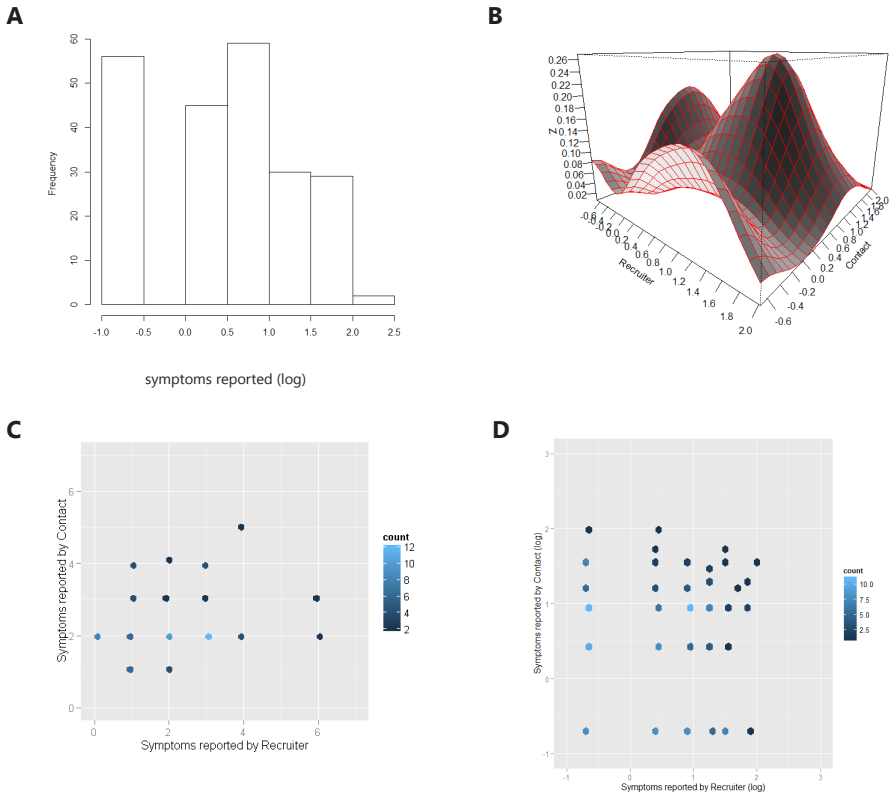


Figure IV. Distribution of self-reported symptoms. A. Distribution of numbers of reported symptoms, log transformed. **B.** The joint probability distribution of log reported symptoms. This distribution did not approximate bivariate normality. **C.** Scatterplot of numbers of reported symptoms by recruiter versus his/her recruited contact person, untransformed. **D.** Scatterplot of log numbers of reported symptoms by recruiter versus his/her recruited contact person.

Text S3. Drivers of the recruitment process.

To investigate which variables are important in the recruitment process, we analysed:
A. the intention to recruit (categorised as "respondent requested" OR "did not request" four invitations for contacts on the last survey page), and
B. the number of contact persons successfully recruited (categorised into "zero" OR "one or more contact persons") by each respondent, using logistic regression.

Confidence intervals for the coefficient estimates were obtained using standard errors. We also used a Pearson's Chi-Squared test (with Yates' continuity correction) to analyse bivariate the independency between the outcome and the categorical predictors. In addition, we investigated how well the model with all predictors fits compared to a null model (i.e., a model with just an intercept). The test statistic is the difference between the residual deviance for the model with predictor and the null model, and is distributed chi-squared with degrees of freedom (df) equal to the differences in df between the model with all predictors and the null model.

Note that our sample size is relatively small for the number of predictor variables that we are analysing, which could result in an unstable model. We produced contingency tables to check for empty cells or cells with a low number of observations (Tables I and IV). In addition, observations are not independently collected with a respondent-driven sampling method, thereby violating a condition for conducting logistic regression.

A. Intention to recruit contact persons

We performed a logistic regression with the intention to recruit as binary outcome, categorised as "a respondent did or did not request for four invitations on the last survey page" (n = 220). The log odds of the outcome was modeled as a linear combination of the variables: age (integer), gender (binary), education (ordinal), household size (integer) and degree (integer, i.e., number of contacts).

Table I. Contingency table of categorical outcome and predictor variables.

		Education		
Intention to recruit	**Gender**	Lower than bachelor	Bachelor	Master
No	Female	9	48	9
	Male	8	27	8
Yes	Female	10	55	24
	Male	14	29	12
	Total			**253**

Table II. Chi-squared test to analyse independence by outcome variable.

	Chi-squared	df	p-value
Gender	0.0053	1	0.9421
Education	3.7457	2	0.1537

Table II shows that the outcome "intention to recruit" is independent of gender and education. The logistic regression coefficients in Table III give the change in log odds of the outcome for one unit increase in the predictor variable. None of the included variables have a statistically significant effect on the outcome, which indicates that none of these variables significantly increase (or decrease) the intention to recruit contact persons by a respondent. The Chi-Square of 8.64 with 6 df and an associated p-value of more than 0.05 (0.1950) showed that the model with predictors does not fit significantly better than the null model, which is in agreement with the output from Table III.

Table III. Output of logistic regression with as binary outcome 'intention to recruit' (A).

	Estimate[a]	SE	z value	Pr(>|z|)	2.5%	97.5%
Constant	0.2414	0.7318	0.3298	0.7415	-1.1929	1.6757
Age	0.0263	0.0261	1.0103	0.3123	-0.0248	0.0774
Male	0.1166	0.3075	0.3791	0.7046	-0.4862	0.7193
Edu Bachelor	-0.6747	0.4534	-1.4881	0.1367	-1.5634	0.2140
Edu Master	-0.0937	0.5792	-0.1618	0.8714	-1.2289	1.0414
Household size	0.0632	0.0506	1.2491	0.2116	-0.0360	0.1623
Degree (number of contacts)	-0.0022	0.0015	-1.4696	0.1417	-0.0051	0.0007

[a]Null deviance: 283.62 (df: 219); residual variance: 274.98 (df: 213) and AIC: 288.98.

B. Number of contact persons successfully recruited

The number of contact persons successfully recruited by each respondent, i.e., contact person also completed the questionnaire, was divided into "zero" or "one or more contact persons" (n = 144). The log odds of this binary outcome (yes/no recruited contact persons) was modelled as a linear combination of the variables: age, gender, education, household size, degree and recruitment option used (binary: Facebook or email).

Table IV. Contingency table of categorical outcome and predictor variables.

Recruited	Gender	Education	Recruitment option used	
			Facebook	Email
No	Female	Lower than bachelor	2	2
		Bachelor	20	6
		Master	7	4
	Male	Lower than bachelor	6	0
		Bachelor	14	2
		Master	4	0
Yes	Female	Lower than bachelor	6	0
		Bachelor	25	4
		Master	9	4
	Male	Lower than bachelor	7	1
		Bachelor	9	4
		Master	7	1
		Total		**144**

Table V. Chi-squared test to analyse independence by outcome variable

	Chi-squared	df	p-value
Gender	0.0084	1	0.9271
Education	3.5934	2	0.1658
Recruitment option used	0.0397	1	0.842

Table V shows that the outcome variable "successfully recruited" is independent of gender, education and recruitment option. The logistic regression coefficients in Table VI give the change in log odds of the outcome for one unit increase in the predictor variable. Of all variables, only household size seems to significantly influence recruitment. For one unit increase in household size, the log odds of successfully recruiting contact persons decreases by -0.1527. However, running a logistic model that only contains household size as a predictor variable does not show a significant effect by this variable. In addition, the Chi-Square of 8.41 with 7 df and an associated p-value of more than 0.05 (0.2980), showed that the model with predictors does not fit significantly better than the null model. The significant influence by household size seen in Table VI is probably caused by the low number of observations (empty cells in the contingency table, see Table IV) that make the model unstable, or due to interference between included predictor variables.

Table VI Output of logistic regression with as binary outcome successfully recruited yes/no.

	Estimate[a]	SE	z value	Pr(>|z|)	2.5%	97.5%
Constant	1.6780	0.8664	1.9368	0.0528	-0.0201	3.3760
Age	-0.0316	0.0301	-1.0467	0.2952	-0.0906	0.0275
Male	-0.0925	0.3631	-0.2546	0.7990	-0.8041	0.6192
Edu Bachelor	-0.1954	0.5062	-0.3860	0.6995	-1.1876	0.7968
Edu Master	0.1946	0.6340	0.3070	0.7588	-1.0480	1.4373
Household size	-0.1527	0.0699	-2.1859	0.0288	-0.2897	-0.0158
Degree (number of contacts)	0.0012	0.0020	0.6109	0.5412	-0.0027	0.0052
Email	-0.0564	0.4536	-0.1243	0.9011	-0.9455	0.8327

[a]Null deviance: 198.93 (df: 143); residual variance: 190.52 (df: 136) and AIC: 206.52.

Text S4. Exploring the first–order Markov assumption.

A respondent-driven (i.e., chain referral) sample can be viewed as a stochastic process in which the social characteristics of each recruiter affect the characteristics of the recruited contact persons. If the sampling conforms to methodological requirements, the proportion of the sample with a certain characteristic is expected to stabilise at a level determined by the characteristics of the population and independent of the characteristics of the seeds[1]. The recruitment process can be modelled as a Markov process, in which the process can assume a limited number of states and is state dependent (i.e., the probability that the recruited contact person comes from a given group depends on the group from which the current recruiter comes). If the recruitment patterns depend only on the recruiter and not on the recruiter's recruiter (or recruiters before that) the recruitment chain corresponds to a first–order Markov process[2]. Although our sample contained only a limited number of waves, we explored whether the resulting recruitment chains are consistent with a first–order Markov process by comparing correlations for age, gender and education found in our sample data with correlations obtained from simulated data.

In the main text we showed the correlations between any two respondents with different link distances in the same network tree (see Table 3 and Figure 4), using Dijkstra's algorithm (which calculates shortest paths between any two persons in the same network tree). The geodesic distance between two nodes (or vertices) is the number of edges in a shortest path connecting them.

To analyse whether recruitment by recruiters in wave 1 and higher is dependent on the seed, we first calculated the correlations between seeds (wave 0) and their contact persons in consecutive waves (maximum up to 3 waves, due to limited number of respondents in the waves ≥4). We then used Monte Carlo techniques to simulate (with n=10.000) a first–order autoregressive process (i.e., an autoregressive Markov Chain of order one) based

on the correlation found in the sample data between seeds and their contact persons in wave 1. For this, we used the function 'arima.sim' of the R–package 'stats' (version 2.15.3). We compared correlations estimated from the sample data (*r*) with correlations estimated from the simulated data (*rr*) with the same geodesic distance.

Table I shows that the sample correlations (*r*) for age and gender, decreased slower than the simulated correlations (*rr*) over the same geodesic distances. For example, for age, the sample correlation between seeds in wave 0 and contact persons in wave 2 is $r = 0.554$, while $rr = 0.277$. This suggests that the recruitment process is a higher order process with regard to the variables age and gender. Thus, correlations found between the recruited contact persons and their recruiter are not only dependent on the direct recruiter, but also on recruiters in previous waves. Due to the limited number of waves in most of our network trees, we were unable to quantify the exact order level. For education, correlations for the geodesic distances 1 and 2 were lower in the data than in the simulations, suggesting a first-order Markov process (i.e., education of contact persons are only dependent on the education of their direct recruiter)

Table I. Correlations between seeds and contact persons in waves 1 to 3, estimated from sample data and simulated data.

		Wave 0	
	Geodesic distance	**Correlations sample data** *(r)*	**Correlations simulated data** (*rr*)
Age (*r*)	**Wave 1**	0.524 [0.350–0.663][a]	0.527 [0.512–0.541]
	Wave 2	0.554 [0.304–0.732][b]	0.277 [0.259–0.295]
	Wave 3	0.320 [-0.143–0.668][c]	0.150 [0.131–0.169]
Gender (r_φ)	**Wave 1**	0.192 [-0.022–0.389][a]	0.204 [0.185–0.223]
	Wave 2	0.172 [-0.135–0.450][b]	0.057 [0.037–0.077]
	Wave 3	0.000 [-0.443–0.443][c]	0.024 [0.004–0.044]
Education (r_{rank})	**Wave 1**	0.701 [0.559–0.811][a]	0.677 [0.664–0.688]
	Wave 2	0.224 [-0.09–0.519][b]	0.465 [0.448–0.481]
	Wave 3	0.000 [-0.167–0.192][c]	0.331 [0.314–0.348]

a) $n_{pairs} = 85$; **b)** $n_{pairs} = 43$; **c)** $n_{pairs} = 20$

References

1. Heckathorn DD (2002) Respondent–Driven Sampling II: Deriving Valid Population Estimates from Chain–Referral Samples of Hidden Populations. Social Problems 49: 11–34.
2. Heckathorn D (1997) Respondent–driven sampling: a new approach to the study of hidden populations. Social Problems 44: 174–199.

Text S5. Sample composition, equilibrium curves and RDS estimates.

To assess whether the sample reached equilibrium (i.e., as the sample grows in size, the composition of the sample ceases to change[1]), we plotted the sample composition over waves for gender, education and age (see Figure S1) and visually assessed whether the process reached equilibrium. The curves in Figure S1 indicate that equilibrium was reached for gender, education and age after a maximum of three waves. However, the achievement of equilibrium over waves for these variables is greatly influenced by the relatively small sample size and the limited number of waves in most network trees (most trees had a maximum of one or two waves).

We also plotted the sample proportion by increasing sample sizes for all variables surveyed in this study (see Figure S2). We used the Volz-Heckathorn (VH) estimator (or 'RDSII' estimator) to provide RDS corrected estimates, thus, the proportion of individuals with characteristic A in the population. The VH-estimator requires only information on the sample compositions and the personal network sizes of the participants. Degree (number of reported contacts) was used to define the participants' personal network size. The Volz-Heckathorn estimator is defined as[2]:

$$\hat{P}_A^{VH} = \frac{\sum_{i \in A \cap S} d_i^{-1}}{\sum_{i \in S} d_i^{-1}}$$

Where d_i is the degree of individual i, and S the set of sampled individuals.

For the variables age (Figure S2E), gender (Figure S2F), education (Figure S2A) and recruitment option (Figure S2C), where the number of observations (n) were 253, we replaced missing degree data with the VH-estimate for degree (average). The VH-estimate for degree was calculated by all those participants who reported a degree (participants with non-missing network data). Degree was censored to a maximum of 500 contacts per day per individual. The VH estimate for degree was 17.3 contacts per participant per day. For all other variables, we removed missing data and used only data from completely filled in questionnaires (n=220).

Judging from the plots in Figure S2, the sample composition did not stabilise for most variables, with the exception of household size, household members with symptoms, contacts while eating and self-reported symptoms, which seem to have stabilised after 200 participants.

All curves in Figure S2 show sudden increases or decreases in VH-estimates along with increasing sample sizes. For variables that are calculated in proportions, these jumps are

related to individuals who reported either a high degree (e.g., more than 100 contacts, resulting in a decrease in VH-estimate) or a low degree (e.g., less than 5 contacts, resulting in an increase in VH-estimate). Similar for age and other variables calculated in averages, participants with a high age (e.g., >40, see Figure S2E) or a high number of contacts (e.g., contacts while eating, see Figure S2G) result in sudden increases or decreases in VH-estimates.

Table I contains the sample means and VH estimates for all variables, estimated over the whole sample. Our sample mainly represents students and their contacts of similar age, and only those individuals that have access to the internet and those who use Facebook (the age groups 20-39 use the internet the most in Thailand[3]). Especially the age group 20-35 years, females and those with a bachelor or master degree are strongly overrepresented in our sample, compared to available Thai demographical estimates, which influenced the VH-estimates. For example, the VH-estimator estimated based on our sample that 82.9% of the Thai population was between 20-34 years, while the actual proportion is estimated to be 22.6%[4]. However, according to the VH-estimate the average household size in Thailand is 3.52, which is close to the average of 3.2 provided by the National Statistical Office Thailand[4]. Also, the VH-estimates for the different household sizes in Thailand were similar to the proportions provided by the ETDA Thailand[3], especially for the household sizes 4, 5 and 6 or more (Table I).

The overrepresentation of certain individuals is likely because we invited only students to act as seeds in our RDS survey, and the number of obtained waves was insufficient for our survey to penetrate into different layers of the Thai population. Although the internet-use (and access) is not equally divided over all age groups in Thailand, a future webRDS survey with a higher number of waves and with seeds selected from different age groups could possibly provide a better representative sample for the entire Thai population.

Table I. Sample proportions and estimated population proportions over the whole sample.

	Category	*n*	Sample mean	VH-estimate	Demographics Thailand[a]
Age (average)		253	26.70 (median: 25.00)	27.85	Median: 35.1 (2013[5]
Age groups (%)	10-14	253	0.79	0.19	6.45%[b][6] IU[c]: (<15): 0.8%[3]
	15-19		3.16	1.44	7.51%[b][6] IU[c]: 5.8%[3]
	20-24		39.13	40.57	7.25%[b][6] IU[c]: 14.9%[3]
	25-29		32.02	29.33	7.32%[b][6] IU[c]: 15.9%[3]
	30-34		13.83	13.02	8.02%[b][6] IU[c]: 18.0%[3]
	35-39		5.93	7.28	8.19%[b][6] IU[c]: 14.7%[3]
	40-44		1.19	1.95	8.37%[b][6] IU[c] (40-49):19.4%[3]
	45-49		2.37	1.97	7.93%[b][6]
	50-54		1.58	4.24	6.85%[b][6] IU[c] (50-59): 9.8%[3]
Male (%)		253	38.74	35.00	49.18 (2010)[7]
Education (%)	J. High S.	253	0.40	0.17	4.07%[b][8]
	High School		15.02	9.39	3.25%[b][8]
	Higher Dipl.		0.79	0.45	3.71%[b][8]
	Bachelor		62.85	67.34	
	Master		20.95	22.65	
Household size (average)		220	4.11	3.52	3.2 (2010)[4]
Household size, categorized (%)	1	220	20.45	22.57	14.7%[3]
	2		9.55	12.28	15.9%[3]
	3		16.36	15.47	19.2%[3]
	4		20.45	25.28	24.2%[3]
	5		14.09	13.03	13.7%[3]
	6 or more		19.09	11.37	12.3%[3]
Proportion used recruitment option (%)	Facebook	253	45.85	40.00	18.5 million users[9]: 28.70%[b]
	Indirect email		5.53	9.76	
	Direct email		5.53	3.92	
	No recruitment		43.08	46.32	

Table I (*continued*). Sample proportions and estimated population proportions over the whole sample

	Category	*n*	Sample mean	VH-estimate	Demographics Thailand[a]
Average number of contacts while eating (%)		220	12.21	6.90	
Proportion household members with symptoms (%)	0	220	57.27	52.78	
	1		15.45	11.02	
	2		8.64	8.13	
	3		3.64	5.25	
	4			2.28	
	5		2.27	5.29	
	6 or more		0.45	0.49	
	Unknown		9.09	14.75	
Number of symptoms (average)		220	1.80	1.63	
Flu/Cold Symptoms (%)	Flu symptoms	220	2.73	1.69	
	Cold symptoms		5.91	5.33	

a) $n_{pairs} = 85$; **b)** $n_{pairs} = 43$; **c)** $n_{pairs} = 20$

Literature

1. Heckathorn DD, Semaan S, Broadhead RS, Hughes JJ (2002) Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25. AIDS and Behavior 6: 55-67.
2. Volz E, Heckathorn DD (2008) Probability based estimation theory for respondent driven sampling. Journal of Official Statistics 24: 79-97.
3. Electronic Transactions Development Agency (Public Organization), Ministry of Information and Communication Technology (2013) Thailand Internet User Profile 2013.
4. National Statistical Office, Ministry of Information and Communication Technology Thailand (2012) The Information and Communication Technology Survey in Household.
5. Central Intelligence Agency (2013) The World Factbook. 25 October 2013 ed.
6. Statistical Forecasting Bureau, National Statistical Office Thailand, Ministry of Information and Communication Technology (2013) Statistical Yearbook Thailand 2013.
7. National Statistical Office, Ministry of Information and Communication Technology Thailand (2012) Key statistics of Thailand 2012.
8. Statistical Forecasting Bureau, National Statistical Office (2013) Thailand's Key Indicators 2013.
9. Socialbakers (2013) Checkfacebook: Facebook statistics of Thailand.

Figure S1. Sample composition over waves for gender, education and age.
The plots above (**A**, **C** and **E**) display the cumulative proportions or averages over waves. The plots below (**B**, **D** and **F**) display the proportions or averages in each wave.

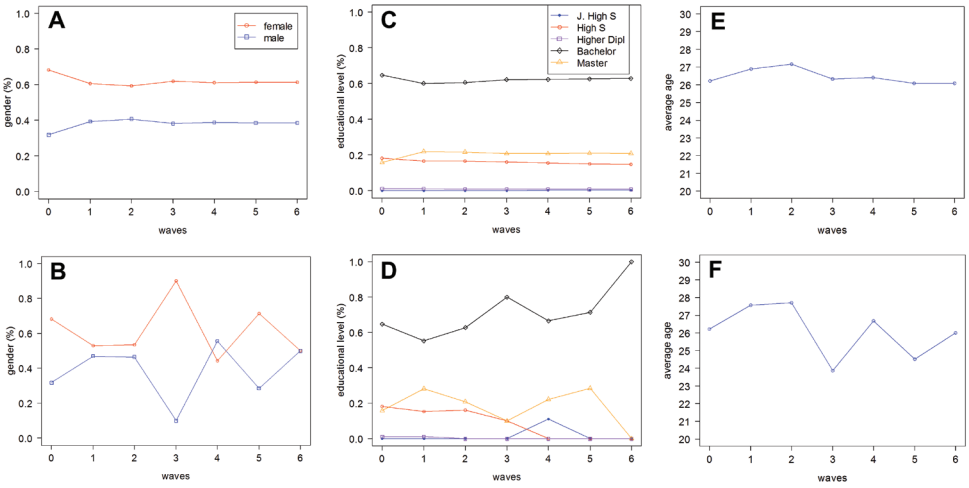


Figure S2 (*next page*). Sample proportions and VH estimates with increasing sample size (not adjusted for network size) for all variables.
The solid lines indicate the raw sample proportions or average, and the dotted lines indicate the VH estimates. (**A**) educational level; (**B**) number of household members (categorised); (**C**) recruitment option used; (**D**) number of household members with symptoms (categorised); (**E**) age (integer); (**F**) male; (**G**) average number of contacts while eating (integer); (**H**) flu (combination of the self-reported symptoms fever, headache and muscle pain) and cold symptoms (combination of the symptoms runny nose, sore throat and cough); (**I**) average number of self-reported symptoms (integer); (**J**) average number of household members (integer).

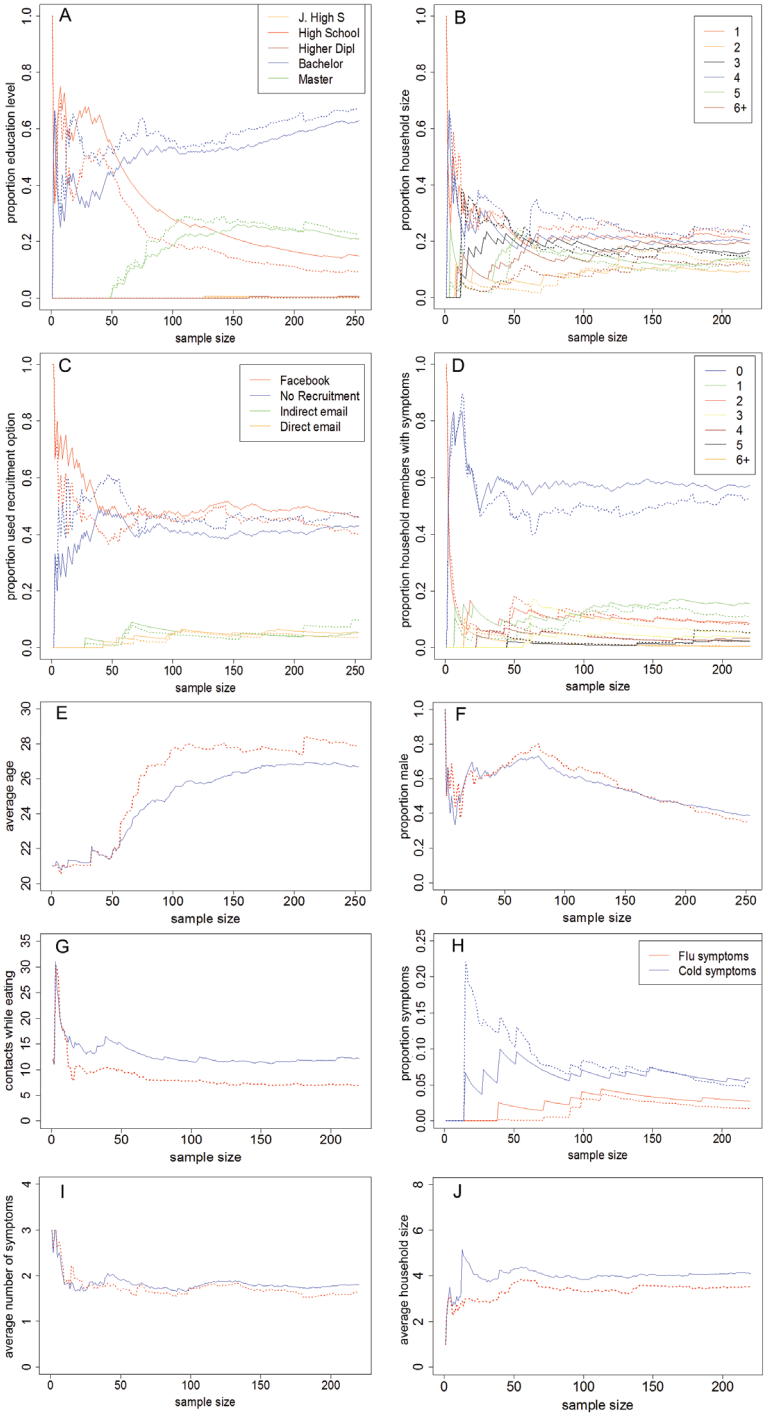

A
J. High S
High School
Higher Dipl
Bachelor
Master
proportion education level
sample size
B
1
2
3
4
5
6+
proportion household size
sample size
C
Facebook
No Recruitment
Indirect email
Direct email
proportion used recruitment option
sample size
D
0
1
2
3
4
5
6+
proportion household members with symptoms
sample size
E
average age
sample size
F
proportion male
sample size
G
contacts while eating
sample size
H
Flu symptoms
Cold symptoms
proportion symptoms
sample size
I
average number of symptoms
sample size
J
average household size
sample size

Table S1. Number of participants and used recruitment options over waves.

	Wave 0	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Total
Number of trees with	45	29	10	3	0	0	2	89
Number of participants in	89	86	43	21	9	7	2	257
Number of incomplete responses (persons who provided DI[a])	9 (7)	12 (11)	6 (6)	7 (6)	2 (2)	1 (0)	0 (0)	37 (32)
Filled in during weekday[b] (% per wave)	65 (73.0)	61 (71.8)	35 (81.4)	17 (85.0)	7 (87.5)	2 (33.3)	1 (50.0)	188 (74.3)
Filled in during weekend[b] (% per wave)	24 (27.0)	24 (28.2)	8 (18.6)	3 (15.0)	1 (12.5)	4 (66.7)	1 (50.0)	65 (25.7)
Yesterday was a weekday[b] (% per wave)	55 (61.8)	57 (67.1)	29 (67.4)	9 (45.0)	5 (62.5)	3 (50.0)	1 (50.0)	159 (62.8)
Yesterday was a weekend day[b] (% per wave)	34 (38.2)	28 (32.9)	14 (32.6)	11 (55.0)	3 (37.5)	3 (50.0)	1 (50.0)	94 (37.2)
Recruited by email[c] (% per wave)	73 (82.0)	22 (25.6)	9 (20.9)	0 (0)	0 (0)	0 (0)	0 (0)	104 (40.5)
Recruited by Facebook[c] (% per wave)	0 (0)	64 (74.4)	34 (79.1)	21 (100)	9 (100)	7 (100)	2 (100)	136 (52.9)
Through registration page[c] (% per wave)	16 (18.0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	16 (6.2)
RO Facebook[d] (% per wave)	46 (51.7)	31 (36.0)	18 (41.9)	9 (42.9)	7 (77.8)	5 (71.4)	0 (0)	116 (45.1)
RO indirect email[d] (% per wave)	10 (11.2)	4 (4.6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (5.4)
RO direct email[d] (% per wave)	9 (10.1)	4 (4.6)	2 (4.6)	0 (0)	0 (0)	0 (0)	0 (0)	15 (5.8)
Did not recruit (% per wave)	24 (27.0)	47 (54.7)	23 (53.5)	12 (57.1)	2 (22.2)	2 (28.6)	2 (100)	112 (43.6)

a) Participants provided demographic info (DI): gender, age and educational level, but no information on their contacts. **b)** Weekday is Monday to Friday; weekend is Saturday and Sunday. **c)** Includes respondents who did not complete the questionnaire. **d)** Recruitment option (RO) used by participant to recruit contacts.

Table S2. The number of contacts while having food (mean, median, SD) younger, same age or older than participant.

Age group	Younger	Same age	Older
14-19	12.4 (0; 27.8)	27.3 (15; 33.5)	27.3 (12; 29.9)
20-29	1.5 (0; 3.4)	6.1 (3; 9.5)	4.6 (2; 7.8)
30-39	4.4 (3; 4.1)	3.0 (1; 5.1)	3.0 (2; 3.9)
40+	5.0 (4; 5.4)	3.0 (0; 5.8)	3.8 (2; 5.7)

Note: Number of contacts while eating was censored to a maximum of 75 contacts per day for each respondent.

Table S3. Reported symptoms.

	Freq.	mean (SD) number of household contacts with symptoms[a]	Did not know whether household contacts had one or more related symptoms[b]
Fever	25	0.48 (0.87)	0
Chills	10	1.50 (1.51)	0
Runny nose	72	0.76 (1.28)	6
Sore throat	59	0.79 (1.13)	7
Cough	39	0.53 (0.88)	4
Headache	83	2.04 (3.25)	10
Muscle pain	88	1.02 (1.39)	7
Diarrhea	18	0.69 (1.14)	2
Other symptoms	6	0.50 (0.84)	0
No symptoms	59	0.79 (1.49)	6

a) Household contacts having one or more related symptoms. **b)** Number of participants who reported the symptom, but did not know whether household members also had one or more related symptoms.

Chapter 3

Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling

Mart L. Stein[1,2*], Jim E. van Steenbergen[2,3], Vincent Buskens[4], Peter G.M. van der Heijden[4,5], Charnchudhi Chanyasanha[6], Mathuros Tipayamongkholgul[7], Anna E. Thorson[8], Linus Bengtsson[8,10], Xin Lu[8,9,10], Mirjam E.E. Kretzschmar[1,2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[3] Centre for Infectious Diseases, Leiden University Medical Centre, Leiden, The Netherlands

[4] Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, The Netherlands

[5] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom

[6] Department of Microbiology, Faculty of Public Health, Mahidol University, Bangkok, Thailand

[7] Department of Epidemiology, Faculty of Public Health, Mahidol University, Bangkok, Thailand

[8] Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

[9] College of Information System and Management, National University of Defense Technology, Changsha, China

[10] Flowminder Foundation, Stockholm, Sweden

PLoS ONE 2014, 9(11):e113711

ABSTRACT

Understanding infection dynamics of respiratory diseases requires the identification and quantification of behavioural, social and environmental factors that permit the transmission of these infections between humans. Little empirical information is available about contact patterns within real-world social networks, let alone on differences in these contact networks between populations that differ considerably on a socio-cultural level. Here we compared contact network data that were collected in the Netherlands and Thailand using a similar online respondent-driven method. By asking participants to recruit contact persons we studied network links relevant for the transmission of respiratory infections. We studied correlations between recruiter and recruited contacts to investigate mixing patterns in the observed social network components. In both countries, mixing patterns were assortative by demographic variables and random by total numbers of contacts. However, in Thailand participants reported overall more contacts which resulted in higher effective contact rates. Our findings provide new insights on numbers of contacts and mixing patterns in two different populations. These data could be used to improve parameterisation of mathematical models used to design control strategies. Although the spread of infections through populations depends on more factors, found similarities suggest that spread may be similar in the Netherlands and Thailand.

INTRODUCTION

The spread of the majority of infectious diseases, including those with pandemic potential, depends on the underlying social contact network of the host population[1,2]. In many countries pandemic preparedness and control measures during large-scale outbreaks are analysed using mathematical models to provide support for public health policy decisions. These models often assume proportionate mixing in a population that is stratified by contact rate[3]. In many models with heterogeneity in contact rates the proportionate mixing assumption is used for convenience and because information about mixing patterns by contact rate is usually not available. In view of results that show that mixing patterns between susceptible individuals greatly influence infectious disease dynamics[4-6], it is pertinent to collect empirical information about mixing patterns that determine the transmission of infectious diseases.

Most studies that investigated contact patterns relevant for the spread of respiratory infections that are spread by close-contact used egocentric data, i.e. responses of participants who were sampled independently of one another[7]. Such studies provide no information about the contact network structure beyond the contact persons reported by participants. Electronic proximity devices can provide more insight in social contact networks, but their use is limited to specific settings[8,9]. Although complete analysis of social contact networks is practically impossible for an entire population, partial insight in degree distributions and contact mixing patterns within these networks can increase our understanding of disease dynamics[10,11]. For example, when individuals primarily have contact with individuals of similar age, infections are likely to spread faster within those subgroups than among them. Furthermore, investigating contact networks can highlight the existence of so-called hubs, i.e. individuals with a relatively high degree[12], who are more likely to become infected and to infect others, and can act as a bridge between communities. Such information can be used to determine optimal control strategies.

Respondent-driven sampling (RDS) is a chain-referral method that was introduced to estimate disease prevalence in hard-to-reach populations, such as drug-users[13,14]. RDS begins with the selection of initial respondents, so-called “seeds”, who are asked to recruit contact persons from their social environment; these contact persons are asked to do the same. Unlike snowball sampling, RDS provides the option to track who recruited whom by means of personal codes. Previously, we proposed online RDS (webRDS) as a method for sampling contacts of contacts and beyond of individuals to study social contact network structures[15]. To investigate the determinants of using this sampling method for eliciting contact information, we conducted webRDS surveys in both the Netherlands and Thailand focusing on contact patterns relevant for the transmission of respiratory pathogens. Note that our objective was not to use RDS for estimating population proportions from our samples.

In this paper, we compare social contact data that were collected with a similar methodology in the Netherlands and Thailand. Earlier, a large population-based study showed striking similarities in mixing patterns across eight European countries[16], and also comparable egocentric data on mixing patterns were collected in countries in Asia[17,18]. However, to our knowledge, no studies performed a direct comparison between contact network data collected in a European country and a country in Asia. Our first aim was to study whether populations with different cultural, demographic and social backgrounds also differ in social mixing patterns. In particular, we aimed to investigate differences and similarities in numbers of contact persons, mixing patterns (i.e., who has contact with whom?) and effective contact rates between both countries. The latter help quantifying the transmission potential of certain directly transmitted infectious diseases through a population. Secondly, our aim was to investigate whether there are differences between the two study populations with respect to factors that drive peer-recruitment via the internet.

METHODS

We conducted webRDS surveys to collect contact network patterns in Thailand and the Netherlands between November 2012 and May 2013 (see Figure S1). Recruitment methods and questionnaires were virtually identical in both countries (apart from different survey languages and minor tailoring to local situations that is elaborated below), in order to carry out a country comparison. Seeds were invited from convenience samples of students from two universities located in Utrecht and Amsterdam (the Netherlands) and two Bangkok Universities (Thailand) (Table 1). The majority of the students in both countries leave the parental home and live in dormitories[19,20]. In the Netherlands it is custom for a student to have an own bedroom but to share all facilities, whilst students in Thailand often share the same room or even the same bed. Seeds in both countries were first informed about the survey, either in physical group meetings or by an information email. Seeds were then invited by email containing a unique hyperlink to the questionnaire or they could register themselves on the survey website using their email address. The questionnaire and invitation emails were provided in local languages and additionally in English.

Contact definition and questionnaire

The questionnaire was kept short to limit the burden for participants; in total it consisted of eleven questions. A 'contact person' was defined as a person sitting or standing within arm's length of the participant (denoted as 'YourSpace', see Figure S2) for 30 seconds or longer. This definition made it fairly simple to recall contact persons, could be easily explained in an online questionnaire, and was not restricted to only face-to-face interactions. Participants were asked to record numbers of contact persons at different settings during one day (namely 'yesterday').

Table 1. Characteristics study populations and recruitment.

		Netherlands	Thailand
Participation	Total participants	358	257
	Complete responses	89.9% (322)	85.6% (220)
	Invited seeds	189	191
	Seeds who filled in questionnaire	48.7% (92)	46.6% (89)
	Seeds who successfully recruited[a] one or more recruitees	32.8% (62)	20.4% (39)
	Pairs of recruiter-recruitee (both completed survey)	233	140
	Maximum number of waves	5	6
	Trees with two or more waves	46.8% (29)	38.5% (15)
Successfully recruited[a] by	Facebook	10.4% (24)	83.6% (117)
	Indirect email invitation	52.6% (121)	5.0% (7)
	Direct email invitation	37.0% (85)	11.4% (16)
Age	Overall mean (range)	33.2 (16-75)	26.7 (14-52)
	Mean age seeds[b] (range)	27.7 (25-53)	26.2 (14-48)
Sex	Female	62.1% (221)	61.4% (156)
	Male	37.9% (135)	38.6% (98)
Education	Higher education	89.9% (320)	83.9% (213)
	Vocational Education / Higher diploma	4.2% (15)	0.8% (2)
	High School / Other education	5.9% (21)	15.3% (39)
Symptoms	Two or more symptoms	39.6% (128)	29.9% (66)
	ILI symptoms[c]	1.9% (6)	3.2% (7)
	Common cold-like symptoms[c]	7.4% (24)	5.9% (13)

[a.] Successfully recruited means that the contact person completed the questionnaire.
[b.] Seeds who successfully[a] recruited one or more contact persons.
[c.] ILI symptoms are a combination of the self-reported symptoms: fever, headache and muscle pain; common cold-like symptoms are a combination of the symptoms: runny nose, sore throat and cough.

We asked participants to record the number of contact persons while travelling and at different locations. Modes of transport and locations were prespecified (i.e., a subset of transport and location types of interest, see Table S1) in the questionnaire and tailored to each country. Participants were asked in a separate section to record the number of contact persons within arm's length while eating, as it is custom in Thai culture to share several small dishes and drinks with friends and / or family in restaurants, which facilitates the potential transmission of infections. For numbers of contact persons at different locations and while eating, the participant was requested to indicate whether the contact person was younger, the same age or older than the participant. We also asked participants for their age, sex, educational level, postal code, and the number and ages (specified in age groups) of persons living in their household during the past seven days. Furthermore, we included a question to record any symptoms (provided in a list, see Table S2) that participants and/or household members

experienced in the past seven days. A combination of the symptoms fever, headache and muscle pain was indicated as influenza-like-illness (ILI) symptoms, and a combination of the symptoms runny nose, sore throat and cough as common cold-like symptoms.

Recruitment and online system

At the end of the questionnaire, we asked participants to recruit four new participants (further referred to as 'recruitees') with whom they had contact according to the above definition in the past 7 days (see also Figure S2). We provided three options for recruitment, namely through Facebook (i.e., send Facebook friends a private message), indirect email (i.e., provide your email address and receive four invitation emails that can be forwarded to recruitees) and direct email (i.e., provide email addresses of your recruitees and the system sends out invitation emails automatically). All invitations contained information on the survey and a personal link and code. All emails and the first page of the questionnaire contained a link to unsubscribe for the survey.

After inviting recruitees, participants were referred to the project website[21] where participants could view a graphical representation of the network components found in the study as a non-material incentive for their participation. These network graphs were anonymous but showed the personal codes provided in the invitations so that participants could identify their own location in the network. Full details of the online RDS survey system can be found in M.L. Stein et al.[15].

Ethical statement

All information about the survey was available on all web pages and could be accessed at any time. All pages contained a log-out button that referred users to a search engine. The system converted IP addresses to a unique anonymous code using a one-way encryption algorithm and original addresses were deleted. Login passwords were only valid for a single participation and could not be used on two computers at the same time. All communication between the users and the survey website was encrypted. We obtained informed consent via the first webpage, on which inter alia study purposes and benefits, and statements on privacy and confidentiality were displayed. Users were able to accept the informed consent form by clicking the 'Start the Survey' button and to continue to the questionnaire, or to deny by clicking the button "No Thanks" whereby users were automatically logged out of the survey. There was no age limit for inclusion included in the study protocols. Due to the setup of the study a selection of participants based on age or a separate consent form for parents/caretakers of minors was not possible. The study received approval from the Medical Ethical Committees from both the University Medical Center Utrecht, the Netherlands (reference: 12-247/C) and the Faculty of Public Health Mahidol University, Thailand (reference: MUPH 2012-187).

Statistical analysis

Using terminology from social network analysis, we defined a participant's degree as the total number of contact persons reported by the participant during one full recording day (i.e., the sum of the numbers of contact persons at different locations and while travelling). We removed two Thai participants who reported a degree of more than 2200, as these extreme degree values are highly unlikely according to our contact definition and their individual answers made clear that they misinterpreted the definition. A negative binomial distribution $N(\mu, k)$ was fitted to the observed frequency distribution of degrees using maximum likelihood methods, where μ denotes the mean and k the dispersion parameter; parametric bootstrapping was performed to estimate confidence intervals. We used a Q-Q plot to graphically compare the degree distributions of the two country samples and used the Anderson-Darling k-sample test to test the hypothesis that these two samples come from one common population.

To investigate mixing patterns within our two samples with respect to the measured variables, we calculated correlation coefficients between pairs of randomly chosen individuals that were one, two, three and four or more link steps away from each other in the same network chain, by using the shortest paths between any two persons in the same network chain[11]. Thus, we calculated correlations between recruiters and their recruitees in consecutive waves. Here we assumed that a recruitment link between two participants can be interpreted as a contact in the sense of our contact definition. We used Pearson's r for integer variables (age, degree, and number of contact persons while eating), phi coefficient (r_{φ}) for binary variables (sex and two or more symptoms reported) and Spearman rank-order (r_{rank}) for ordinal variables (education).

Logistic regression analysis was used to investigate which measured characteristics are important for online peer-recruitment. We defined the binary outcome "intention to recruit" as a respondent that requested invitations for recruitees on the last survey page (versus a respondent who did not request invitations). We used this binary outcome as the RDS system only registered whether or not a participant clicked the button to request for four invitations, and not how many of the four invitations were actually sent out to recruitees. We repeated the logistic regression analysis for the sampled data without seeds (data without wave 0), to investigate the influence of seeds on the outcome. See supplementary file Text S1 for a more detailed description. In RDS methodology, samples are weighted for individual degrees as persons with a high degree theoretically have a higher chance to get invited than persons with a low degree[22]. For illustration purposes, we used the output of the logistic regression model to estimate for individual subjects in each country the probability of the intention to recruit as a function of degree; adjusted for age, sex, education and household size. Confidence intervals (95%) were obtained using standard errors.

In addition, we assessed for each country sample the validity of the first-order Markov assumption[23], i.e., that correlations found between recruiter and recruitee are only dependent on the direct recruiter. This was done by calculating the correlations for age and sex between seeds (wave 0) and their recruitees in consecutive waves (maximum up to 3 waves, due to limited number of participants in waves ≥4). For the numeric variable 'age' we compared $r_{waves0\text{-}3}$ with $r_{waves0\text{-}1} * r_{waves1\text{-}2} * r_{waves2\text{-}3}$; for the categorical variable "sex" we raised the 1-step transition matrix to the third power to obtain $r_{waves0\text{-}3}$.

Effective contact rate

In a population that is heterogeneous with respect to contact rates, the basic reproduction number depends on the so-called effective contact rate C. Assume that a population is stratified into n subgroups with contact rates c_i, i=1, 2, 3,..., n. Furthermore, assume that mixing is proportionate, i.e., random mixing weighted by the contribution of a subgroup to the total number of contacts in the population. Then it can be shown that the basic reproduction number R0 is proportional to the effective contact rate given by

$$C = \overline{c} + \frac{v}{\overline{c}} \qquad (1)$$

where $\overline{c}$ denotes the population mean of the c_i and v the variance of the c_i (see Anderson and May[3]; p 233).

Here we used the concept of the effective contact rate to quantify the heterogeneity in contact rates found in our sample and to assess their possible impact on transmission of infection. Based on the assumption that degree distributions observed in our sample are representative for the degree distribution in the population, we computed C with data on degree. In addition, we computed an effective contact rate with contact persons at different locations, contact persons while travelling, contact persons while eating and household members in order to compare and assess the contribution of each of these categories. This provides some indication for the possible effects of control measures that reduce contact rates such as school closure or cancellation of mass gathering events on transmission of infection[24]. We are aware that the assumption of representativeness of our sample of the degree distribution in the population is most likely not fulfilled, but want to demonstrate how such data can be used to investigate the impact of interventions on social network connectivity.

We also note that the applied statistical tests do not take the interdependence structure within our sampled data into account. The RDS data files are available online, doi: 10.6084/m9.figshare.1147465. R version 3.0.3 was used for statistical analysis and RGraphviz for creating Figure 1. See supplementary file Text S2 for the full R source code and applied libraries.

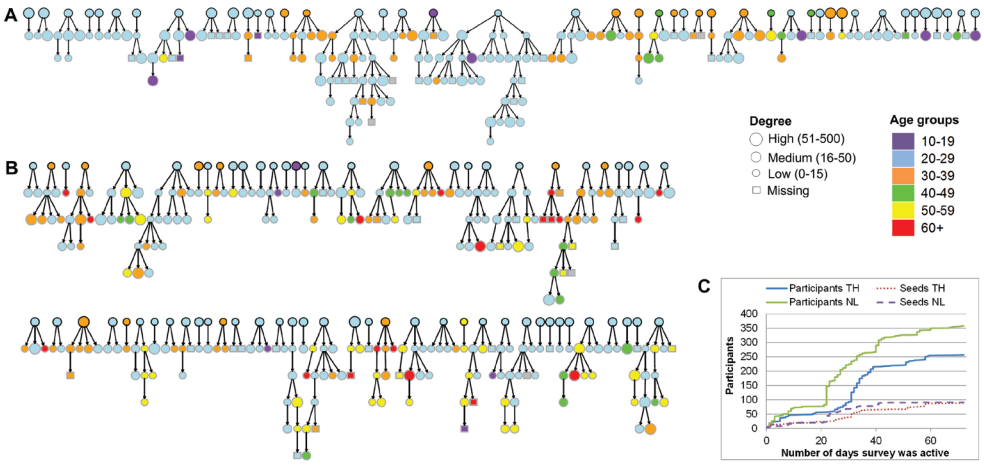


Figure 1 RDS networks and participation over time. The recruitment networks obtained in (**A**) Thailand and (**B**) the Netherlands. Seeds are indicated with black borders; seeds who did not recruit other participants are not displayed. The circle sizes reflect the total number of contact persons (degree) reported by each participant. The colours indicate the different age groups, illustrating the mixing by age. (**C**) Displays the cumulative number of participants and seeds over time the surveys were active in each country. The jumps in the cumulative number of participants are caused by newly invited seeds.

RESULTS

A total of 358 individuals participated in the Netherlands, compared to 257 participants in Thailand. Although in both countries the survey's operating days, numbers of invited seeds and seeds who filled in the questionnaire were similar, seeds in the Netherlands were more successful in inviting recruitees who also completed the questionnaire (see Table 1). Nevertheless, in Thailand we reached up to six waves (in two trees) of recruitees, compared to five waves (in three trees) in the Netherlands (see Figure 1). There were 233 pairs of recruiter-recruitee in the Netherlands and 140 in Thailand. In both countries, more than half of all participants invited recruitees, namely 55.4% (198) in the Netherlands and 56.4% (145) in Thailand. Of all seeds, 87.0% (80) in the Netherlands invited one or more recruitees compared to 73.0% (65) in Thailand. The majority of the Thai participants (45.1% of 257 participants) used Facebook to invite recruitees, while in the Netherlands invitations were mainly sent by email (Figure 2).

Characteristics of participants

The age distributions of seeds in both countries differed significantly (K-S test p=0.018); with seeds in the Netherlands having a mean age of 26.7 (median: 26.0, range: 19-53) and in Thailand 26.2 (median: 24.0, range: 14-48). In the Netherlands the relatively younger participants more often recruited individuals from an older generation, which led to a wider range of participating age classes. Consequently, the entire Dutch sample had a higher overall mean age (33.2, median: 28.0) compared to Thailand (26.7, median: 25.0). The age distributions of all participants in both countries differed significantly (K-S test p<0.001).

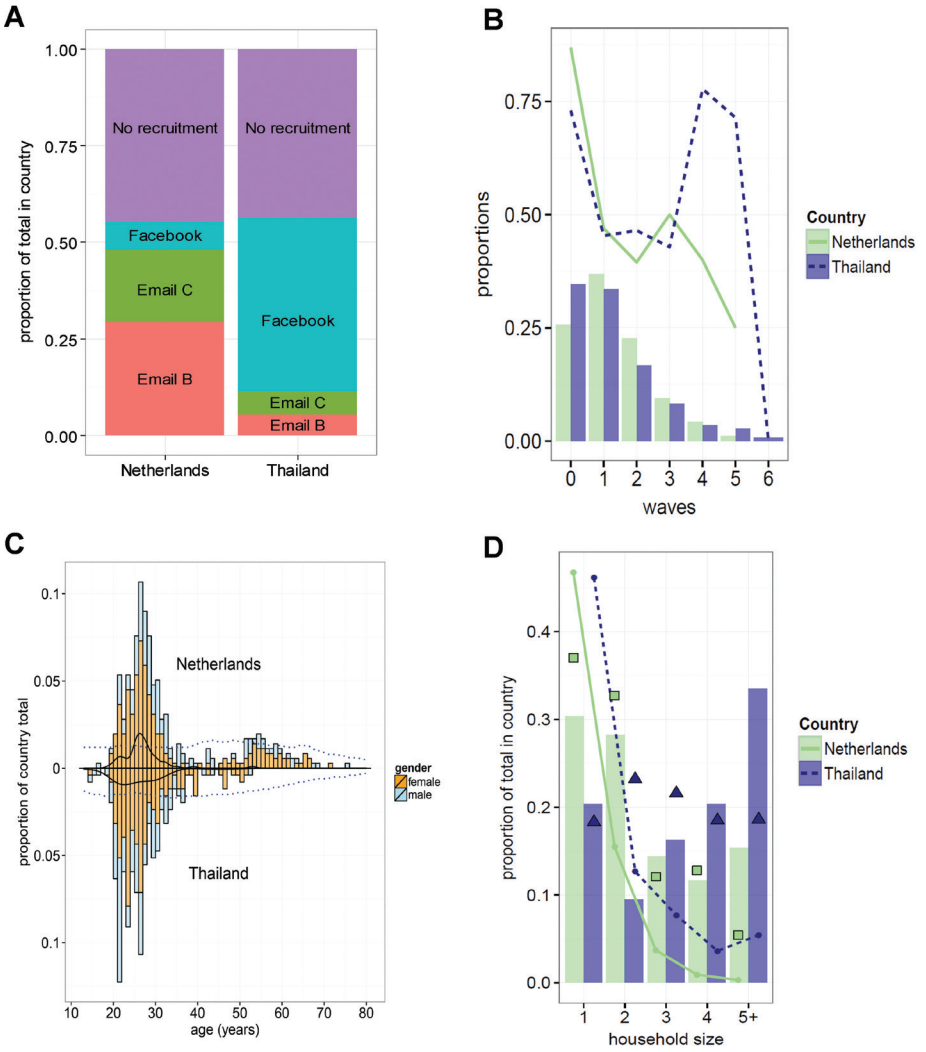


Figure 2 Characteristics of study participants and recruitment. (**A**) Recruitment options used by participants. (**B**) Distribution of participants over waves (in proportions of country totals). The lines indicate the proportion of participants, in each wave, who also invited recruitees (these participants requested four invitations on the last page of the survey). In both countries the proportion of seeds (wave 0) that invited recruitees was more than 70%, and on average more than half of all participants invited one or more recruitees. (**C**) The age distributions in each country for all participants, with colour indicating sex. The solid lines (black) display the age distributions for seeds only. The blue dotted lines show the estimated Dutch and Thai population estimates for 2012. (**D**) The household sizes reported by participants in each country (bars). The squares (for the Netherlands) and triangles (Thailand) show the household sizes as reported by the national census bureaus. The lines in the graph indicate the number of household members with one or more symptoms (as a proportion of the country totals), as reported by participants (30 participants in the Netherlands did not known whether household members had any symptoms, compared to 20 in Thailand).

The majority of participants in both countries was female and highly educated (Table 1). The Dutch sample contained 11.6% more females compared to Dutch national census data, in the Thai sample there were 10.4% more females than in the national census data from Thailand[25]. Figure 2 displays the age distributions of all participants, the age distribution of all seeds

who invited one or more recruitees, and the distributions by sex of all participants in both countries. Dutch participants reported more often living in households consisting of only one or two members (by 58.6% of 326), while in Thailand household sizes of three or more persons were more often reported (by 70.1% of 221). The distribution of reported household sizes was roughly in agreement with national census data (Figure 2d). However, compared to census data, both country samples contained higher proportions of participants living in a household of five or more members, and the Thai sample contained relatively low proportions of participants living in a two-member household[25,26].

Self-reported symptoms

Participants in the Netherlands reported 15.7% more symptoms and 1.5% more common cold-like symptoms than Thai participants, but in Thailand participants reported 1.3% more ILI symptoms (Table 1). In Thailand, 33.5% (of 221) reported one or more household contacts with symptoms, compared to 31.6% (of 323) in the Netherlands.

Reported numbers of contact persons

A total of 12812 contact persons were reported in Thailand, compared to 8336 in the Netherlands. Figure 3a shows that the medians of the distributions of numbers of contact persons at different locations, while travelling and while eating were higher in Thailand than in the Netherlands. Summing contact persons reported for each location showed no significant differences between days of the week in both samples (see Figure S3). In Thailand there were significant differences in distributions between days of the week for contact persons while travelling, e.g., between Monday and Friday (p=0.036) and Saturday and Sunday (p=0.014), these differences were not observed in the Netherlands (see Figure S4).

Figure 3b shows the fitted negative binomial distributions to the observed frequency distributions of degree in the Netherlands (μ = 25.65 [23.15–28.51]; k = 1.00 [0.87–1.16]) and Thailand (μ = 58.51 [50.30–67.24]; k = 0.83 [0.72–1.02]). The non-overlapping confidence intervals of the parameter estimates μ indicate that the sampled degree distributions were not similar, which is confirmed by visually examining the Q-Q plot (Figure 3c) and the Anderson-Darling test (p<0.001). Table 2 displays the means, variances and effective contact rates for the various contact categories. The higher means and variances of the observed degree distributions in Thailand indicate not only more heterogeneity, but also overall more contact persons compared to the Dutch sample. Consequently, the effective contact number computed with degree was almost two times higher in Thailand than in the Netherlands (respectively 205.5 and 111.9). Similar differences in means, variances and C between both countries were observed for contact persons at different locations, while travelling and while eating.

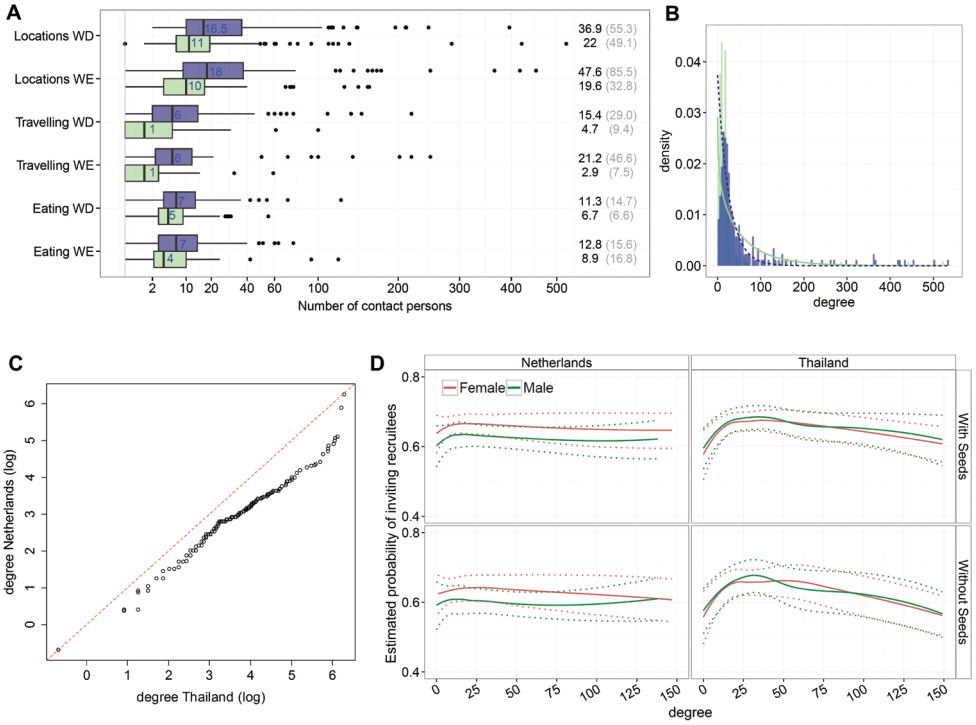


Figure 3 Egocentric contact data. (**A**) Box and whisker plot showing the median, quartiles and 95 percentiles of numbers of contact persons reported by participants for travelling, at different locations and while eating, during weekdays (WD, Monday-Friday) and weekend days (WE, Saturday-Sunday). Dutch sample is indicated in green and Thai sample in purple. (**B**) Distribution of the overall reported numbers of contact persons ("degree") in each country. The green solid line (the Netherlands) and the dotted purple line (Thailand) indicate the fitted negative binomial distributions for reported degree. (**C**) Quantile-Quantile plot of the degree distributions displayed in plot 3B. (**D**) The probability estimates of inviting recruitees ("intention to recruit") as a function of degree, specified by sex (and additionally adjusted for variables age, education and household size). Dotted lines indicated confidence intervals (95%). Values were obtained using logistic regression analysis for the full sample and the sample excluding seeds (i.e. sample without wave 0).

Drivers of online recruitment

Figure 3d shows the estimated probability of inviting recruitees as a function of degree. In Thailand, degree seemed to influence participants' intention to invite recruitees. However, when excluding Thai seeds there was no significant influence of degree, and only participant's age slightly influenced the intention to recruit. In the Netherlands, female participants requested more often the four invitations to invite their recruitees compared to male participants (p=0.009, adjusted for degree, age, education and household size). Excluding seeds in the analysis showed for the Dutch sample similar results with respect to those differences between males and females. See supplementary file Text S1 and Figure S5 for the extended analysis.

Mixing patterns

Figure 4a displays for each age group the proportions of contact persons younger than, same age or older than the participants at different locations (reported number of contact persons at different locations aggregated) and separately while eating. In both study populations,

Table 2. Effective contact rates *C*.

		The Netherlands	Thailand
Degree (total number of contact persons)	Mean	25.6	58.5
	Variance	2212.2	8601.9
	C	111.9	205.5
Contact persons at different locations	Mean	21.4	40.9
	Variance	2057.6	4671.4
	C	117.6	155.0
Contact persons while travelling	Mean	4.2	17.6
	Variance	80.8	1346.9
	C	23.3	94.2
Contact persons while eating	Mean	7.3	11.8
	Variance	106.5	225.0
	C	21.9	30.8
Household members	Mean	3.0	4.0
	Variance	10.1	8.6
	C	6.4	6.1

participants showed a strong tendency to invite recruitees of similar age, sex and education, although the Dutch sample showed lower correlations for mixing by education (Table 3). The assortative recruitment patterns by age, sex and education disappeared in both samples after a distance of three or more link steps between any two persons in the same network chain. However, in the Thai sample, assortative recruitment by education persisted until a distance of three link steps. By contrast, mixing by degree was observed not to be assortative (or disassortative) in both countries, but random for all link distances between two persons in the same chain. Figure 4b illustrates the mainly assortative recruitment patterns by age in each country, but also some recruitment of participants from different age classes, especially in Dutch sample. Figure 4c illustrates for both countries the random mixing by degree of recruiters-recruitees.

First-order Markov assumption

In both the Netherlands and Thailand, the sample correlations $r_{waves0\text{-}3}$ for age (respectively 0.171 and 0.320, see Table S3) were not in agreement with $r_{waves0\text{-}1} * r_{waves1\text{-}2} * r_{waves2\text{-}3}$ (respectively 0.086 and 0.147). Also for sex, sample correlations (respectively 0.106 and 0.000) slightly deviated from correlations obtained from the three-step transition matrices (respectively 0.029 and 0.013). This suggests for peer-recruitment, with regard to age and sex, a higher-order process in both countries. Thus, correlations found between the recruiter-recruitee are not only dependent on the direct recruiter, but also on recruiters in previous waves. Due to a limited number of waves in most network trees, we were unable to quantify the exact order level.

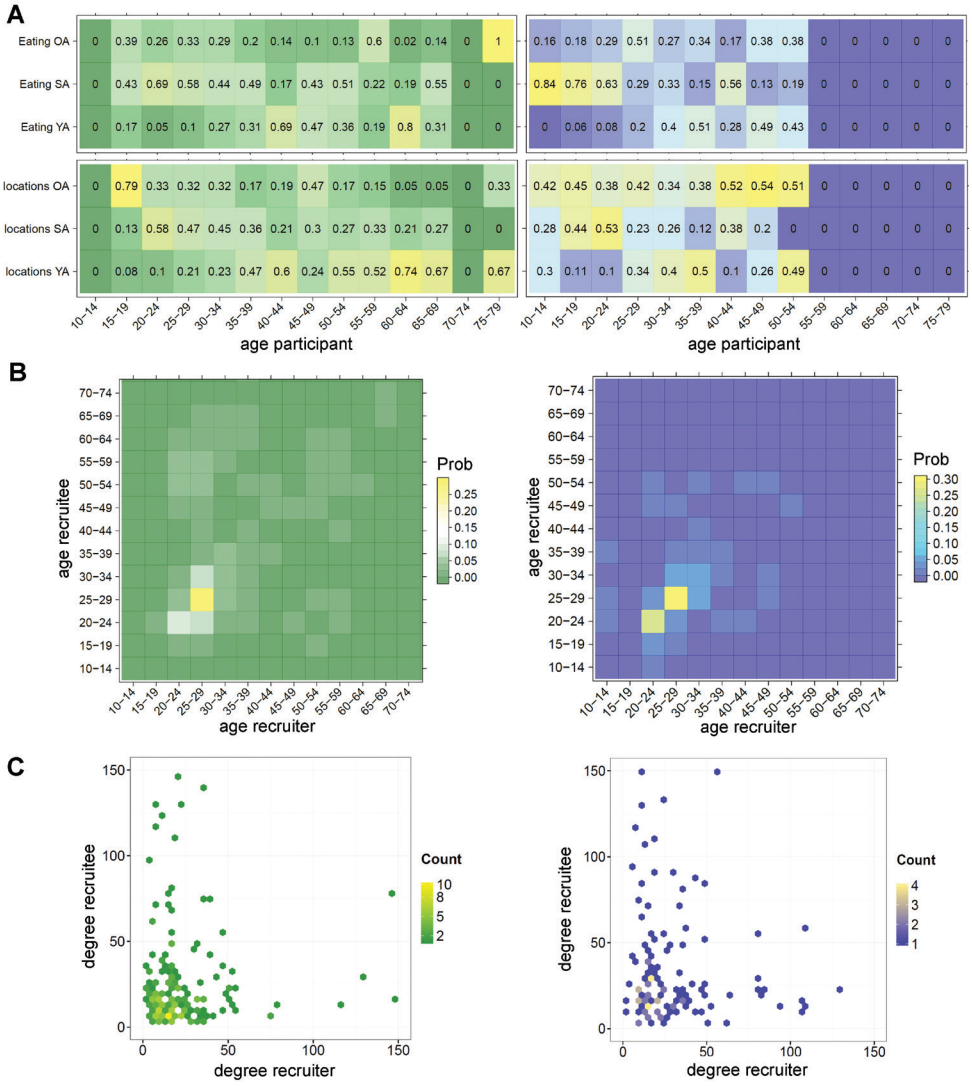


Figure 4 Who has contact with whom? The green coloured plots on the left are all based on data collected in the Netherlands and purple coloured plots on the right on data collected in Thailand. Plots in (**A**) display for each age group the proportion of reported contact persons that were younger, the same age or older than the participants. Displayed values are proportions that were calculated separately for the categories "contact persons while eating" and "contact persons at different locations". The brighter coloured cells indicate higher proportions. (**B**) The contact intensity matrices for recruiter-recruitees pairs with respect to age (displayed as proportions of country totals) and (**C**) degree (displays actual reported degree; outliers above 150 are not displayed). The brighter coloured cells/points in B-C illustrate higher contact recruitment rates and the darker cells/points lower rates.

Table 3. Correlations between any two linked individuals in the same network tree.

	Country[a]	Number of link steps between any two individuals in the same network			
		1	2	3	≥4[b]
Age (*r*)	**NL**	0.469 [0.369–0.558]*	0.158 [0.048–0.265][£]	-0.071 [-0.193–0.053]	0.021 [-0.102-0.143]
	TH	0.555 [0.439–0.652]*	0.350 [0.224–0.465]*	0.087 [-0.050–0.222]	0.040 [-0.039–0.118]
Sex (*φ*)	**NL**	0.289 [0.174–0.396]*	0.055 [-0.057–0.165]	0.103 [-0.021–0.223]	0.066 [-0.057–0.187]
	TH	0.205 [0.054–0.346][£]	0.132 [-0.005–0.264]	-0.045 [-0.180–0.092]	0.023 [-0.054–0.100]
Education (r_{rank})	**NL**	0.133 [-0.018–0.293][£]	-0.001 [-0.095–0.114]	0.037 [-0.087–0.174]	-0.050 [-0.139–0.056]
	TH	0.553 [0.362–0.735]*	0.336 [0.141–0.531]*	0.188 [-0.005–0.392][£]	-0.071 [-0.112– -0.018]
Degree log (*r*)	**NL**	0.081 [-0.047–0.207]	0.049 [-0.071–0.169]	0.028 [-0.109–0.164]	-0.009 [-0.144–0.127]
	TH	-0.055 [-0.220–0.113]	-0.103 [-0.251–0.049]	0.119 [-0.038–0.270]	0.042 [-0.050–0.133]
Degree - categories (r_{rank})	**NL**	0.046 [-0.085–0.180]	0.128 [0.012–0.244][£]	0.022 [-0.115–0.164]	-0.076 [-0.199–0.055]
	TH	0.049 [-0.119–0.223]	-0.027 [-0.186–0.122]	0.087 [-0.068–0.234]	0.005 [-0.080–0.095]
Number of contact persons while eating log (*r*)	**NL**	0.051 [-0.077–0.177]	0.020 [-0.100–0.138]	-0.011 [-0.146–0.125]	0.091 [-0.045–0.223]
	TH	0.167 [-0.000–0.325]	-0.007 [-0.159–0.145]	-0.020 [-0.176–0.136]	-0.002 [-0.094–0.090]
Two or more symptoms reported (*φ*)	**NL**	-0.018 [-0.142–0.111]	0.1314 [0.013–0.257][£]	-0.019 [-0.152–0.115]	-0.010 [-0.148–0.120]
	TH	-0.059 [-0.208–0.115]	0.000 [-0.152–0.157]	-0.111 [-0.257–0.038]	0.001 [-0.095–0.089]

[a] NL: the Netherlands; TH: Thailand.
[b] Distances of four or more links were lumped together.
* p-value < 0.001.
[£] p-value < 0.05.

DISCUSSION

Here we have reported on results from similar webRDS surveys conducted in the Netherlands and Thailand. Our study provides, to our knowledge, the first comparison of social contact data relevant for infections transmitted by the respiratory or close-contact route from countries in Europe and South-East Asia. Even though these two countries differ in many aspects of their cultural, social and demographic determinants, comparison of the country samples showed clear similarities in contact and mixing patterns. Mixing was assortative by age, sex and education, and random by degree. We reached a similar number of waves in both countries using the same non-material incentives. By contrast, differences between both country samples were observed for age classes reached, age of daily contact persons and levels of dispersion in overall degree distributions. Moreover, seeds in the Netherlands were more successful in inviting recruitees, and female participants in the Netherlands showed a significantly higher intention to recruit; this sex difference was not observed for the Thai sample. Thai participants primarily used Facebook, which is very popular among young Thai[27], to invite recruitees while standard email was more preferred in the Netherlands.

Despite comparable high population densities in Bangkok (5294 persons per square kilometer[26]) and the Dutch regions (3353 in Utrecht and 4767 in Amsterdam[28]), where we invited our seeds, Thai participants reported on average higher numbers of contact persons. Both study samples showed a strong heterogeneity in numbers of contact persons per individual, which was also found earlier in other populations[17,29]. Theoretical studies have demonstrated that heterogeneity in numbers of contacts influences both the infection attack rate[30] and the basic reproduction number[31], thereby impacting on the effectiveness of control measures. It should be noted that the adopted contact definition was different from the definition generally used by contact diary surveys (i.e., two-way conversation in close proximity or a physical contact like shaking hands or kissing). This has implications for the comparability with other studies, especially when comparing location-dependent contact behaviour[32].

Our study samples showed random mixing by degree, which gives support to the proportionate mixing assumption underlying the derivation of the effective contact rate as computed in equation (1). Regarding the effectiveness of control measures our data provide some indications for assessing possible effects of interventions that reduce numbers of contacts. For instance, our data illustrate that by asking all individuals in a population to stay at home during an outbreak would reduce the effective contact rate from 111.9 to 6.4 in the Netherlands and from 205.5 to 6.1 in Thailand.

However, the question remains how well correlations between recruiters and recruitees describe the mixing patterns in social networks; this depends mainly on the randomness of

peer-recruitment. In Thailand, participants invited more recruitees of similar age, sex and education, suggesting that the survey spread in a more homogeneous group compared to the Netherlands. This might be due to inequalities between urban and rural areas in Thailand, especially with respect to education, income and occupation opportunities[33]. Educated Thai are therefore more likely to connect to similarly educated peers. In the Netherlands, participants from the younger age groups more frequently invited recruitees from older age classes that led to recruitment within these older age groups. Such links between younger and older age groups were less visible in Thailand. This could either be due to social-cultural differences or to the low proportion of internet-users among Thai aged 35 and older.

Our Dutch recruiter-recruitee matrix for age showed a similar assortative pattern (including links between younger and older age groups), as the contact matrix that was collected earlier in the Netherlands during a large egocentric survey (POLYMOD, a population-based survey on social contact patterns conducted in eight European countries[34]). These similarities in contact matrices imply that recruitment links may be representative for the social network links in our Dutch study population, at least with regard to age. However, young children (aged 0-12) and network links between these children and adults are missing in our country samples. We were unable to make a similar comparison of contact matrices for Thailand, as there are no Thai contact data available yet in the published literature. A large household-structured survey conducted in Vietnam[17] (a middle-income country, also densely populated in urban areas) demonstrated similar assortative mixing patterns by age as was seen in our Thai sample.

It is important to recognize the limitations of the presented data. Our results are based on small sample sizes and are not representative of the general population. In both countries we solely invited students of which the majority was aged between 20 and 30 years. The survey remained mainly within that age group, therefore limiting the generalisability of our results to other age groups. For the analyses of mixing patterns, the numbers of contact persons that were reported by recruiters and recruitees for one single day might not be a good representation of their contact frequency during the entire week. The data do contain responses for all days of the week, such that a comparison between weekdays and weekends is possible, but these are cross-sectional data and not longitudinal information per participant. We do not have information about individual variation in numbers of contacts during the week. In addition, in both countries only a small number of recruitment trees reached more than 4 waves, which may be a consequence of using solely a non-material incentive. The use of a small material incentive to stimulate online peer-recruitment should be further investigated as was done previously for hard-to-reach populations[35,36].

Although we chose to use an aggregated contact diary design to limit the burden for each participant and to stimulate online peer-recruitment, this prevented us from collecting other

determinants relevant for infection transmission, such as contact duration and intensity. Earlier studies have shown the importance of contact duration for understanding the transmission dynamics of close-contact pathogens[37,38]. Also, contact duration influences the probability that a contact is reported by a participant[39,40]. Although it is possible to derive contact durations from earlier conducted diary surveys, we cannot preclude the effect of heterogeneities in motivation or recall capabilities on the reported degrees, between both country samples, and between participants in same country (e.g., differences in reporting quality between females and males[39]).

The applied survey methodology was also unable to provide information on clustering of contact persons within small subgroups of the population. Theoretically it has been demonstrated that clustering of contacts influences transmission dynamics[5,6,41,42]. We did provide our participants with a separate link to indicate repeated invitations, but we received no reports of survey clustering. It remains challenging to motivate persons to undertake action in case they receive multiple invitations, as most persons instantly delete repetitive emails. In a next stage, we plan to experiment with a design in which contact persons need to be reported by name in order to measure clustering, like was done in some diary studies[39,40]. However, this introduces new privacy issues that may prevent participants from inviting contact persons, and thereby limit the possibility to study correlations between connected individuals.

The proportions of ILI symptoms found in our samples are broadly in agreement with estimates on the incidence of seasonal influenza, at least for the Netherlands. Previously, a study estimated that the overall symptomatic infection attack rate was 2.5% for seasonal influenza in the Netherlands[43]. If all symptomatically infected persons would be symptomatic at the same time, we would expect to find similar numbers in our sample. However, infections are spread over several months (see Figure S1) and therefore the 1.9% ILI cases found in the Dutch sample seems reasonable. This proportion seems even high when comparing the period of survey administration to the total period of the influenza season, this may indicate clustering of the influenza virus. There are several factors that influence the detection of symptoms via the applied RDS methodology, e.g., immunity in host populations can be clustered. Although natural immunity is hard to measure, clustered influenza vaccination patterns have been described in literature[44,45]. Also, participants were only asked to report any symptoms that they had in the past 7 days and it is possible that they experienced symptoms before this 7-days period or in the days after filling in the questionnaire. We are currently conducting new webRDS surveys in which we extended the 7-days period to 14 days, and added a follow-up measurement 3 weeks later to investigate whether participants developed any symptoms in the meantime.

Our findings from the comparison of data collected with RDS surveys in two countries provide

new insight on contacts and mixing patterns within social networks. Information on correlations between linked individuals can be used to improve parameterisation of mathematical models used to design optimal control strategies and lend support to the often used proportionate mixing assumption. Although the spread of infections through a population depends on more factors than just contacts and mixing patterns, similarities found between both countries suggest that the spread of directly transmitted respiratory infections may be similar in the Netherlands and Thailand, and in other countries with comparable contact patterns.

Acknowledgements

This study was conducted within the Utrecht Center for Infection Dynamics. We thank all survey participants in the Netherlands and Thailand. We are grateful to Martin Camitz (Sweden), Titan Tang (China) and Adam Ju (China) for programming the RDS survey system. Furthermore, we also thank Albert Wong from the Department of Statistics of the National Institute for Public Health and the Environment (RIVM) in the Netherlands for his assistance with R. Finally, we thank Wasamon Sabaiwan (Thailand) for her help with the Thai translations of the questionnaire and data collection in Thailand.

REFERENCES

1. 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.

2. May RM (2006) Network structure and the biology of populations. Trends Ecol Evol 21: 394-399.

3. Anderson RM, May RM (1991) Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press.

4. Newman ME (2002) Spread of epidemic disease on networks. Phys Rev E Stat Nonlin Soft Matter Phys 66: 016128.

5. Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: homogeneous and network models in epidemiology. J R Soc Interface 4: 879-891.

6. Volz EM, Miller JC, Galvani A, Ancel Meyers L (2011) Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol 7: e1002042.

7. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA (2012) Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect 140: 2117-2130.

8. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 107: 22020-22025.

9. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS One 5: e11596.

10. Newman MEJ, Girvan M (2002) Mixing patterns and community structure in networks. In: Pastor-Satorras R, Rubi J, Diaz-Guilera A, editors. Statistical Mechanics of Complex Networks. Berlin, 2003: Springer.

11. Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89: 208701.

12. Adamic LA, Lukose RM, Puniyani AR, Huberman BA (2001) Search in power-law networks. Phys Rev E Stat Nonlin Soft Matter Phys 64: 046135.

13. Heckathorn D (1997) Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems 44: 174-199.

14. Wejnert C, Heckathorn DD (2008) Web-Based Network Sampling: Efficiency and Efficancy of Respondent-Driven Sampling for Online Research. Sociological Methods and Research 37: 105–134.

15. Stein ML, Van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, et al. (2014) Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PLOS One 9: e85256.

16. 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: e74.

17. Horby P, Pham QT, Hens N, Nguyen TT, Le QM, et al. (2011) Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS One 6: e16965.

18. Fu YC, Wang DW, Chuang JH (2012) Representative contact diaries for modeling the spread of infectious diseases in Taiwan. PLoS One 7: e45113.

19. Kaewyeam K (1996) The current housing situation and demand for the students of King Mongkut's Institute of Technology Thonburi. Bangkok: Chulalongkorn University. 269 p.

20. ABF Research (2012) Landelijke monitor studentenhuisvesting 2012. Delft: KENCES.

21. Utrecht Center for Infection Dynamics (2013) YourSpace. Utrecht: Julius center for health sciences and primary care, University Medical Center Utrecht.

22. Salganik MJ, Heckathorn D (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34: 193-239.

23. Goel S, Salganik MJ (2009) Respondent-driven sampling as Markov chain Monte Carlo. Stat Med 28: 2202-2229.

24. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M (2008) Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol Infect 136: 813-822.

25. Centraal Bureau voor de Statistiek (2013) Huishoudens; grootte, samenstelling, positie in het huishouden, 1 januari. StatLine.

26. National Statistical Office Thailand (2010) The 2010 Population and Housing Census. Thailand: Ministry of Information and Communication Technology.

27. Socialbakers (2013) Checkfacebook: Facebook statistics of Thailand.

28.Nationale Atlas Volksgezondheid (2012) Bevolkingsdichtheid per gemeente 2012. 11 december 2012 ed. Bilthoven: National Institute for Public Health and the Environment (RIVM).

29. Danon L, House TA, Read JM, Keeling MJ (2012) Social encounter networks: collective properties and disease transmission. J R Soc Interface 9: 2826-2833.

30. Duerr HP, Schwehm M, Leary CC, De Vlas SJ, Eichner M (2007) The impact of contact structure on infectious disease control: influenza and antiviral agents. Epidemiol Infect 135: 1124-1132.

31. Farrington CP, Unkel S, Anaya-Izquierdo K (2013) Estimation of basic reproduction numbers: individual heterogeneity and robustness to perturbation of the contact function. Biostatistics 14: 528-540.

spread of infectious diseases in Taiwan. PLoS One 7: e45113.

32. Bolton KJ, McCaw JM, Forbes K, Nathan P, Robins G, et al. (2012) Influence of contact definitions in assessment of the relative importance of social settings in disease transmission risk. PLoS One 7: e30893.

33. Hanks LM (1962) Merit and Power in the Thai Social Order. American Anthropologist 64: 1247-1261.

34. Wejnert C (2010) Social Network Analysis with Respondent-Driven Sampling Data: A Study of Racial Integration on Campus. Soc Networks 32: 112-124.

35. Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, et al. (2012) Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. PLoS One 7: e49417.

36. Truong HH, Grasso M, Chen YH, Kellogg TA, Robertson T, et al. (2013) Balancing theory and practice in respondent-driven sampling: a case study of innovations developed to overcome recruitment challenges. PLoS One 8: e70344.

37. De Cao E, Zagheni E, Manfredi P, Melegaro A (2014) The relative importance of frequency of contacts and duration of exposure for the spread of directly transmitted infections. Biostatistics 15: 470-483.

38. Smieszek T (2009) A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. Theor Biol Med Model 6: 25.

39. Smieszek T, Barclay VC, Seeni I, Rainey JJ, Gao H, et al. (2014) How should social mixing be measured: comparing web-based survey and sensor-based methods. BMC Infect Dis 14: 136.

40. Smieszek T, Burri EU, Scherzinger R, Scholz RW (2012) Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiol Infect 140: 744-752.

41. Miller JC (2009) Spread of infectious disease through clustered populations. J R Soc Interface 6: 1121-1134.

42. Ball F, Britton T, Sirl D (2013) A network with tunable clustering, degree correlation and degree distribution, and an epidemic thereon. J Math Biol 66: 979-1019.

43. McDonald SA, Presanis AM, De Angelis D, van der Hoek W, Hooiveld M, et al. (2014) An evidence synthesis approach to estimating the incidence of seasonal influenza in the Netherlands. Influenza Other Respir Viruses 8: 33-41.

44. Barclay VC, Smieszek T, He J, Cao G, Rainey JJ, et al. (2014) Positive network assortativity of influenza vaccination at a high school: implications for outbreak risk and herd immunity. PLoS One 9: e87042.

45. Salathe M, Khandelwal S (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 7: e1002199.

46. National Influenza Centres, Global Influenza Surveillance and Response System (2014) FluNet. South-East Asia: World Health Organization.

47. Brandsema P, Dijkstra F, Euser SM, Van Gageldonk-Lafeber AB, de Lange MMA, et al. (2013) Jaarrapportage surveillance respiratoire infectieziekten 2012. Bilthoven: National Institute for Public Health and the Environment,. pp. 90.

SUPPLEMENTARY MATERIALS

Text S1. Drivers of online peer-recruitment: logistic regression.

We investigated which characteristics influence online peer-recruitment. We defined "intention to recruit" as a respondent that requested invitations for recruitees on the last survey page. To analyse which characteristics influence the intention to recruit (categorised as recruiter "requested" or "did not request" four invitations) we used logistic regression analyses, with intention to recruit as binary outcome. The log odds of the binary outcome was modeled as a linear combination of the variables degree (integer), age (integer), education (binary: higher or lower education), sex (binary) and household size (integer). We conducted this analysis for full samples obtained in the Netherlands and Thailand, as well as for the samples excluding seeds.

First, we investigated one by one whether the relations between the probability of requesting invitations (log odds = "LOG probability/1-probability") and the predictor variables (integer) were linear. We used "rcspline.plot" of the R package "Hmisc" to plot the estimated restricted cubic spline (RCS[1], with knots = 4) function, which relates a single variable - the "predictor"- to the outcome for a logistic model (see Figure S5). Knots were placed at fixed quantiles (0.05, 0.35, 0.65, 0.95) of the predictor's marginal distribution, ensuring that enough points are available in each interval and also guards against letting outliers overly influence knot placement[1]. We also used a Pearson's Chi-Squared test (with Yates' continuity correction) to analyse bivariate the independency between the outcome and the categorical predictors (Table I).

Table I. Chi-squared test to analyse independence by outcome variable.

		Chi-squared	df	p-value
The Netherlands	Sex	5.4153	1	0.01996
	Education	0	1	1
Thailand	Sex	0.0053	1	0.9421
	Education	0.0032	1	0.9549

Figure S5 illustrates for both countries that the relations between the log odds of intention to recruit and the predictor variables age, degree (log transformed) and household were not linear. Log-transformation or other similar transformations did not result in linear relations. We therefore added age, degree (log) and household size with RCS (with knots = 3, as this resulted in a logistic model with the lowest Akaike's information criterion, AIC) to the logistic regression model. Table 1 shows for the Dutch sample that the outcome "intention to recruit" is dependent on the variable "sex". In order to compare both country samples and to analyse the relationship between degree and the intention to recruit (e.g., do individuals with a high

degree have a higher probability to invite their recruitees compared to individuals with a low degree?), we used a logistic regression model containing all "predictor" variables (i.e., a full model).

Table II. Output of logistic regression for the Netherlands, full sample

	estimate[a]	SE	z value	Pr(>|z|)	2.50%	97.50%
Constant	-1.009	1.292	-0.781	0.435	-3.542	1.524
rcs degree(log)_S0	0.230	0.221	1.039	0.299	-0.204	0.664
rcs degree(log)_S1	-0.274	0.271	-1.013	0.311	-0.805	0.256
rcs age_S0	0.054	0.043	1.242	0.214	-0.031	0.138
rcs age_S1	-0.149	0.114	-1.300	0.194	-0.373	0.076
higher education	-0.216	0.415	-0.521	0.603	-1.030	0.598
male	-0.629	0.241	-2.613	***0.009***	-1.100	-0.157
rcs household size_S0	0.024	0.193	0.123	0.902	-0.355	0.403
rcs household size_S1	-0.028	0.291	-0.097	0.923	-0.598	0.542

[a]Null deviance: 434.91 (df: 324); residual variance: 425.18 (df: 316) and AIC 443.18.

Table II shows the regression coefficients for the full Dutch sample. "rcs" indicate the cubic spline terms that we added to the model (k-1 regression parameters, not including the intercept). An RCS function with three knots includes two splines: S0 (a linear part) and S1 (a non-linear part). Estimates of S>0 are virtually not interpretable[2].

Table III. Output of logistic regression for the Netherlands, sample without seeds

	estimate[a]	SE	z value	Pr(>|z|)	2.50%	97.50%
Constant	-2.255	1.398	-1.613	0.107	-4.994	0.485
rcs degree(log)_S0	0.155	0.270	0.574	0.566	-0.374	0.685
rcs degree(log)_S1	-0.232	0.366	-0.634	0.526	-0.949	0.485
rcs age_S0	0.089	0.046	1.912	0.056	-0.002	0.180
rcs age_S1	-0.216	0.131	-1.651	0.099	-0.473	0.040
higher education	-0.443	0.441	-1.004	0.315	-1.307	0.421
male	-0.625	0.279	-2.240	***0.025***	-1.171	-0.078
rcs household size_S0	0.089	0.219	0.408	0.683	-0.340	0.519
rcs household size_S1	-0.155	0.336	-0.461	0.645	-0.814	0.504

[a]Null deviance: 327.17 (df: 235); residual variance: 316.36 (df: 227) and AIC 334.36.

In the Dutch sample, males had a significantly lower log odds (-0.629, p=0.009) of intending to invite recruitees compared to female participants when adjusted for degree, age, education and household size. We obtained similar results for the Dutch sample without seeds, see Table III. Although in the sample without seeds also age seemed to slighty influence the intention

to recruit, besides sex. Table IV and V displays the regression coefficients for respectively the full Thai sample and the Thai sample without seeds. In the Thai sample, participants' age and degree seemed to influence the intention to recruit.

Table IV. Output of logistic regression for Thailand, full sample

	estimate[a]	SE	z value	Pr(>|z|)	2.50%	97.50%
Constant	-4.122	1.906	-2.162	0.031	-7.859	-0.385
rcs degree(log)_S0	0.645	0.281	2.299	***0.022***	0.095	1.195
rcs degree(log)_S1	-1.189	0.441	-2.697	***0.007***	-2.053	-0.325
rcs age_S0	0.135	0.075	1.804	0.071	-0.012	0.281
rcs age_S1	-0.121	0.093	-1.301	0.193	-0.303	0.061
higher education	-0.738	0.513	-1.439	0.150	-1.744	0.267
male	0.150	0.318	0.471	0.637	-0.474	0.774
rcs household size_S0	0.131	0.157	0.833	0.405	-0.177	0.439
rcs household size_S1	-0.008	0.153	-0.051	0.960	-0.307	0.292

[a]Null deviance: 278.50 (df: 216); residual variance: 262.11 (df: 208) and AIC 280.11

Table V. Output of logistic regression for Thailand, sample without seeds

	estimate[a]	SE	z value	Pr(>|z|)	2.50%	97.50%
Constant	-6.236	2.511	-2.483	0.013	-11.157	-1.314
rcs degree(log)_S0	0.566	0.335	1.688	0.091	-0.091	1.224
rcs degree(log)_S1	-0.974	0.541	-1.801	0.072	-2.035	0.086
rcs age_S0	0.231	0.107	2.155	***0.031***	0.021	0.440
rcs age_S1	-0.238	0.136	-1.742	0.081	-0.505	0.030
higher education	-0.898	0.710	-1.264	0.206	-2.289	0.494
male	0.089	0.391	0.228	0.819	-0.678	0.856
rcs household size_S0	0.036	0.180	0.199	0.842	-0.318	0.390
rcs household size_S1	0.061	0.168	0.364	0.716	-0.268	0.390

[a]Null deviance: 186.69 (df: 136); residual variance: 172.35 (df: 128) and AIC 190.35.

We used the fitted logistic regression model to estimate probabilities of 'intention to recruit' for individual subjects as a function of degree. Thus, probabilities were estimated based on actual obtained data ('data-driven'); not by keeping the other predictors 'constant' by using means. Confidence intervals (95%) were obtained using standard errors. Lines and confidence intervals in Figure 3d were smoothed using 'geom_smooth' of the R-package "ggplot2".

Literature

1. Harrell F.E. 2001 Restricted Cubic Splines. In Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis (ed. Springer), pp. 20-26. New York, Springer-Verlag New York.
2. Desquilbet L., Mariotti F. 2010 Dose-response analyses using restricted cubic spline functions in public health research. Statistics in medicine 29(9), 1037-1057.

Text S2. R source code.

The R source code can be found online as supplementary file to the article ML Stein et al. PLoS ONE 9(11): e113711.

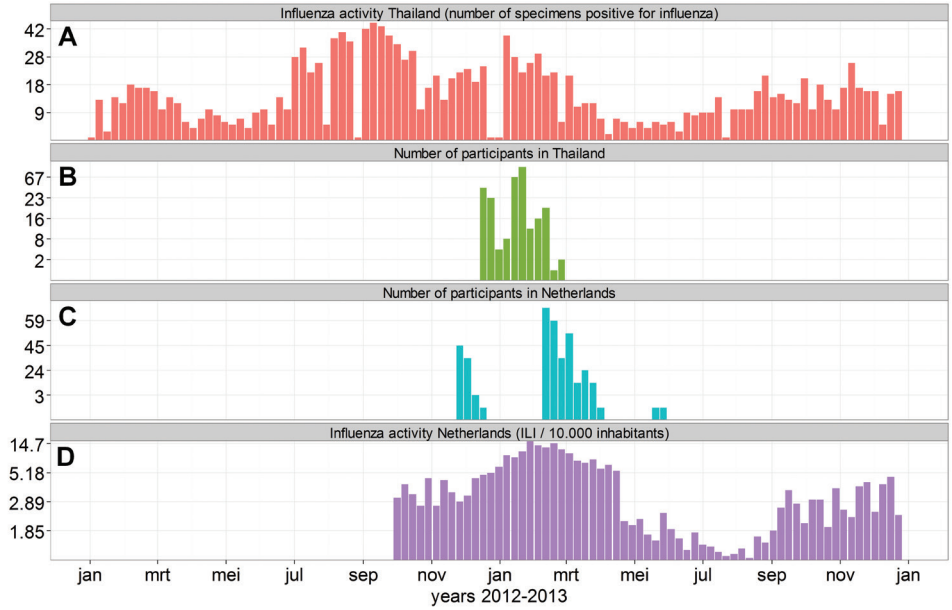


Figure S1. Period of survey administration and influenza activity in each country. (**A**) displays the influenza activity in Thailand in numbers of specimens that were found positive for influenza (as provided by FluNet[46]). (**B**) Number of participants in Thailand collected between December 2012 and March 2013. (**C**) Number of participants in the Netherlands collected between end of November 2012 and beginning of May 2013. Most participants filled in the questionnaire before end of March 2013. Two participants filled in the questionnaire end of May 2013, after being invited by other participants. (**D**) displays the influenza activity in the Netherlands in number of persons per 10000 inhabitants that visited the general practioner with influenza-like-illness (ILI) (as provided by NIVEL[47]).

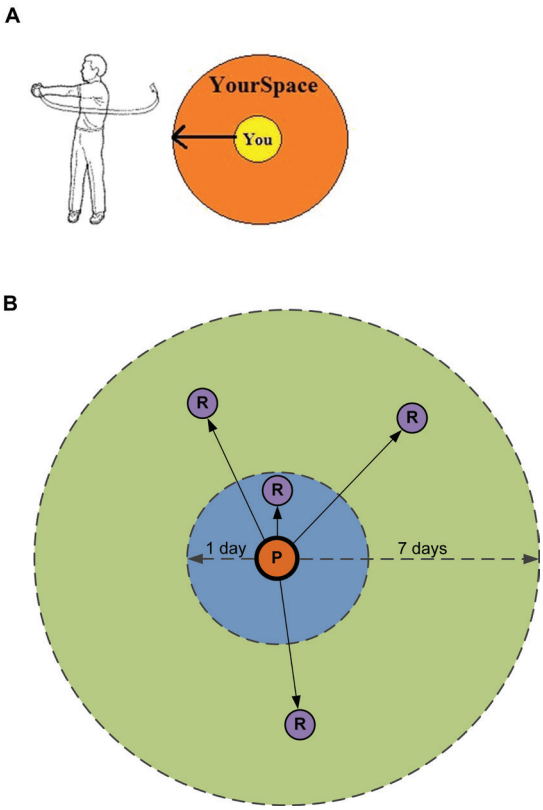


Figure S2. Graphical illustrations of contact definition and RDS recruitment.

(**A**) Figure illustrating contact definition: a person sitting or standing within arm's length of the participant, which was denoted as "YourSpace", for 30 seconds or longer. This figure was displayed in the online questionnaire to clarify to participants who they had to count as a contact person.

(**B**) Figure illustrating difference between 'contact persons' and 'recruitees'. We asked a participant (P) to invite four recruitees (R) who he/she had met according to the contact definition (within 'YourSpace') in the past 7 days. The blue circle illustrates 'contact persons' met by the participant 1 day before the day of filling in the questionnaire ('yesterday'); these contact persons were recorded in the questionnaire. The green circle illustrates persons met 2-7 days before the day of filling in the questionnaire; we did not collect information on these persons. Participants could either have met recruitees 'yesterday' or 2-7 days before the participation day.

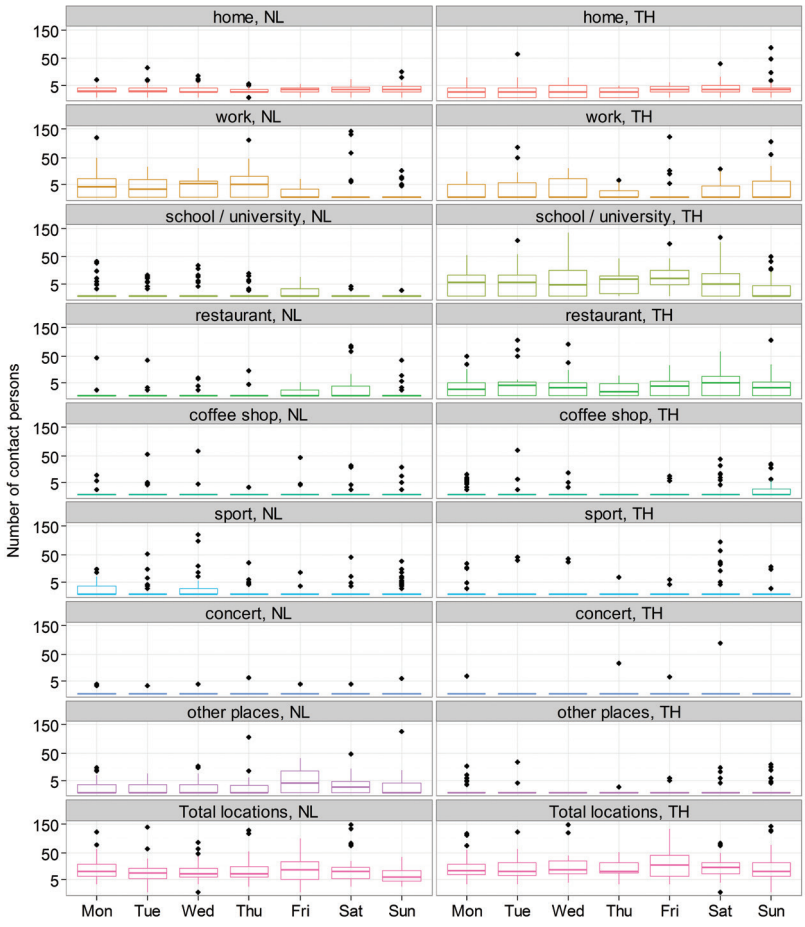


Figure S3. Distributions of numbers of contact persons across days of the week by each location and all locations together.

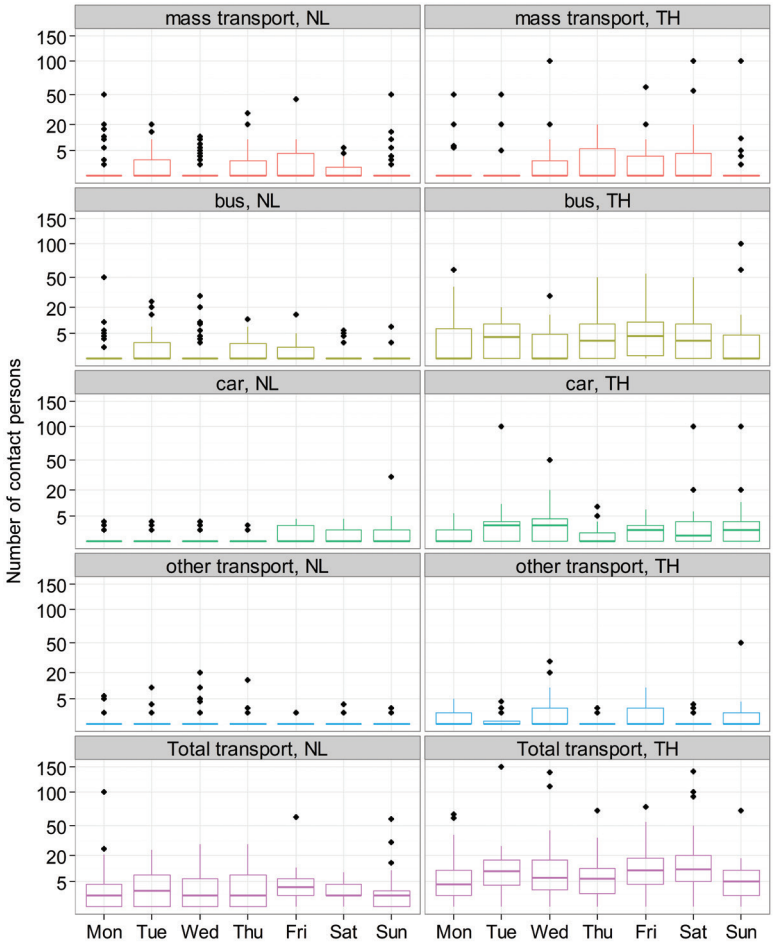


Figure S4. Distributions of numbers of contact persons across days of the week by transport vehicle and all transport vehicles together.

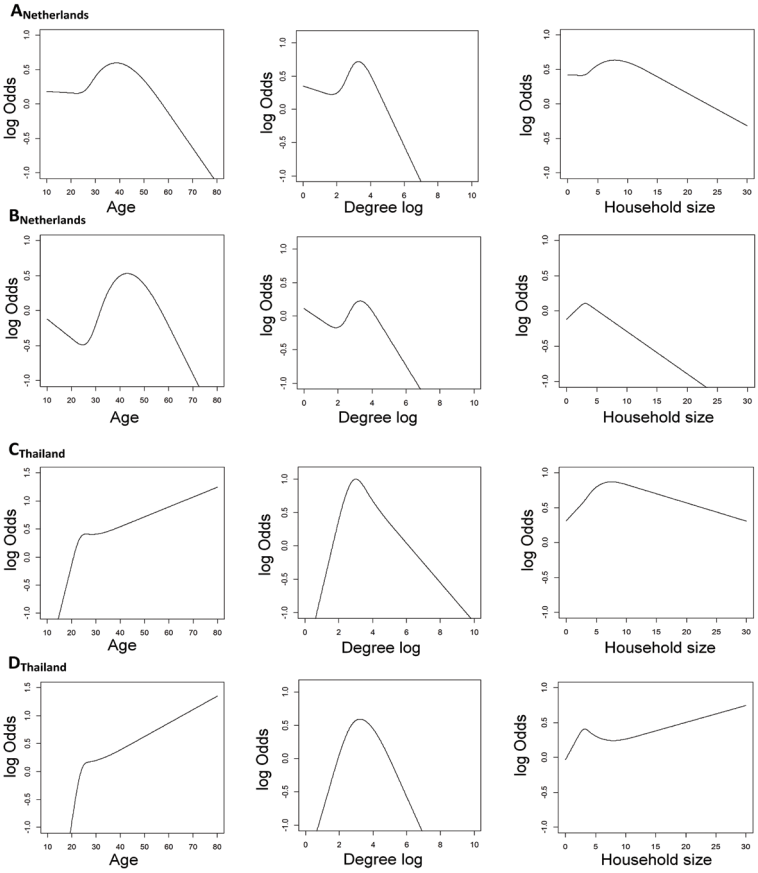


Figure S5 Investigating the relation between the outcome and the independent (integer) variables. The plots display the relation between the outcome and age, degree and household size. Plots **A** are based on the full Dutch sample ($n_{Netherlands}$=356); plots **B** are based on the Dutch sample without seeds (data without wave 0, $n_{Netherlands}$=264). Plots **C** are based on the full Thai sample ($n_{Thailand}$=251); plots **D** on the Thai sample without seeds ($n_{Thailand}$=163).

Table S1. Subset of transport and location type of interest as shown in questionnaire.

	Displayed for participants in the Netherlands	Displayed for participants in Thailand
Locations	at home	at home
	at work	at work
	at school / university	at school / university
	in a restaurant	in a restaurant
	in a coffee shop	in a coffee shop
	at sports / leisure	at sports / leisure
	at a concert / theater / cinema	at a concert / theater / cinema
	at other places	at other places
Transport vehicles	Train and/or airplane	Skytrain and/or subway and/or airplane
	Bus and/or metro and/or tram	Bus and/or minibus and/or shuttle boot
	Car	Car and/or taxi
	Other transport vehicle	Motorbike and/or Tuk-Tuk

Table S2. Symptoms displayed in questionnaire.

Have you had any of the following complaints in the past week?
No complaints
Fever
Chills
Runny or blocked nose
Sore throat
Cough
Headache
Muscle/joint pain
Diarrhea
Other, namely

Table S3. Exploring the first-order Markov assumption: correlations in country samples.

	Geodesic distance[a]	0 – 1	1 – 2	2 – 3	0 – 3
Netherlands	age (r)	0.295 [0.131–0.443]*	0.614 [0.455–0.734]*	0.473 [0.160–0.699][£]	0.171 [-0.177–0.481]
	gender (r_φ)	0.227 [0.059–0.383][£]	0.382 [0.177–0.555]*	0.336 [-0.002–0.606]	0.106 [-0.241–0.428]
Thailand	age (r)	0.524 [0.350–0.663]*	0.621 [0.394–0.777]*	0.451 [0.011 – 0.745][£]	0.320 [-0.143–0.668]
	gender (r_φ)	0.192 [-0.022–0.389]	0.347 [0.052–0.586][£]	0.192 [-0.273 – 0.585]	0.000 [-0.443–0.443]

[a] Link distance between seeds in wave 0 and contact persons in consecutive waves (waves 1 to 3). * p-value < 0.001 [£] p-value < 0.05

Chapter 4

Tracking social contact networks with online respondent-driven detection: who recruits whom?

Mart L. Stein[1,2*], Peter G.M. van der Heijden[3,4], Vincent Buskens[5], Jim E. van Steenbergen[2,6], Linus Bengtsson[7,8], Carl E. Koppeschaar[9], Anna E. Thorson[7] and Mirjam E.E. Kretzschmar[1,2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[3] Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, the Netherlands

[4] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom

[5] Department of Sociology, Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, the Netherlands

[6] Centre of Infectious Diseases, Leiden University Medical Centre, Leiden, the Netherlands

[7] Department of Public Health Sciences-Global Health, Karolinska Institutet, Stockholm, Sweden

[8] Flowminder Foundation, Stockholm, Sweden

[9] Science in Action BV, Amsterdam, the Netherlands

BMC Infectious Diseases 2015, 15:522

ABSTRACT

Background

Transmission of respiratory pathogens in a population depends on the contact network patterns of individuals. To accurately understand and explain epidemic behaviour information on contact networks is required, but only limited empirical data is available. Online respondent-driven detection can provide relevant epidemiological data on numbers of contact persons and dynamics of contacts between pairs of individuals. We aimed to analyse contact networks with respect to sociodemographic and geographical characteristics, vaccine-induced immunity and self-reported symptoms.

Methods

In 2014, volunteers from two large participatory surveillance panels in the Netherlands and Belgium were invited for a survey. Participants were asked to record numbers of contacts at different locations and self-reported influenza-like-illness symptoms, and to invite 4 individuals they had met face to face in the preceding 2 weeks. We calculated correlations between linked individuals to investigate mixing patterns.

Results

In total 1560 individuals completed the survey who reported in total 30,591 contact persons; 488 recruiter-recruit pairs were analysed. Recruitment was assortative by age, education, household size, influenza vaccination status and sentiments, indicating that participants tended to recruit contact persons similar to themselves. We also found assortative recruitment by symptoms, reaffirming our objective of sampling contact persons whom a participant may infect or by whom a participant may get infected in case of an outbreak. Recruitment was random by sex and numbers of contact persons. Relationships between pairs were influenced by the spatial distribution of peer recruitment.

Conclusions

Although complex mechanisms influence online peer recruitment, the observed statistical relationships reflected the observed contact network patterns in the general population relevant for the transmission of respiratory pathogens. This provides useful and innovative input for predictive epidemic models relying on network information.

BACKGROUND

For infectious diseases, such as influenza, severe acute respiratory syndrome and measles, proximity and social contact between individuals are major factors for person-to-person transmission. Knowledge on contact patterns is therefore important for the design of optimal outbreak control strategies[1-4]. To accurately understand and explain epidemic dynamics, information is required on the underlying contact network of a host population, i.e., a network that contains all contact persons potentially at risk for infection. For example, the number of contacts an infectious individual has with susceptible persons determines among others the basic reproduction number R0 (i.e., the number of secondary cases one case generates in an susceptible population)[5].

Contact networks are complex and highly dynamic (i.e., not constant over time)[6]. Previous empirical studies of contact patterns used different methods of data collection, including direct observation, contact diaries and electronic proximity sensors, to quantify social mixing behaviour for a variety of populations[7, 8]. For example, the POLYMOD study, a large randomized study in eight European countries, used contact diaries to analyse mixing patterns of independent respondents[9]. Despite controversies on the different modes of transmission of respiratory infectious diseases[10], face-to-face conversations and physical contact are often used as proxies for potential infectious contacts[9,11]. Close contact persons such as family, friends and colleagues are thereby assumed to capture the majority of contacts for potential transmission events[12].

A social network approach can provide relevant epidemiological data on numbers of contacts and the strength and dynamics of contacts between pairs of individuals in a population[13, 14]. Respondent-driven detection, a method of detection derived from snowball sampling, is a chain recruitment method that allows for systematic sampling of contact persons of participants. Previously, we demonstrated that under certain conditions such a recruitment method can be applied online to extract topological properties of contact networks in an anonymous manner[15, 16]. This approach provides novel insights in contact network structures compared to studies that sampled participants independently of one another and collected no information about the network beyond the contact persons reported by participants[7]. In these earlier studies 'seed' individuals of similar age groups and backgrounds were sampled at convenience[15, 16]. Furthermore, complex mechanisms may play a role when participants choose from amongst their contact persons and when contact persons decide whether to join the survey[11]. For example, with an offline (i.e., paper based) chain recruitment method participants have a tendency to recruit spatially proximal peers[17]. This determines the type of contact networks being sampled. Note that we distinguish respondent-driven detection from respondent-driven sampling as our main objective was to study contact networks, and not to

estimate population proportions from the sample.

Earlier we reported on a study in which we combined online respondent-driven detection with participatory surveillance, i.e., an Internet-based system that captures voluntarily submitted data on influenza-like-illness (ILI) symptoms from the general public[18]. We showed that such respondent-driven approach can be used to improve the detection of symptomatic cases[19]. In this paper we were interested in the contact networks of respondents and the association with self-reported disease. In particular, we aimed to determine correlations between participants linked by recruitment chains (i.e., who recruits whom) with respect to sociodemographic characteristics, vaccine-induced immunity and self-reported symptoms. In addition, we investigated the effect of spatial peer recruitment on these correlations. If recruitment of contact persons by participants is random, these statistical relationships reflect the underlying contact networks in the general population that are relevant for the transmission of respiratory pathogens.

METHODS

Study design

Volunteers of two participatory surveillance panels were invited via the organizations' electronic newsletters for an online and anonymous survey between November 2013 and May 2014. The first panel focused on ILI, operates in the Netherlands and Dutch speaking Flanders (Belgium), and had 16942 active volunteers. The second panel focused on pneumonia, operates only in the Netherlands, and had 1691 active volunteers. After completion of the questionnaire, participants were asked to recruit 4 contact persons (e.g., family members, friends, acquaintances) whom they had met face to face in the past 2 weeks. Invited contact persons were asked to do the same. Online peer recruitment was followed by means of unique codes that were automatically generated. Participants could invite contact persons via standard email, via a private message on Facebook, or by sharing a unique link (i.e., a Uniform Resource Locator that includes a personal code). A 'seed' indicates a volunteer from the surveillance panels who participated in our survey and a 'recruit' is a contact person recruited by a survey participant. By 'waves' we refer to consecutive subsamples, with seeds in wave 0, recruits invited by seeds in wave 1, and so forth. 'Recruitment trees' refers to chains of participants connected via recruitment. Invited contact persons could opt-out of the survey and provide reasons for not participating.

After completion of the questionnaire, participants were referred to a research website that displayed the latest results (e.g., anonymous network trees). Participants recruited via the first panel who completed the survey had the opportunity to join a raffle for 1 of 10 gift cards of €25. This incentive only slightly increased peer recruitment as was shown in Stein et al.[19] For

details on the software system and information on the 171 non-responders we also refer the reader to Stein et al.[19].

We obtained ethical approval from the Medical Ethical Committee of the University Medical Center Utrecht, the Netherlands (13-664/C). Informed consent was obtained before survey participation.

Questionnaire

We defined 'contact' as touching a person (e.g., shaking hands or hugging) or talking to a person within a distance of about one arm's length (duration of conversation did not matter). Participants were asked to report as precisely as possible the number of contact persons that they had during one full day ('yesterday') at four predefined locations, namely at home, at work or educational institute (school or university), at the house of family or friends or other acquaintances, and at other places (e.g., during sports, shopping or travelling, or in a restaurant or cafe). Participants were asked to specify the age group of the contact person (namely 0–11 years, 12–18 years, 19–60 years and older than 60 years); multiple contacts with the same person during the course of the day needed to be counted only once. 'Degree' denotes the total number of contact persons reported by a participant.

Participants were asked to report any symptoms (provided in a list) that they had experienced in the past 2 weeks. If any symptoms were reported, we asked additional disease related questions and whether they knew any contact persons with similar symptoms. Symptomatic participants were asked about the type of disease that they thought to have experienced (e.g., influenza or common cold); we further refer to this as self-reported influenza or common cold. We used the definition of the World Health Organization to define ILI that includes having fever (excluding questions on a body temperature of ≥ 38°C) and cough with an onset within the last 10 days. Participants were also asked whether they had received an invitation to get an influenza vaccination and whether they had received influenza vaccination in the past 12 months. This information was used as a proxy for the possible immune status of participants. As earlier studies described clustered patterns of influenza vaccination uptake and sentiments concerning vaccination, we asked participants whether they believed that the influenza vaccine protects them against influenza[20, 21]. Lastly, for each participant we collected information on age, sex, educational level, household members and their age, four digit postal code, and work or study location. Parents could fill in the questionnaire for their child.

Statistical analysis

First we assessed the main effects of covariates (age, sex, household size and ILI) on degree using a Poisson Inverse-Gaussian regression model (see also Additional file). This model is an alternative to a negative binomial model and has the potential for modelling highly dispersed

count data due to the flexibility of the Inverse Gaussian distribution[22, 23].

We investigated mixing patterns within our sample by analysing shortest paths between pairs of any two individuals that were one, two, or three or more link steps away from each other in the same recruitment tree[24]. Correlation coefficients with respect to the same measured variable were calculated for pairs of recruiter and recruit in consecutive waves. Pearson's *r* was used for integer variables (age, degree and household size), phi coefficient (r_{φ}) for binary variables (sex, vaccination status, symptoms) and Spearman rank-order (r_{rank}) for ordinal variables (education, vaccination beliefs). These correlations provide both insight in recruitment patterns, as well as in clustering (i.e., contact persons of an individual with the same characteristic(s) are recruited or infected with a probability that is higher than expected if the distribution was random) of disease, vaccination status and sentiments.

We compared the sampled recruiter-recruit age matrix with the participant-contact age matrix collected in the Netherlands during POLYMOD (Van de Kassteele J, Van Eijkeren J, Wallinga J: Efficient estimation of age-specific social contact rates between men and women, in preparation)[9]. If we assume that POLYMOD data accurately reflects all contact persons of an individual, then by a comparison we can investigate to what extent recruitment links between two participants can be interpreted as a contact in the sense of our contact definition, at east with respect to age. Firstly, we used the two-sample Kolmogorov-Smirnov (KS) test to compare column wise for each participant's age group the (integer) age distribution of recruits sampled in our study, with those of contact persons recorded in POLYMOD. Secondly, we used a homogeneous uniform association model (i.e., a model that assumes that all strata in two-way contingency tables have a common local odds ratio, OR) to test whether there is a statistical difference between both entire matrices[25–27].

To analyse the spatial spread of recruitment we converted the registered 4-digit postal codes into coordinates using geocoding and computed the distance between a recruiter and their recruit with the great-circle distance. We also computed the distance a participant commutes between home and the work or study location. We investigated the co-occurrence of a characteristic separately for recruiter-recruit pairs that had the same postal code, and between pairs that lived 1 to 10 km and more than 10 km away from each other. The equality of correlation coefficients, calculated for integer variables, was tested using Fisher z-transformation[28]. The equality of odds ratios, calculated for binary variables, was tested using a log-linear model. Finally, we used a logistic regression model to estimate for individuals living in four different regions in the Netherlands the probability of recruiting a contact person at the work or study location (see also Additional file 1). Statistical analyses were performed in R (version 3.1.1).

RESULTS

Description of sample

A total of 1560 individuals completed the survey at least once, of which 1105 seeds (wave 0) who were invited via the panels, and 455 recruits (waves 1 to 6) who were invited by participants. Neither participatory surveillance panel was representative of the general population in terms of basic demographic characteristics. However, through peer recruitment the sample representativeness slightly improved in terms of age and sex (see also Stein et al.[19]). Overall, 64.7% of the participants were female, 55.5% were aged between 50–69 years (mean age: 53.6; range: 3–97 years), 57.4% obtained a bachelor degree or higher, 41.5% had a two-person household and 41.9% received an influenza vaccine in the past 12 months (Table 1). Less than half of all seeds (45.8%) reported at least one symptom, while more than half of the recruits (on average 57.8% in waves 1 to 6) reported symptoms. Of all participants, 8.3% self-reported they had influenza of which 32.3% had received the influenza vaccine, resulting in an OR of 0.64 (95% confidence interval (CI) 0.42–0.95) for self-reported influenza by vaccinated individuals (compared to non-vaccinated).

Reported contact persons

A total of 30591 contact persons were reported by 1531 participants, with a mean degree of 19.6 per participant (median: 11.0; standard deviation (SD): 35.3). Twenty-nine participants reported zero contact persons. Figure 1A displays the sampled degree distribution, which showed strong over-dispersion. A Poisson Inverse-Gaussian distribution with mean μ = 19.6 (95% CI 18.3–21.1) and dispersion parameter λ = 2.0 (95% CI 1.8–2.1) best fitted the empirical degree distribution. Analysis of degree with a multiple regression model showed a lower contact frequency for those aged ≥ 65 years compared to participants between 0 and 39 years old (Table 2). A larger household size was associated with a higher number of contact persons. Participants with ILI had less contact persons than persons without these symptoms. Such reduction in numbers of contacts has also been observed among ILI cases during the 2009 influenza epidemic and may be explained by people staying at home and avoiding social activities when ill[30]. Weekdays were associated with 33%–84% more contact persons than Sundays (see also Additional file for the distribution of contact persons by days of the week), which is in accordance with results from other studies on contact patterns[9, 30].

Recruitment trees

Figure 1B shows the size of 1105 recruitment trees. Most recruitment trees consisted of only one node (i.e., seeds who did not recruit contact persons). There were 206 recruitment trees with at least two nodes (i.e., trees with at least two participants and one recruitment wave), and two of these trees consisted of 11 nodes each. One recruitment tree reached 6 waves of recruits. The majority of the recruits responded the same day they were invited by their

Table 1. Sample characteristics overall and per recruitment wave.

		Wave 0 (n: 1105)		Wave 1 (n: 310)		Wave 2 (n: 93)		Waves 3-6 (n: 52)		Total (n: 1560)	
		n	%	n	%	n	%	n	%	n	%
Country	Netherlands	1018	92.1	295	95.2	86	92.5	52	100	1451	93.0
	Belgium	87	7.9	15[a]	4.8	7	7.5	0	0	109	7.0
Sex	Male	387	35.0	122	39.4	31	33.3	10	19.2	550	35.3
	Female	718	65.0	188	60.6	62	66.7	42	80.8	1010	64.7
Age[b]	0-39	139	12.5	91	29.3	26	28.0	13	25.0	268	17.2
	40-49	189	17.1	43	13.9	18	19.3	6	11.5	256	16.4
	50-64	496	44.9	106	34.2	32	34.4	22	42.3	656	42.1
	65+	281	25.5	70	22.6	17	18.3	11	21.2	379	24.3
Education	Bachelor or higher	651	58.9	166	53.5	56	60.2	23	44.2	896	57.4
	Lower than bachelor	144	41.1	29	46.5	37	39.8	29	55.8	664	42.6
Household[c]	1-person	280	25.3	78	25.2	22	23.7	10	19.2	390	25.0
	2-persons	478	43.3	110	35.5	38	40.9	22	42.3	648	41.5
	3-persons	145	13.1	35	11.3	6	6.4	6	11.5	192	12.3
	4 or more	202	18.3	87	28.0	27	29.0	14	26.9	330	21.2
Work or Study	yes	775	70.1	228	73.5	73	78.5	41	78.8	1117	71.6
	no	330	29.9	82	26.5	20	21.5	11	21.2	443	28.4
Vaccinated[d]	yes	516	46.7	104	33.5	19	20.4	15	28.8	654	41.9
	no	589	53.3	206	66.5	74	79.6	37	71.2	906	58.1
Symptoms	yes	506	45.8	172	55.5	56	60.2	35	68.3	769	49.3
	no	599	54.2	138	44.5	37	39.8	17	32.7	791	50.7
Self-reported common cold	yes	175	15.8	60	19.4	27	29.0	10	19.2	272	17.4
	no	930	84.2	250	80.6	66	71.0	42	80.8	1288	82.6
Self-reported influenza	yes	96	8.7	24	7.7	7	7.5	3	5.8	130	8.3
	no	1009	91.3	286	92.3	86	92.5	49	94.2	1430	91.7
ILI	yes	34	3.1	2	0.6	2	2.2	2	3.8	40	2.6
	no	1071	96.9	308	99.4	91	97.8	50	96.2	1520	97.4

[a]One participant lived in Germany
[b]One participant provided an invalid age
[c]Note: 48 participants who completed the survey did not provide information on their household size and were assumed to live alone
[d]Vaccinated against influenza in the past 12 months

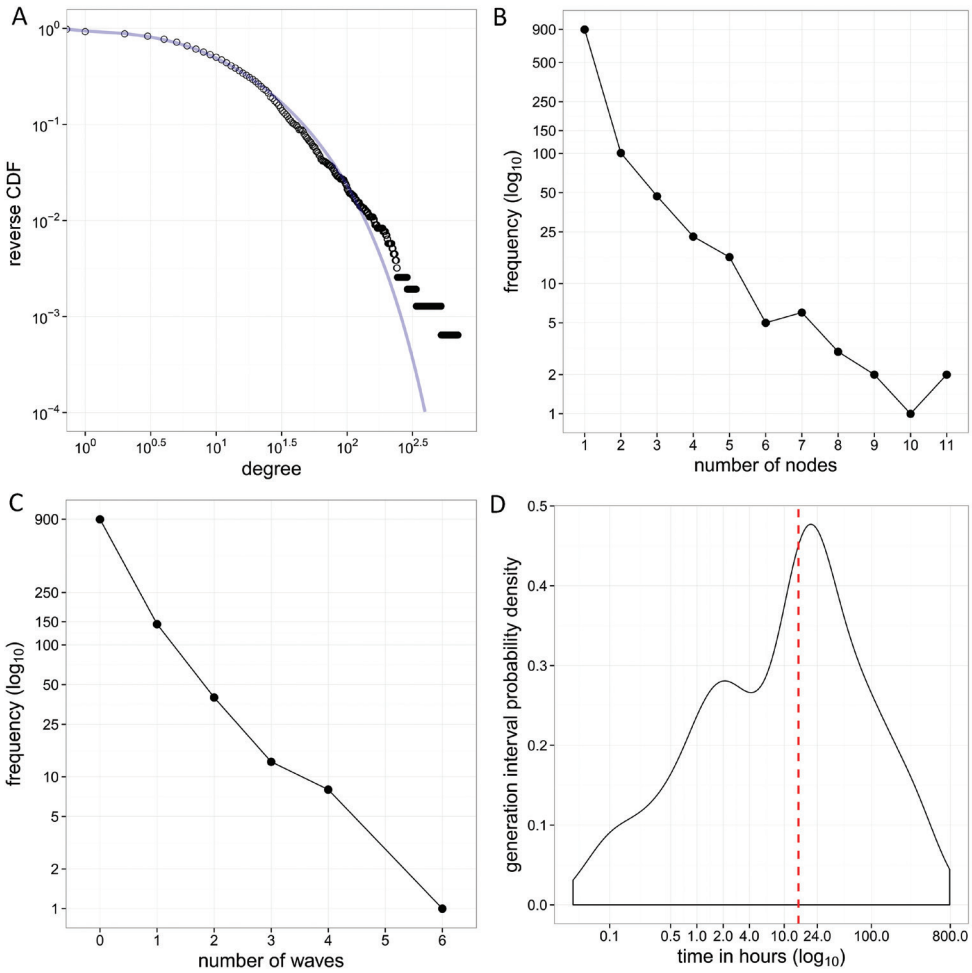


Figure 1 Reported contact persons and recruitment trees. (A) The empirical reversed cumulative distribution of degree (number of contact persons per participant) is indicated with black circles. The line is the fitted theoretical Poisson inverse-Gaussian distribution with mean μ: 19.6 (95 % CI 18.3–21.1) and dispersion parameter λ: 2.0 (95 % CI 1.8–2.1). **(B)** Number of participants (nodes) per recruitment tree. Most recruitment 'trees' only consisted of one participant (the seed), two trees consisted of 11 participants. **(C)** Number of waves that recruitment trees reached by peer recruitment, with seeds in wave 0. One recruitment tree reached 6 waves of recruits. **(D)** Recruitment generation interval. Red line indicates median generation interval

recruiter, giving a median generation interval (i.e., the time between invitation by a recruiter and participation by his/her recruit) of 14.6 hours (mean: 50.7; SD: 100.0) (Figure 1D). Overall, the larger the proportion of women or individuals with a bachelor's degree or higher in a recruitment tree, the larger the tree size was on average. Seed characteristics did not appear to influence the number of nodes in a recruitment tree (see also Additional file).

Table 2. Number of reported contact persons per participant per day by different characteristics and relative number of contacts from the Poisson inverse-Gaussian Regression model.

Category	Covariate	Number of participants	Mean (standard deviation) of number of reported contacts	Relative number of reported contacts (95% CI)[a]
Age of participant	0-39	268	20.98 (24.88)	1.00
	40-49	256	25.35 (37.24)	0.97 (0.80-1.17)
	50-64	656	19.94 (35.16)	0.93 (0.79-1.09)
	65+	379	14.19 (39.63)	0.69 (0.58-0.83)
Sex of participant	Female	1010	18.94 (30.78)	1.00
	Male	549	20.83 (42.41)	1.05 (0.94-1.18)
Household size	1	389	17.85 (29.49)	1.00
	2	648	15.73 (23.91)	1.02 (0.89 -1.17)
	3	192	26.54 (58.17)	1.44 (1.20-1.73)
	4	218	24.93 (43.10)	1.55 (1.29-1.87)
	≥5	112	25.92 (37.37)	1.81 (1.43-2.29)
ILI	No	1519	19.93 (35.68)	1.00
	Yes	40	7.25 (9.70)	0.37 (0.25-0.53)
Days of the week	Sunday	224	16.68 (51.25)	1.00
	Monday	414	17.94 (32.15)	1.33 (1.12-1.59)
	Tuesday	249	24.27 (36.80)	1.84 (1.52-2.23)
	Wednesday	192	22.41 (31.73)	1.60 (1.30-1.96)
	Thursday	182	21.16 (28.29)	1.61 (1.31-1.99)
	Friday	117	18.76 (28.11)	1.42 (1.12-1.81)
	Saturday	181	16.65 (29.16)	1.27 (1.03-1.57)

[a]Dispersion parameter λ = 1.7 (95 % CI 1.4–2.1). The Poisson Inverse-Gaussian model is appropriate for modelling correlated counts with long sparse extended tails. The over-dispersion parameter in the model was significantly different from zero, indicating the necessity to use this model instead of a generalised Poisson model. Comparing AIC statistics, the Poisson Inverse-Gaussian model gave a better fit as opposed to a negative binomial model and a generalised Poissonmodel[22].

Recruitment mixing patterns

Overall, we obtained 455 pairs between a recruiter and his/her recruit whereby both participants completed the survey. For an additional 33 pairs we solely obtained basic demographic information.

We observed assortative recruitment patterns by age (r = 0.36 [95% CI 0.28-0.44]), education (r_{rank} = 0.31 [95% CI 0.23-0.40]) and household size (r = 0.22 [95% CI 0.13-0.30]), indicating that participants tend to recruit contact persons similar to themselves (Table 3). Recruitment was random (i.e., not assortative, nor disassortative) by sex (r_{φ} = 0.07 [95% CI -0.02-0.16]) and degree (r= 0.07 [95% CI-0.03-0.16]).

Table 3. Homophily in network components for different link steps.

	Variables (type of correlation coefficient)	1 link step[a]	p value	2 link steps[a]	p value	3-6 link steps (lumped together)[a]	p value
Type of contact network	Age (r)	0.36 [0.28–0.44]	<0.001 (df: 486)	0.13 [-0.03–0.28]	0.109 (df: 156)	0.23 [-0.01–0.43]	0.058 (df: 70)
	Sex (r_{φ})	0.07 [-0.02–0.16]	0.107 (df: 486)	0.25 [0.09–0.39]	0.002 (df: 156)	0.17 [-0.07–0.38]	0.167 (df: 70)
	Education (r_{rank})	0.31 [0.23–0.40]	<0.001 (n: 488)	0.08 [-0.08–0.24]	0.293 (n: 158)	-0.01 [-0.25–0.21]	0.951 (n: 72)
	Household size (r)	0.22 [0.13–0.30]	<0.001 (df: 486)	0.18 [0.02–0.33]	0.025 (df: 156)	0.03 [-0.20–0.26]	0.785 (df: 70)
	Degree LOG (r)	0.07 [-0.03–0.16]	0.153 (df: 468)	-0.02 [-0.18–0.14]	0.808 (df: 149)	-0.03 [-0.26–0.21]	0.838 (df: 67)
Clustering of vaccination and disease	Vaccinated (r_{φ})	0.23 [0.14–0.32]	<0.001 (df: 453)	0.02 [-0.14–0.18]	0.817 (df: 143)	0.07 [-0.17–0.30]	0.567 (df: 67)
	Belief vaccination protects (r_{rank})	0.26 [0.18–0.35]	<0.001 (n: 455)	0.02 [-0.14–0.18]	0.812 (n: 145)	0.11 [-0.13–0.32]	0.387 (n: 69)
	One or more symptoms (r_{φ})	0.11 [0.02–0.20]	0.018 (df: 453)	0.11 [-0.05–0.27]	0.179 (df: 143)	0.15 [-0.09–0.37]	0.231 (df: 67)
	Self-reported common cold (r_{φ})	0.04 [-0.06–0.13]	0.455 (df: 453)	-0.08 [-0.24–0.08]	0.333 (df: 143)	-0.11 [-0.33–0.14]	0.389 (df: 67)
	Self-reported influenza (r_{φ})	0.26 [0.17–0.34]	<0.001 (df: 453)	0.03 [-0.13–0.20]	0.691 (df: 143)	-0.04 [-0.27–0.20]	0.764 (df: 67)

[a]Coefficients and 95% confidence intervals are shown.

Pairs showed frequently a similar influenza vaccination status (r_{φ} = 0.23 [95% CI 0.14-0.32]) and the same beliefs on vaccine effectiveness (r_{rank} = 0.26 [95% CI 0.18-0.35]). To a lesser extent, we observed assortative recruitment by self-reported symptoms (r_{φ} = 0.11 [95% CI 0.02-0.20]). There were 150 (33.0%) pairs where both individuals reported at least one symptom compared to 104 (22.9%) where both did not report any symptoms.

The assortative correlations by age persisted between any two participants that were two or more link steps away from each other in the same network chain, indicating that the survey mainly spread among individuals of similar age. Having one or more symptoms also seemed to cluster within the same recruitment trees.

Comparison with POLYMOD

Figure 2A shows the recruiter-recruit matrix by age that visualizes the strong tendency of participants to recruit contact persons of similar age. This pattern is most pronounced in those aged 50-65 years. We observed two sub-diagonals that represent recruitment across generations. A column wise comparison with the contact mixing matrix by age of POLYMOD showed comparable distributions for participants aged between 20-39 years (Figure 2B). This suggests that recruitment links might be representative for the contact persons recruiters encounter in daily life, at least with respect to age. However, the number of recruitments by participants in this age group was likely insufficient for a proper comparison of samples. A statistical comparison of the entire two matrices showed a significant difference ($p < 0.001$).

Overall, the strong assortative recruitment by age resulted in higher sample proportions of recruits of similar ages, while pairs of individuals with different ages were underrepresented compared to POLYMOD. The average numbers of contact persons by age reported in the questionnaire by participants were consistent with the assortative recruitment patterns. This was most apparent for participants aged between 19-60 years who reported mainly contact with persons of the same age group (Figure 2C).

Participants below the age of 65 years mostly reported contacts at work or university, while those aged ≥ 65 years reported mostly contacts at other places. The number of persons contacted at different locations was similar in POLYMOD, although participants in our sample reported slightly less contact persons at home (Figure 2D).

In the Additional file we displayed the mixing matrices by age of our sample and of POLYMOD separately, as well as the absolute number of self-reported symptoms and a visualisation of the mixing patterns by degree.

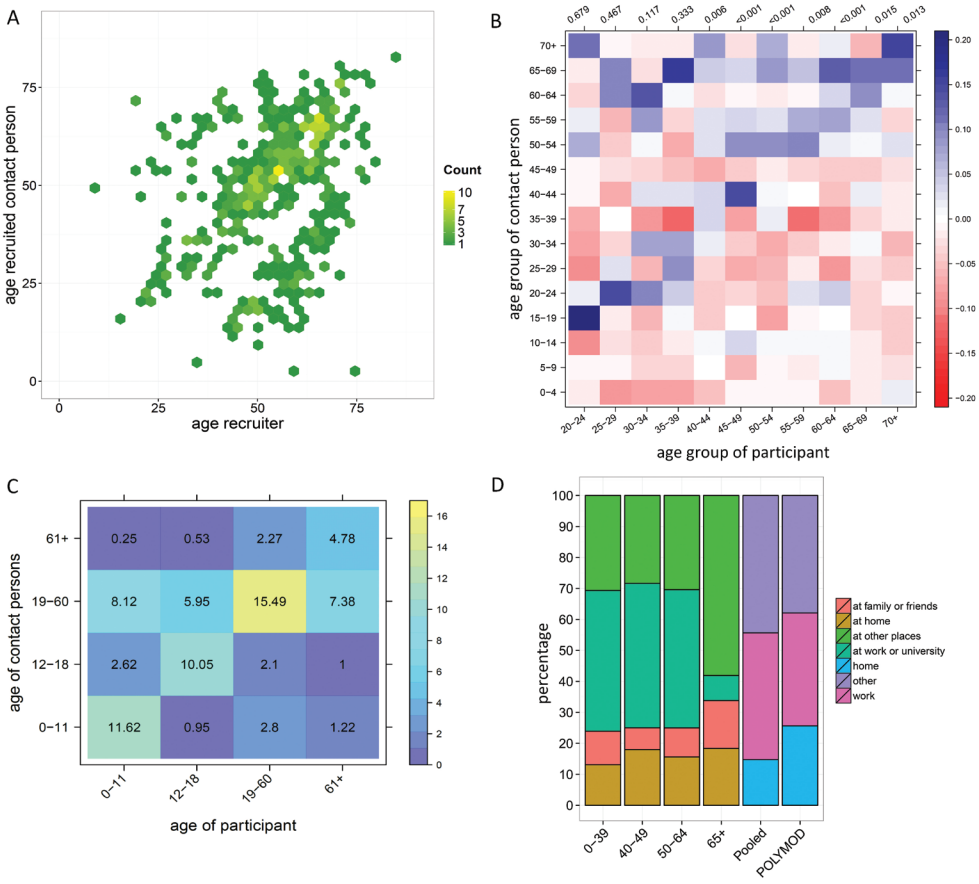


Figure 2 Recruitment and contact persons by age. **A** Recruitment patterns by age (n_{pairs}: 488). **B** Difference between recruitment matrix and contact matrix by age of Dutch POLYMOD. Colours and scale indicate for each cell the proportional difference between both matrices, for the particular participant's age group and his/her contact person's age group (note: recruitment matrix minus POLYMOD matrix). For each participant's age group, integer counts of contact persons were compared with POLYMOD using a two-sample KS test, the p values are shown above each column. **C** Contact persons reported in questionnaire by participants, values indicate the average number of contact persons in an age group recorded per day by participants. **D** Contact location by age groups and pooled for comparison with POLYMOD. The first four columns show the locations as displayed in the questionnaire. For comparison with POLYMOD, the sample was weighted for the size of POLYMOD age groups (weights are displayed in Additional file 1), and the category "at the home of family and friends" was combined with "other". POLYMOD was regrouped as "home", "work" (at work and at school combined) and "other" (leisure, travel and other combined), frequency of contact with the same person was ignored and for contact at multiple locations only the first entry was counted (equivalent to our questionnaire)

Spatial recruitment

The median geographical distance between a recruiter and recruit was 3.0 km (mean: 21.0; SD: 38.5) (Figure 3). There were 180 recruits with the same postal code as their recruiter, which suggests recruitment of nearby residents including household members. Seeds and their recruits lived on average further away from each other than pairs of participants in consecutive waves. The mean distance decreased from 22.4 km (SD: 40.1) between participants in waves 0 and 1, to 14.6 km (SD: 27.1) between participants in waves 2 and 3.

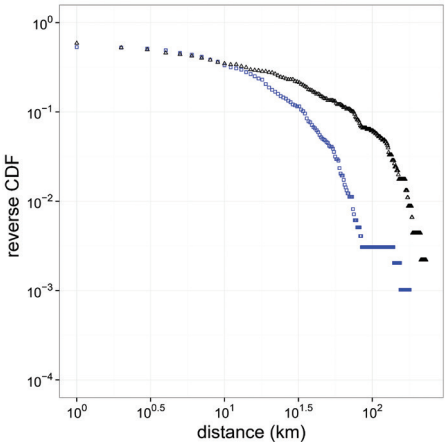


Figure 3 Distribution of recruitment and commuting distances. Black triangles indicate distances between recruiters and their recruits, with median 2.8 km (mean: 20.7; SD: 38.3). Blue squares indicate distances participants commute to work, with median: 3.4 km (mean: 11.0; SD: 18.1).

Of all recruitments, 76.4% took place within the same Dutch province (i.e., the Netherlands counts 12 provinces that represent the administrative layers between the national government and the local municipalities) or within Belgium (included as one 'province'), which corresponds to the 87.7% of all participants that work or study within their province of residence (Figure 4). The estimated probabilities of recruiting a contact person in the municipality where the recruiter both lived and worked varied between 0.56-0.77 (see also Additional file).

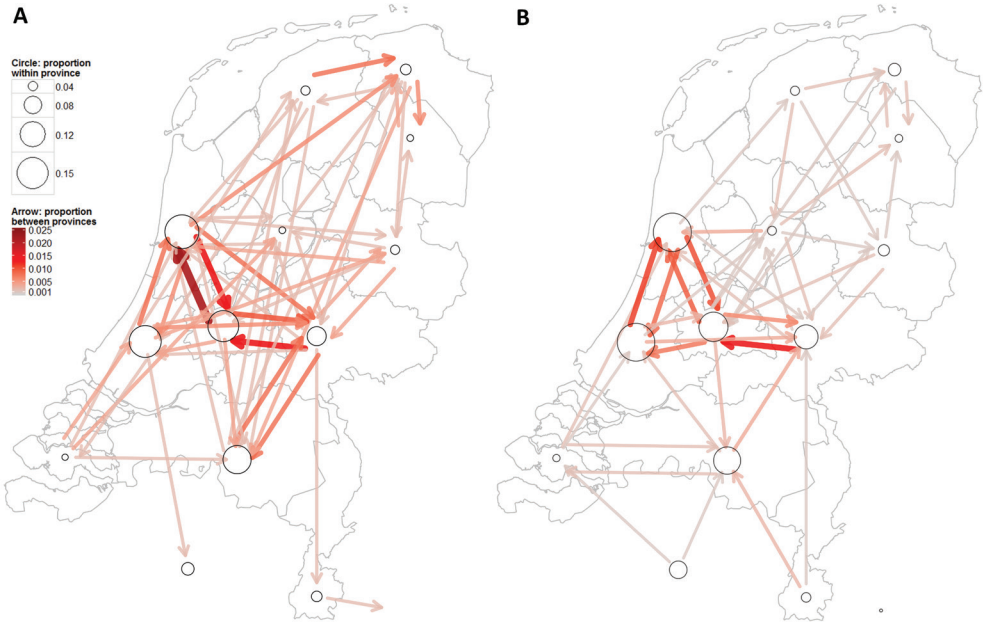


Figure 4 Spatial recruitment and commuting network structure. (A) Peer recruitment within the Netherlands and (between) Belgium. Arrows indicate recruitment between provinces and circles recruitment within a province. **(B)** Commuting network: directions that participants daily commute to work or study. Arrows indicate commuting across provinces, and circles commuting within a province. Sizes of arrows and circles are weighted for the total number of recruitments/commuters, with darker colours/larger circles indicating higher proportions. The maps were created with a shapefile (.shp file) that was extracted from GADM, an online geographic database of global administrative areas that is freely available for academic and other non-commercial use[46].

The distance between a recruiter and recruit determined the type of contact networks being sampled. Recruitment of persons with same postal code was stronger assortative by age, education, household size, degree, vaccination status and vaccination beliefs, and strongly disassortative by sex, compared to recruitment of persons who lived 1 km or further away. These patterns may reflect recruitment of individuals within the same household, such as partners. Participants were more likely to recruit persons of the same sex who lived 1 km or further away. Recruitment was strongly assortative by vaccination beliefs for pairs living >10 km away from each other, and by one or more symptoms and self-reported influenza for pairs living 1 to 10 km away from each other (Table 4).

DISCUSSION

In this study we explored social contact networks arising from a respondent-driven survey conducted in the Netherlands and parts of Belgium during the winter season 2013-2014. We have shown that an online respondent-driven method in combination with participatory surveillance can be used to (i) study contact networks relevant for the spread of infectious diseases that transmit via close contact between individuals, (ii) detect clustering of these diseases in a contact network, and (iii) reach within short time and with large spatial coverage a diverse group of individuals in the general population. Furthermore, we found that the spatial distribution of recruitment influences the type of contact networks being sampled.

We analysed a large number of recruiter-recruit pairs and of individuals with different ages and backgrounds. This enabled us to investigate the distribution of numbers of contact persons and to quantify the strength of network ties that allow the transmission of diseases that spread via close contact or airborne droplets. Such information can inform mathematical models of infectious disease epidemics[31-34]. Symptomatic participants showed a tendency to recruit other symptomatic participants, at least for one or more symptoms and self-reported influenza. This observation lends some support to our hypothesis that via respondent-driven recruitment we reached contact persons whom a participant may infect or by whom a participant may get infected in case of an infectious disease outbreak. The self-reported symptom data by pairs of participants provides an indication on disease clustering in contact networks. Such information can be quickly obtained with online respondent-driven detection as the recruitment generation interval was less than one day. We also observed clustering of the same influenza vaccination status and reported sentiments about vaccination in recruitment trees. Such clustering of similar health behaviour has been described before and provides an indication of clustering of vaccine-induced immunity in a population[20, 21]. Clustering of negative vaccination statuses or sentiments about vaccination leads to clusters of unprotected individuals that increase the likelihood of disease outbreaks[21]. Such information could be used to design intervention messages for vulnerable populations.

Table 4. Effect of geographical distance on recruiter-recruit[a] relationship.

Variable	correlation / odds ratio	Same postal code[b]	p value	1 to 10 km[b]	p value	>10 km[b]	p value	Overall test
Age	*r*	0.50 [0.39–0.61]	<0.001 (df: 177)	0.40 [0.25–0.53]	<0.001 (df: 144)	0.21 [0.06–0.35]	0.008 (df: 160)	0.008
Education	r_{rank}	0.33 [0.19–0.47]	<0.001 (n: 179)	0.26 [0.09–0.41]	0.001 (n: 146)	0.32 [0.15–0.47]	<0.001 (n: 162)	0.770
Household size	*r*	0.40 [0.26–0.51]	<0.001 (df: 177)	0.08 [-0.09–0.24]	0.363 (df: 144)	0.14 [-0.01–0.29]	0.067 (df: 160)	0.004
Degree LOG	*r*	0.16 [0.01–0.30]	0.034 (df: 173)	-0.02 [-0.18–0.15]	0.855 (df: 136)	0.04 [-0.11–0.20]	0.583 (df: 154)	0.264
Belief vaccination protects	r_{rank}	0.19 [0.04–0.35]	0.012 (n: 169)	0.17 [-0.00–0.33]	0.056 (n: 131)	0.41 [0.27–0.55]	<0.001 (n: 154)	0.041
Sex	OR	0.35 [0.14–0.79]	0.006 (n: 179)	4.86 [2.13–11.39]	<0.001 (n: 146)	1.91 [0.93–3.93]	0.054 (n: 162)	<0.001
Vaccinated	OR	4.94 [2.30–11.07]	<0.001 (n: 169)	3.54 [1.50–8.67]	0.001 (n: 131)	1.36 [0.66–2.81]	0.366 (n: 154)	0.025
One or more symptoms	OR	1.09 [0.57–2.11]	0.771 (n: 169)	3.03 [1.39–6.80]	0.002 (n: 131)	1.36 [0.68–2.72]	0.349 (n: 154)	0.093
Self-reported common cold	OR	1.27 [0.47–3.23]	0.585 (n: 169)	1.31 [0.33–4.35]	0.635 (n: 131)	1.10 [0.25–3.79]	0.874 (n: 154)	0.974
Self-reported influenza	OR	8.01 [1.98–31.38]	<0.001 (n: 169)	9.32 [1.22–59.64]	0.001 (n: 131)	4.90 [0.73–25.05]	0.052[c] (n: 154)	0.814

[a]Number of pairs with same postal code (n: 180 pairs), with same Internet Protocol (IP) address (n: 86), and number of pairs with both same postal code and same IP address (n: 72).
[b]Correlation coefficients / odds ratios with 95% confidence intervals are shown.
[c]Fisher's exact test was used for contingency tables containing small values (n < 10).

Compared to a paper-based approach[17], online peer recruitment was spatially wider dispersed and covered a larger geographical area. A stratification on distance of the relationships between recruiter-recruit pairs showed differences in the type of recruited contact persons. There may be several explanations why a participant invited certain contact persons[35]. For example, symptomatic participants may have been biased towards inviting symptomatic contact persons who lived further away than contact persons whom they more frequently meet. A proper assessment would require to investigate the 'pool of contact persons' from which a recruiter can choose, and which contact persons were invited but did not join the survey. Furthermore, identifying different types of relations (e.g., family members, friends or colleagues) by asking recruits about their recruiter would allow further clarification of the observed correlations. Such information can only be collected with a non-anonymous survey design, which would also make it possible to measure transitivity, i.e., the extent to which contact persons of a participant are also contact persons of each other[36]. This network property is known to reduce the rate at which an infection can spread through a network[36-38].

The 'who recruited whom' matrix stratified by age showed qualitatively similar structures as the contact matrix by age reported in POLYMOD[9]. In addition, proportions of contact persons at different locations were similar to POLYMOD and the regression analysis showed similar covariates such as age, household size and days of week to affect degree. This suggests for online recruitment that invited contact persons are in general representative for the contact persons daily encountered by participants and that respondent-driven detection can indeed provide accurate information on the underlying contact network. However, despite the fact that recruitment criteria were set the same for all participants, regardless of whether they reported symptoms, we cannot preclude a bias in how participants chose from their contact persons. The age matrices were statistically not comparable. There may be several explanations for this statistical discrepancy, such as a difference in the age distributions of the samples and the fact that POLYMOD participants were able to report an unrestricted number of contact persons, while our survey participants could only invite a maximum of four contact persons.

This study has limitations. By using participatory surveillance panels for recruitment of seeds, we reached a diverse group of individuals within a short period of time. However, the volunteers in these panels are not representative for the general population; some groups like women and highly educated persons are overrepresented[19]. Such overrepresentations are common in participatory surveillance systems[18]. We did reach all age groups, but due to strong assortative peer recruitment certain age classes were represented more in the sample and the young age classes were reached less with our survey, therefore limiting the generalisability of our results to the young age groups.

To reduce the participation burden and stimulate recruitment at the end of the questionnaire, we applied an aggregated contact diary design, i.e., a participant did not need to report on each contact separately. The mean number of contact persons per participant was therefore likely higher than in previous studies[9, 39]. More importantly, we did not collect information on contact intensity and duration. The probability of transmission between individuals requires different levels of contact for different infectious diseases, e.g., influenza and measles require only spatial proximity between individuals to transmit, while Ebola is believed to require physical contact to cause infection[7, 14].

Note that the survey did not include questions on other potentially important transmission routes, such as exposure not involving physical contact or conversation (e.g., sneezing passenger in public transport) or indirect fomite transmission from shared contaminated objects[7]. Earlier studies explicitly linked contact intensity and duration with infection risk and showed their importance for understanding transmission dynamics[40, 41]. Contact duration also influences the likelihood that a certain contact is reported, e.g., contacts of long duration are substantially more likely to get reported than contacts of short duration[42, 43]. It is possible to derive these contact metrics from earlier studies, but not to exclude the effect of heterogeneities in motivation or recall capabilities on reported numbers of contacts, e.g., between male and female participants[42].

In a future survey volunteers of participatory surveillance panels could be selected according to specific characteristics to obtain seeds that are in some sense representative for the general population. Furthermore, it may be useful to conduct a similar study in other countries where comparable participatory surveillance systems are in place, such as the United Kingdom, Italy and France, to allow for a country comparison[44].

CONCLUSIONS

In this study we used online respondent-driven detection to study the distribution of the number of contact persons and mixing patterns within contact networks. The observed contact patterns are relevant for the transmission of respiratory pathogens that spread via close contact between individuals. We found that the spatial distribution of recruitment influenced the type of contact networks being sampled. Even though complex mechanism influence peer recruitment, the observed statistical relationships reflected the observed contact network patterns in the general population. This provides useful and innovative input for predictive epidemic models relying on network information.

Abbreviations

CI: confidence interval
ILI: influenza-like-illness
IP: Internet Protocol
km: kilometres
KS: Kolmogorov-Smirnov
OR: odds ratio
SD: standard deviation

Competing interests

Carl E. Koppeschaar started 'deGroteGriepmeting.nl' in the Netherlands and Belgium in 2003 and received an educational grant by Pfizer to develop 'deGroteLongontstekingmeting.nl'. He was co-beneficiary of the European FP7 program EPIWORK to develop Influenzanet.eu.

Authors' contributions

MLS and MEEK conceived and designed the study; MLS, JES and MEEK developed the questionnaires; MLS, LB and AT developed the online survey system; MLS, CEK and MEEK performed the study; MLS, PGMH, VB and MEEK analysed the data; MLS, PGMH, VB, JES and MEEK helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study was conducted within the Utrecht Center for Infection Dynamics. The Swedish Research Council (vr.se) has financed the development of the online survey system. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to Martin Camitz, Antwan Wiersma and Ronald Smallenburg for their help with the survey launches and to Jan van de Kassteele and Albert Wong for their help with the statistical analyses and the comparison with the Dutch POLYMOD data.

REFERENCES

1. Musher DM. How contagious are common respiratory tract infections? New England Journal of Medicine. 2003;348:1256-66.

2. Rea E, Lafleche J, Stalker S, Guarda BK, Shapiro H, Johnson I et al. Duration and distance of exposure are important predictors of transmission among community contacts of Ontario SARS cases. Epidemiology and infection. 2007;135(6):914-21.

3. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 2005;437(7056):209-14.

4. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(7):2825-30.

5. Wallinga J, Edmunds WJ, Kretzschmar M. Perspective: human contact patterns and the spread of airborne infectious diseases. Trends in microbiology. 1999;7(9):372-7.

6. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, Vespignani A. Dynamics of person-to-person interactions from distributed RFID sensor networks. PloS one. 2010;5(7):e11596.

7. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiology and infection. 2012;140(12):2117-30.

8. Barrat A, Cattuto C, Tozzi AE, Vanhems P, Voirin N. Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases. Clin Microbiol Infect. 2014;20(1):10-6.

9. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine. 2008;5(3):e74.

10. Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M. Transmission of influenza A in human beings. The Lancet Infectious diseases. 2007;7(4):257-65.

11. Eames K, Bansal S, Frost S, Riley S. Six challenges in measuring contact networks for use in modelling. Epidemics. 2014. doi:10.1016/j.epidem.2014.08.006.

12. Christakis NA, Fowler JH. Social network sensors for early detection of contagious outbreaks. PloS one. 2010;5(9):e12948.

13. Christakis NA, Fowler JH. The spread of obesity in

a large social network over 32 years. The New England journal of medicine. 2007;357(4):370-9.

14. Read JM, Eames KT, Edmunds WJ. Dynamic social networks and the implications for the spread of infectious disease. Journal of the Royal Society, Interface / the Royal Society. 2008;5(26):1001-7.

15. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PGM, Chanyasanha C, Tipayamongkholgul M et al. Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling. PloS one. 2014;9(11):e113711.

16. Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PGM et al. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PloS one. 2014;9(1):e85256.

17. Jenness SM, Neaigus A, Wendel T, Gelpi-Acosta C, Hagan H. Spatial Recruitment Bias in Respondent-Driven Sampling: Implications for HIV Prevalence Estimation in Urban Heterosexuals. AIDS and Behavior. 2014;18(12):2366-73.

18. Wojcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerging themes in epidemiology. 2014;11:7.

19. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Koppeschaar CE, Bengtsson L et al. Enhancing Syndromic Surveillance With Online Respondent-Driven Detection. American journal of public health. 2015;105(8):e90-7.

20. Barclay VC, Smieszek T, He J, Cao G, Rainey JJ, Gao H et al. Positive network assortativity of influenza vaccination at a high school: implications for outbreak risk and herd immunity. PloS one. 2014;9(2):e87042.

21. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS computational biology. 2011;7(10):e1002199.

22.Hilbe JM. Alternative variance parameterizations: Poisson inverse Gaussian regression. Negative Binomial Regression. 2nd ed. New York: Cambridge University Press; 2011. p. 341-3.

23. Dean C, Lawless JF, Willmot GE. A mixed poisson-inverse-Gaussian regression model. The Canadian Journal of Statistics. 1989;17(2):171-81.

24. Newman ME. Assortative mixing in networks. Physical review letters. 2002;89(20):208701.

25. Homogeneous KM, Association U. Contigency Table Analysis: Methods and Implementation Using R. New York: Springer; 2014. p. 187–91.

26. Lang JB. Maximum Likelihood Fitting of Multinomial-Poisson Homogeneous (MPH) Models for Contingency Tables using MPH.FIT. 2009. http://

homepage.stat.uiowa.edu/~jblang/mph.fitting/mph.fit.documentation.htm. Accessed 12 March 2015.

27. Lang JB. Multinomial-Poisson Homogeneous Models for Contingency Tables. Ann Stat. 2004;32:340–83.

28. Arsham H. Test for Equality of Several Correlation Coefficients. 2015. http:// home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/MultiCorr.htm. Accessed

12 March 2015.

29. Van Kerckhove K, Hens N, Edmunds WJ, Eames KT. The impact of illness on social networks: implications for transmission and control of influenza. Am J Epidemiol. 2013;178(11):1655–62.

30. Beutels P, Shkedy Z, Aerts M, Van Damme P. 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. 2006;134(6):1158–66.

31. Ferguson NM, Keeling MJ, Edmunds WJ, Gani R, Grenfell BT, Anderson RM, et al. Planning for smallpox outbreaks. Nature. 2003;425(6959):681–5.

32. Eubank S, Guclu H, Kumar VS, Marathe MV, Srinivasan A, Toroczkai Z, et al. Modelling disease outbreaks in realistic urban social networks. Nature.

2004;429(6988):180–4.

33. Longini Jr IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DA, et al. Containing pandemic influenza at the source. Science. 2005;309(5737):1083–7.

34. Germann TC, Kadau K, Longini Jr IM, Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci U S A. 2006;103(15):5935–40.

35. Wejnert C, Heckathorn DD. Web-Based Network Sampling Efficiency and Efficacy of Respondent-Driven Sampling for Online Research. Sociol

Methods Res. 2008;37(1):105–34.

36. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–2.

37. Keeling MJ. The effects of local spatial structure on epidemiological invasions. Proc Biol Sci. 1999;266(1421):859–67.

38. Volz EM, Miller JC, Galvani A, Ancel Meyers L. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol. 2011;7(6):e1002042.

39. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. Social encounter networks: characterizing Great Britain. Proc Biol Sci. 2013;280(1765):20131037.

40. De Cao E, Zagheni E, Manfredi P, Melegaro A. The relative importance of frequency of contacts and duration of exposure for the spread of directly transmitted infections. Biostatistics. 2014;15(3):470–83.

41. Smieszek T. A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. Theor Biol Med

Model. 2009;6:25.

42. Smieszek T, Barclay VC, Seeni I, Rainey JJ, Gao H, Uzicanin A, et al. How should social mixing be measured: comparing web-based survey and sensor-based methods. BMC Infect Dis. 2014;14:136.

43. Smieszek T, Burri EU, Scherzinger R, Scholz RW. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiol Infect. 2012;140(4):744–52.

44. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet

participatory surveillance experience. Clin Microbiol Infec. 2014;20(1):17–21.

45. The GADM project. GADM version 1.0: a geographic database of global administrative areas. 2009. http://www.gadm.org. Accessed 22 June 2014.

SUPPLEMENTARY MATERIALS

This file contains supporting information for the results presented in the manuscript "*Tracking social contact networks with online respondent-driven detection: who recruits whom*?". The supportive information is presented in the order as it is discussed in the main manuscript.

In chapter 1 we explained in detail how numbers of contacts and the effects of covariates were analysed. In chapter 2 we investigated the influence of participants' characteristics on the size of a recruitment tree. In chapter 3 we displayed the mixing matrices by age of our sample and of the Dutch POLYMOD separately. Here we also provided the absolute number of self-reported symptoms and a visualisation of the mixing patterns by degree. In chapter 4 we analysed the distance between recruiters and their recruits, and quantified the extent to which a recruiter who lives in a certain region in the Netherlands invited contact persons that live in the same municipality as the recruiter is working and/or studying.

Chapter 1 Numbers of contact persons and the effect of covariates

In the questionnaire participants were asked for number of contact persons during one full day ('yesterday'), this number was defined to be a participant's 'degree'. First, we looked at the distribution of degree, stratified by days of week and the locations that were predefined in the questionnaire (Figure A1). For at home and at other places the distributions of degree were fairly similar. During weekdays participants reported more contact persons at work or university, then during the weekend. There were no large differences in the total degree distributions (see 'at all locations') between weekdays and weekends.

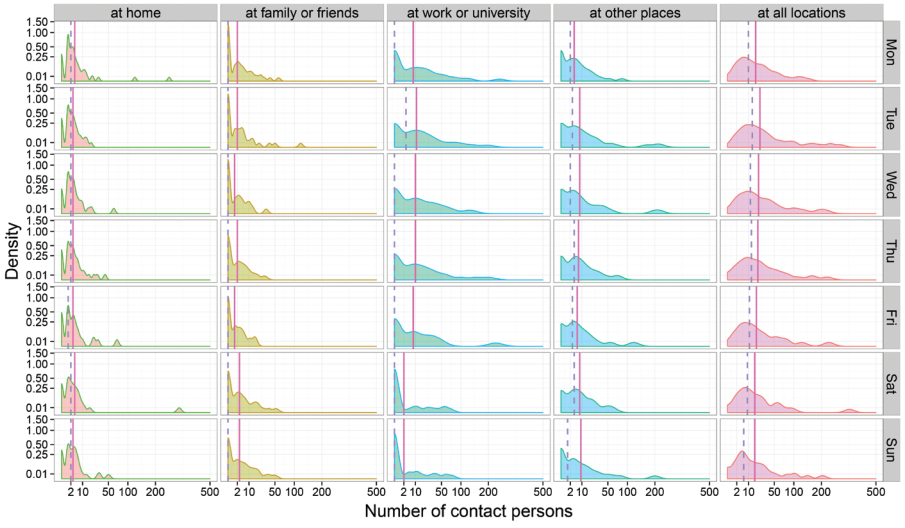


Figure A1. Contact persons stratified by days of the week and location of contact. In each distribution the mean (solid line) and median (dotted line) is displayed. Participants with more than 500 contact persons are not displayed.

We investigated which covariates influence degree using a regression model. Firstly, we investigated which theoretical distribution best fitted the empirical distribution using the R package 'GAMLSS'. The degree distribution showed strong over-dispersion, with a mean degree of 19.6 per participant (median: 11.0; SD: 35.3). Table A1 displays the parameter estimates and AIC's of the various fitted distributions. Note that the power-law was fitted with the GAMLSS function "PARETO2". Cumulative distributions with a power-law form are sometimes said to follow a Pareto distribution (or Zipf 's law)[1]. Figure A2 displays the various distributions in a reverse cumulative probability distribution plot (Log_{10} transformed).

Table A1. Theoretical distributions fitted to empirical degree distribution.

	parameter(s)	AIC
Pareto 2 (α; x_m)[a]	49.46; 0.28	12179
Log normal (μ; σ^2)[a]	11.67; 1.03	12184
Poisson Inverse Gaussian (μ; λ)	19.62; 1.96	12207
Negative binomial (μ; k)	19.61; 1.12	12475
Geometric (discrete) exponential (p)	19.61	12485
Poisson (λ)	19.61	48667

[a] Continuous distributions fitted to a discrete distribution.

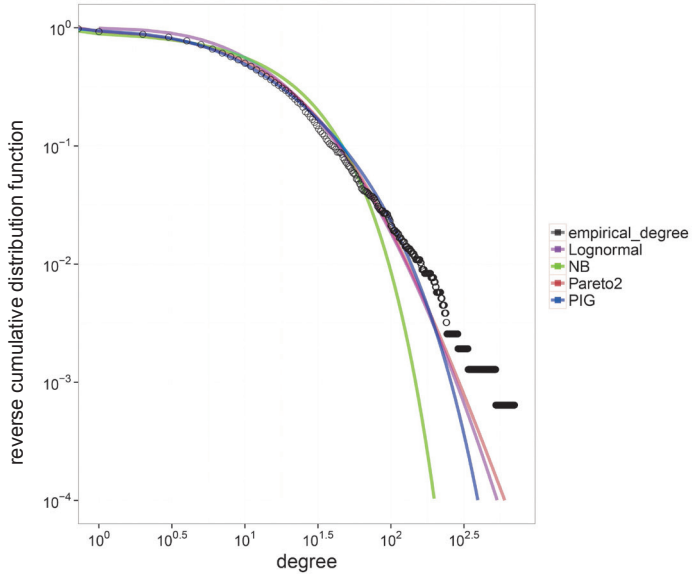


Figure A2. Fitted theoretical distributions to empirical degree distribution. The empirical degree distribution (number of reported contact persons) is displayed with black circles. The figure displays four theoretical distributions that were fitted to the empirical degree distribution, namely: the Poisson-inverse Gaussian distribution (PIG, discrete distribution), the Negative Binomial distribution (NB, discrete distribution), the Pareto 2 distribution (continuous distribution) and the Log-normal distribution (continuous distribution).

Based on the AIC's, the continuous distributions Log-normal and Power-law best fitted the empirical degree distribution[2]. However, these are continuous distributions fitted to a discrete distribution. Therefore, we chose the first best fitted discrete distribution: the Poisson-inverse Gaussian (PIG). The PIG distribution, an alternative to negative binomial, has the potential for modelling highly dispersed count data due to the flexibility of Inverse Gaussian distribution[3, 4]. We applied the PIG distribution in the regression analysis.

We used a PIG regression model to investigate the effect of the following covariates on degree: age, sex, household size and ILI, and days of the week. The reference categories were the 0–39 age group, females, one-person households, no self-reported ILI, and Sunday. Table A2 shows the output of the regression model. IRR stands for incidence rate ratio that are standard provided when conducting a PIG regression analysis.

Table A1. Theoretical distributions fitted to empirical degree distribution.

	IRR[a]	SE	t value	Pr(>|z|)	2.5%	97.5%
Intercept	13.055	0.109	23.527	0.000	10.540	16.171
40-49	0.969	0.094	-0.338	0.735	0.805	1.166
50-64	0.928	0.081	-0.918	0.359	0.791	1.089
65+	0.692	0.093	-3.985	0.000	0.577	0.829
male	1.053	0.058	0.887	0.375	0.940	1.180
household size:2	1.019	0.070	0.263	0.792	0.887	1.170
household size:3	1.441	0.094	3.896	0.000	1.199	1.732
household size:4	1.552	0.094	4.655	0.000	1.290	1.867
household size:5 or more	1.809	0.121	4.888	0.000	1.426	2.294
ILI	0.367	0.188	-5.320	0.000	0.254	0.531
Monday	1.334	0.089	3.237	0.001	1.120	1.588
Tuesday	1.837	0.098	6.192	0.000	1.515	2.226
Wednesday	1.597	0.105	4.469	0.000	1.301	1.961
Thursday	1.615	0.107	4.470	0.000	1.309	1.993
Friday	1.423	0.122	2.897	0.004	1.121	1.806
Saturday	1.269	0.109	2.197	0.028	1.026	1.570

[a] IRR: incidence rate ratio. Number of observations: 1559, df: 17 AIC: 12050.81, Global deviance: 12016.81

Chapter 2 Numbers of contact persons and the effect of covariates.

We conducted a descriptive analysis to investigate which characteristics of individuals in a recruitment tree influence the total size of a tree. Firstly, we plotted the number of nodes, i.e., participants who completed the questionnaire, stratified by characteristics of seeds (Figure A3).

The characteristics of seeds did not appear to influence the number of nodes in a recruitment tree. Figure A3-B did show a slight increase in tree size for a larger proportion of trees with a female seed, e.g., of all trees with a node size of 4 more than 75% had a female seed. This effect of female seeds was not shown for trees with a size of 5 or more nodes, which is probably due to the lower number of trees with those sizes.

In Figure A4 we investigated the relationship between tree size and the composition of the entire recruitment tree. Overall, the larger the proportion of women or individuals with a bachelor's degree or higher in a recruitment tree, the larger the tree size was on average (see Figures A4-A to C). The average age in a recruitment tree did not appear to influence the number of nodes in a recruitment tree (Figure A4-D).

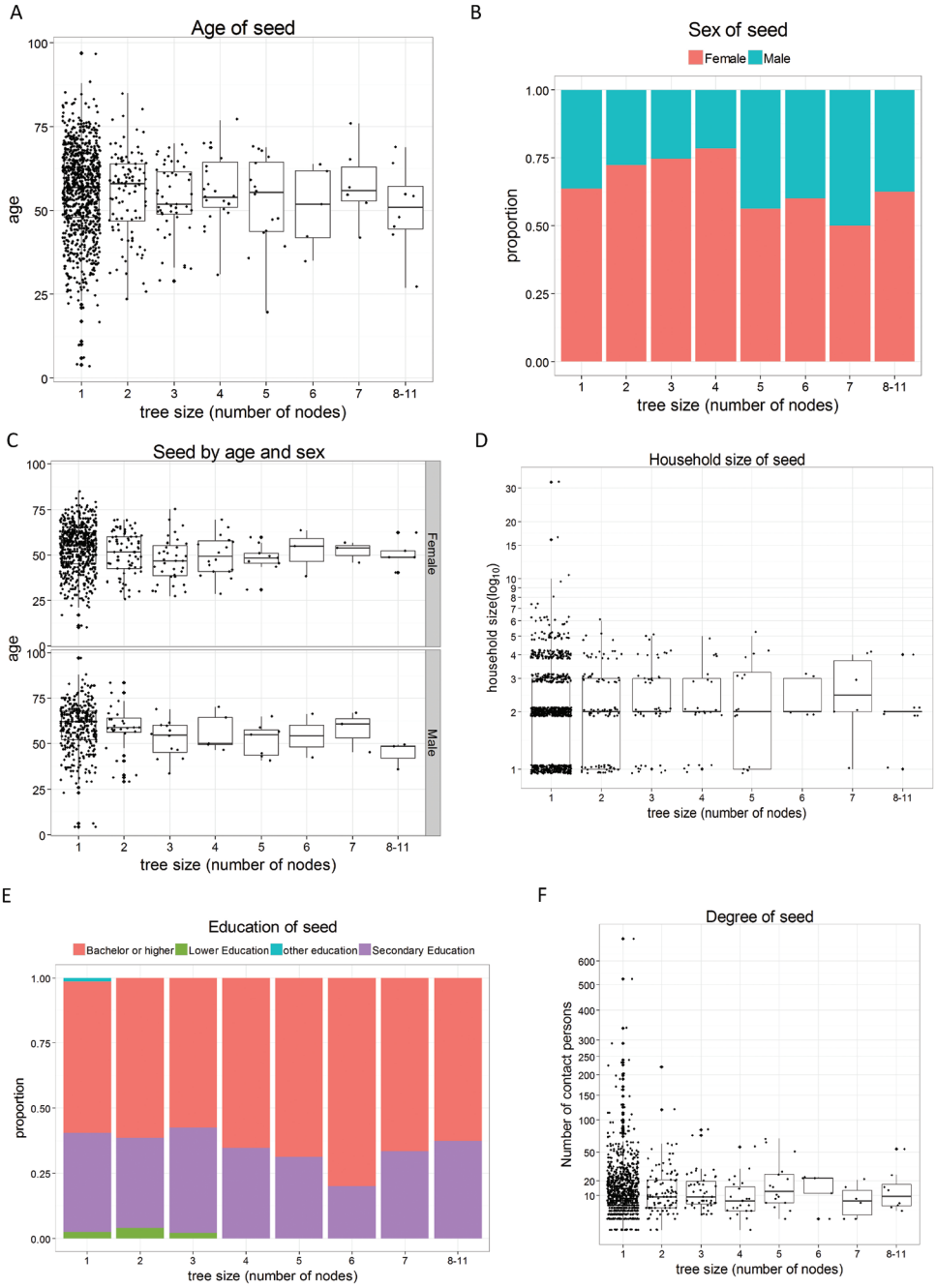


Figure A3. Relation between tree size and characteristics of *seeds*. With **(A)** age of seeds, **(B)** sex of seeds, **(C)** age and sex, **(D)** household size, **(E)** educational level of seeds, **(F)** number of contacts of seeds (degree).

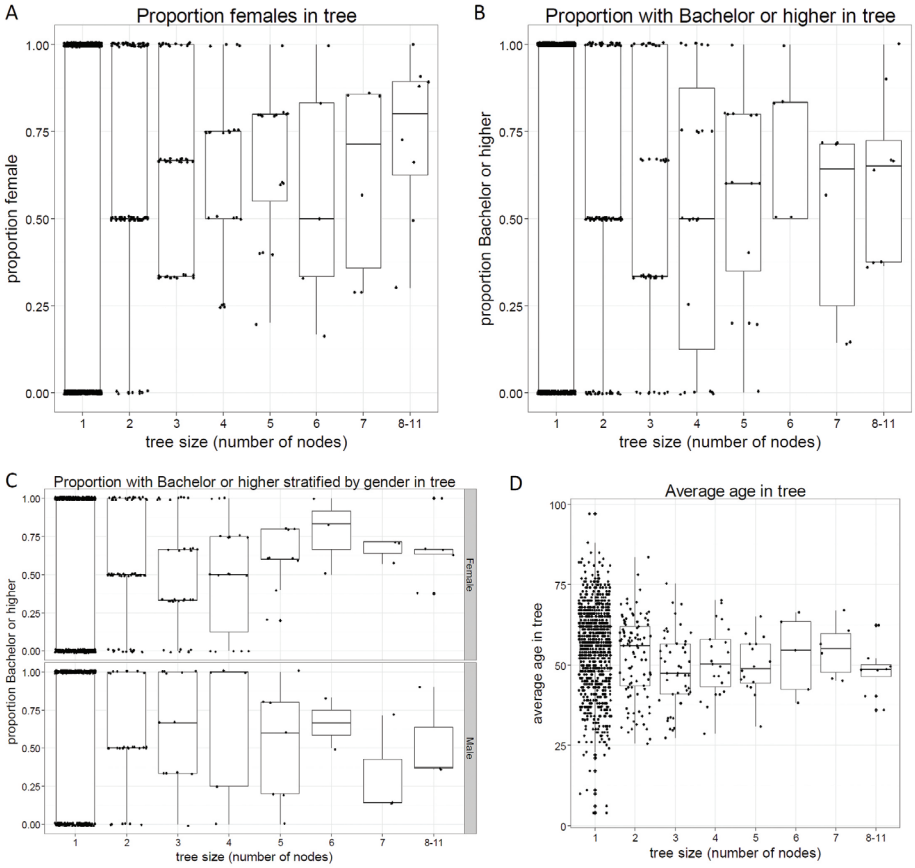


Figure A4. Relation between tree size and composition of trees. With **(A)** proportion of females in a tree, **(B)** proportion of individuals with a bachelor's degree or higher, **(C)** tree size stratified by number of individuals in a tree with a bachelor's degree or higher and sex, **(D)** the average age of individuals in one tree.

Chapter 3 Recruitment mixing patterns

3.1 Mixing by age and comparison with POLYMOD

We compared the recruiter-recruit matrix stratified by age with the participant-contact matrix by age collected during the Dutch POLYMOD study (see Figure A5)[5, 6]. We used the Dutch POLYMOD data that was corrected for digit preference by participants for the age of contact persons, details on this correction can be found in Van de Kassteele, J., et al.[6].

Strong assortative mixing patterns by age were observed in both matrices. However, in our sampled recruiter-recruit matrix the younger age groups (below 20 years of age) were not represented. In the Dutch POLYMOD study these younger age groups were purposely oversampled to be able to analyse their contact patterns, as the hypothesis is that children play a central role in the transmission dynamics of influenza pandemics[6]. Children have frequent contact within their own groups and they have a wide range of contacts, therewith connecting all age groups[7].

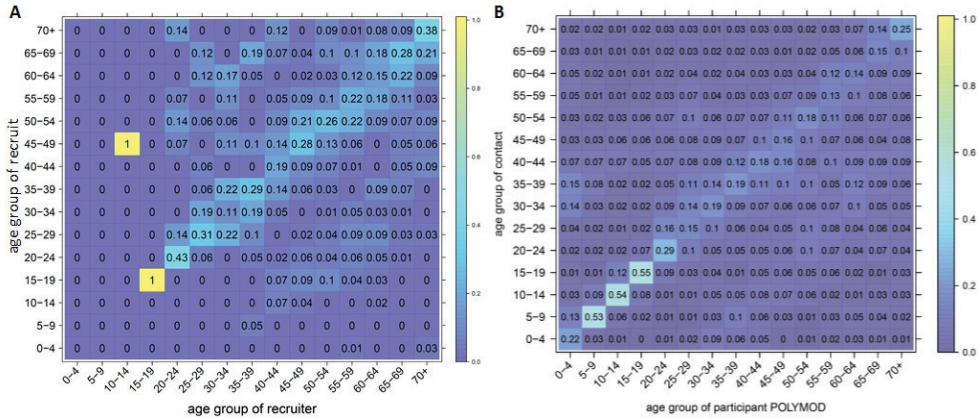


Figure A5. Comparison with the Dutch POLYMOD study. We compared the recruiter-recruit matrix stratified by age with the participant-contact matrix by age collected during the Dutch POLYMOD study[6]. (**A**) Sampled Recruiter–Recruit matrix, with respect to age; the values and colours indicate for each age group of recruiters the proportion of contact persons recruited. Thus, e.g., 0.43 for recruits between "20-24" indicates that from all recruits recruited by a recruiter from age group "20-24", 43% was between 20-24 years. (**B**) POLYMOD participant–contact matrix, the values and colours indicate for each age group of POLYMOD participants the proportion of contact with persons of different age groups. POLYMOD had a correlation for age of 0.47 [0.45-0.48], p-value < 2.2e-16; df: 10186.

In Figure 2D (see main manuscript) we compared the number of contact persons reported at different locations by our participants, with the contacts reported at different locations in the Dutch POLYMOD study[6]. For this comparison the sample was weighted for the size of age groups in the POLYMOD study. The applied weights can be found in Table A3.

Table A3. Weighting sample for a comparison with POLYMOD

Age group	POLYMOD	SAMPLE[a]	Weight
0-39	439 (56.1%)	268 (17.2%)	3.27
40-49	76 (9.7%)	256 (16.4%)	0.59
50-64	132 (16.9%)	656 (42.1%)	0.40
65+	135 (17.3%)	379 (24.3%)	0.71

[a] The weights for age groups in our sample were obtained by dividing the POLYMOD proportion by the corresponding sample proportion.

3.2 Mixing by age and comparison with POLYMOD

We plotted the recruiter-recruit matrix by degree (see Figure A6). We observed random recruitment by degree between a recruiter and a recruit, which corresponds to the correlation coefficients for degree displayed in Table 3 in the main manuscript. As the dots clustered in the left corner, we looked more closely at the distribution up till an individual degree of 50. Furthermore, we plotted the distribution on a $\log_{10}$ scale, which also illustrated random mixing.

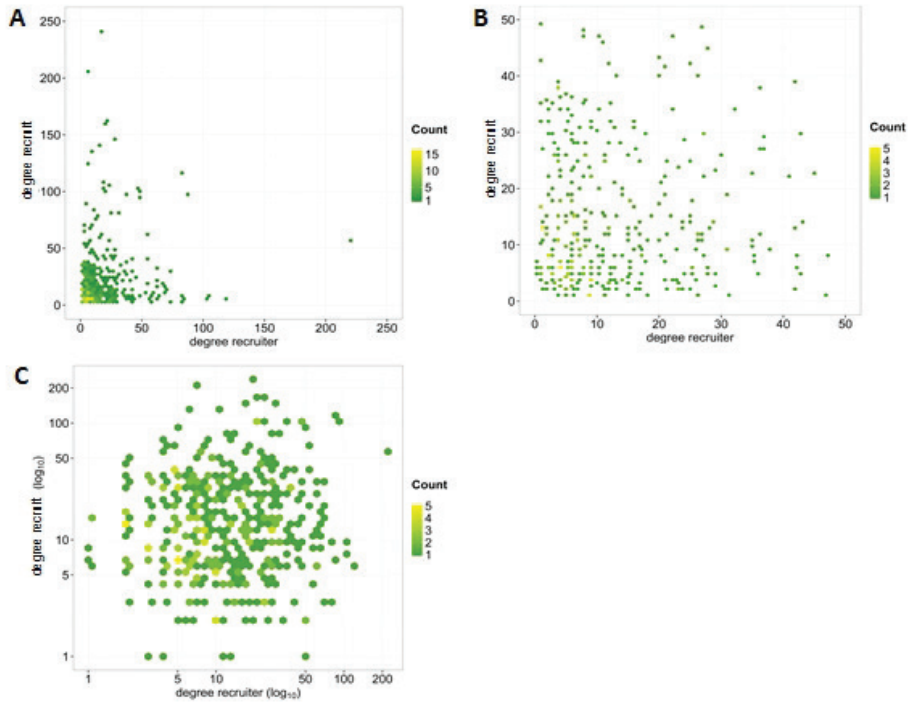


Figure A6. Recruiter degree versus recruit degree. (A) Untransformed **(B)** Untransformed but axes limited to a degree of 50 **(C)** Log_{10} transformed.

3.3 Mixing by self-reported symptoms

Table A4 displays the absolute number of times a symptom was reported by participants. Runny or blocked nose was most frequently reported and vomiting the least frequent. The second column displays the number of recruiter-recruit pairs where only one of them reported the concerning symptom. Third column displays the number of pairs where both the recruiter and recruit reported the particular symptom. Runny and blocked nose was again most frequently reported, by either one of them or by both individuals in a pair.

Table A4. Self-reported symptoms and recruiter-recruit pairs with symptoms.

	$N_{symptom\ reported}$	$N_{pairs\ only\ 1\ had\ symptom}$	$N_{pairs\ both\ had\ symptom}$
Fever	104	44	4
Chills	134	73	5
Runny or blocked nose	422	162	41
Earache	74	36	1
Sore throat	267	136	15
Cough	338	125	27
Stuffiness	146	68	4
Headache	360	153	36
Muscle / joint pain	292	149	30
Diarrhea	88	51	7
Vomit	26	14	0
Other symptoms	91	59	3

Chapter 4 Distance between recruiter-recruit pairs

We investigated the distance between recruiters and recruits based on the provided 4-digit postal codes. The distribution of the distance between recruiter-recruit pairs was right-skewed (Figure A7). Based on the distribution in Figure A7, we categorised for Table 4 (see main manuscript) distance into three groups: same postal code, 1-10 km and >10km. It appeared that for the first group (same postal code), recruiters invited slightly more similar aged recruits compared to the other two distance groups (see Figure A8). This is confirmed by the correlation coefficients for age in Table 4 in the main manuscript.

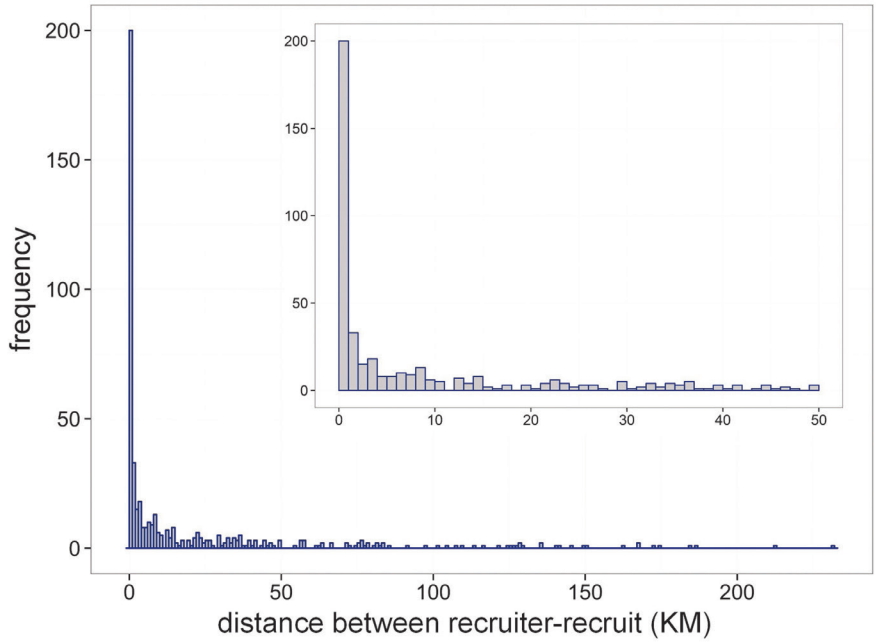


Figure A7. Histograms of distances between recruiter-recruit pairs in kilometres (km).

The mean distance between seeds (wave 0) and their recruits in wave 1 was higher than the mean distance between recruiter-recruit pairs in consequent waves. Figure A9 displays the distances for different link steps, as seen from the seed. Thus link step 1 is between seeds and their recruits in wave 1.

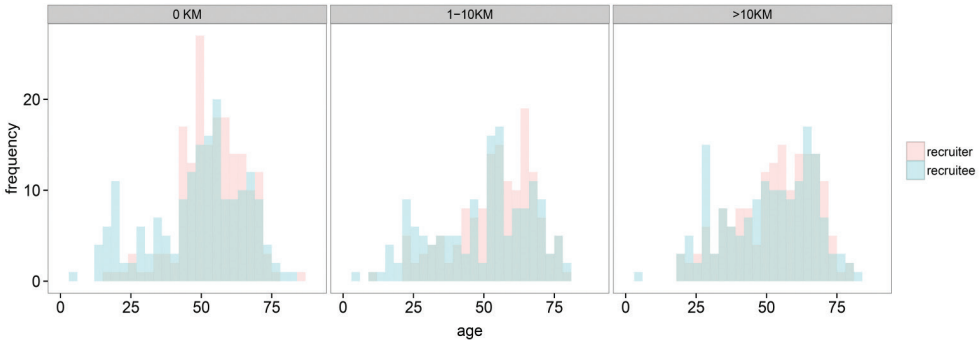


Figure A8. Age distributions of recruiters and recruits stratified by distance. Pink colour: recruiters; blue colour: recruits.

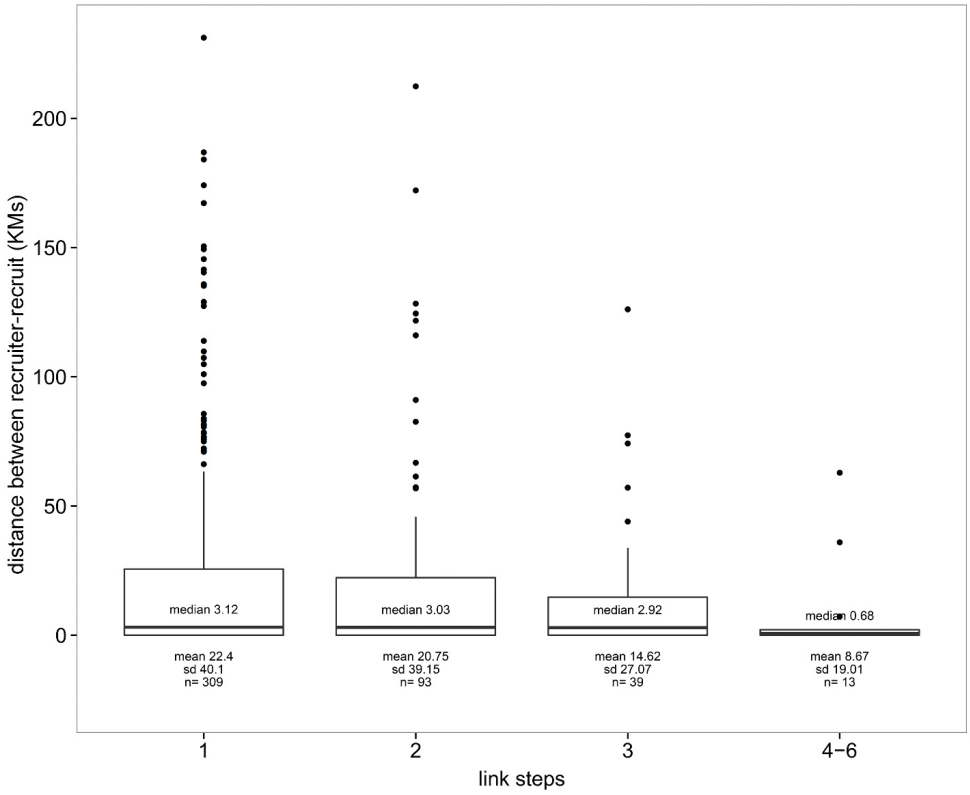


Figure A9. Distance between recruiter-recruit for different link steps. Link steps here indicate links steps as seen from the seed. Thus, link step '1' indicates the link between seeds and recruits-in-wave-1; link step 2 the link between recruiters-in-wave-1 and recruits-in-wave-2, and so on.

4.1 A logistic regression analysis

The Netherlands counts 12 provinces that represent the administrative layers between the national government and the local municipalities (i.e., subdivisions of a province). We categorised the Netherlands into four regions: North-Netherlands (Friesland, Groningen, Drenthe), Middle-Netherlands (Overijssel, Flevoland, Gelderland, Utrecht), West-Netherlands (North-Holland, South-Holland) and South-Netherlands (North-Brabant, Limburg, Zeeland).

Figure 4 demonstrated similar patterns between the recruitment trees (Figure 4A) and the commuting network (Figure 4B). Therefore, we investigated for Dutch participants the relationship between the geographical locations where a recruiter works and/or studies, and the location where their recruited contact person lives. We excluded participants living in Belgium. Home location was defined by the provided 4-digit postal code. The work location was defined by the city or village that was provided in the questionnaire.

Our goal was to quantify the extent to which a recruiter who lives in a certain region (four

regions defined, see above) invites contact persons that live in the same municipality as the recruiter is working/studying.

We used a mixed effect logistic regression model to estimate the binary outcome:
- recruiter did not invite a recruit who lives in the same municipality as the recruiter is working or studying (0)
- recruiter invited a recruit who lives in the same municipality as the recruiter is working or studying (1)

This outcome variable was created through recoding:
- "municipality where recruiter works/studies" ≠ "municipality where his/her recruit lives" [1]
- "municipality where recruiter works/studies" = "municipality where his/her recruit lives" [2]

The log odds of the binary outcome was modelled as a linear combination of the variables "region of residence recruiter" (four regions) and "recruiter lives and works in same municipality" (binary: yes/no), with the region West-Netherlands and 'recruiter not working in the same municipality as he/she is living' as a reference group. The random intercept was provided by recruiter ID, to correct for differences between recruiters, e.g., in numbers of contact persons invited per recruiter and type of recruited contact persons.

Table A5. Frequency table

Province of residence	works and lives in same municipality	Recruiter *did not* invite recruit in same municipality as he/she is working/studying[1]	Recruiter invited recruit in same municipality as he/she is working/studying[2]	Total
South-Netherlands	yes	15 (40.5%)	22 (59.5%)	37
	no	19 (79.2%)	5 (20.8%)	24
Middle-Netherlands	yes	33 (44.6%)	41 (55.4%)	74
	no	55 (90.2%)	6 (9.8%)	61
West-Netherlands	yes	28 (44.4%)	35 (55.6%)	63
	no	52 (96.3%)	2 (3.7%)	54
North-Netherlands	yes	5 (27.8%)	13 (72.2%)	18
	no	10 (90.9%)	1 (9.1%)	11
Total recruiters*		217	125	342

* 20 recruiters from Belgium were excluded. 144 recruiters indicated retirement, but a majority of them also indicated a location where they (still) work.

Table A5 displays the 342 recruiters that were analysed stratified by outcome, location, and whether or not they live and work in the same municipality. For these 342 entries, the recruiter and his/her recruit:

- indicated that they lived in the Netherlands
- provided a work or study location in the questionnaire.

We used the fitted logistic regression model to estimate probabilities of the outcome (2) for the four regions in the Netherlands (i.e., predictions based on not knowing what recruiter ID is being predicted). Confidence intervals (95%) were calculated by both using fixed-effects uncertainty only, as well as by using fixed effects uncertainty + random effect variance.

Table A6 shows the output of the mixed effect logistic regression. The variable working and living in same municipality significantly influenced the outcome. Table A7 displays the estimated probabilities. Participants living in the North of the Netherlands had the highest probability, namely 0.774 [95% CI 0.433–0.939], to invite a recruit in the municipality where they also work and live.

Table A6. Fixed effects

	Estimate[a]	SE	z value	Pr(>|z|)
Intercept	-3.632	0.749	-4.851	1.231E-06
Middle-Netherlands	0.288	0.521	0.553	5.802E-01
South-Netherlands	0.904	0.641	1.410	1.587E-01
North-Netherlands	0.982	0.819	1.198	2.308E-01
Work and live in same municipality	3.882	0.764	5.082	***3.730E-07***

[a] AIC 351.9; BIC 375.0; logLik -170.0; deviance 339.9; Random effect variance 3.319 (SE: 1.822); subjects: 223

Table A7. Estimated probability of inviting a contact person in the municipality where recruiter works or studies.

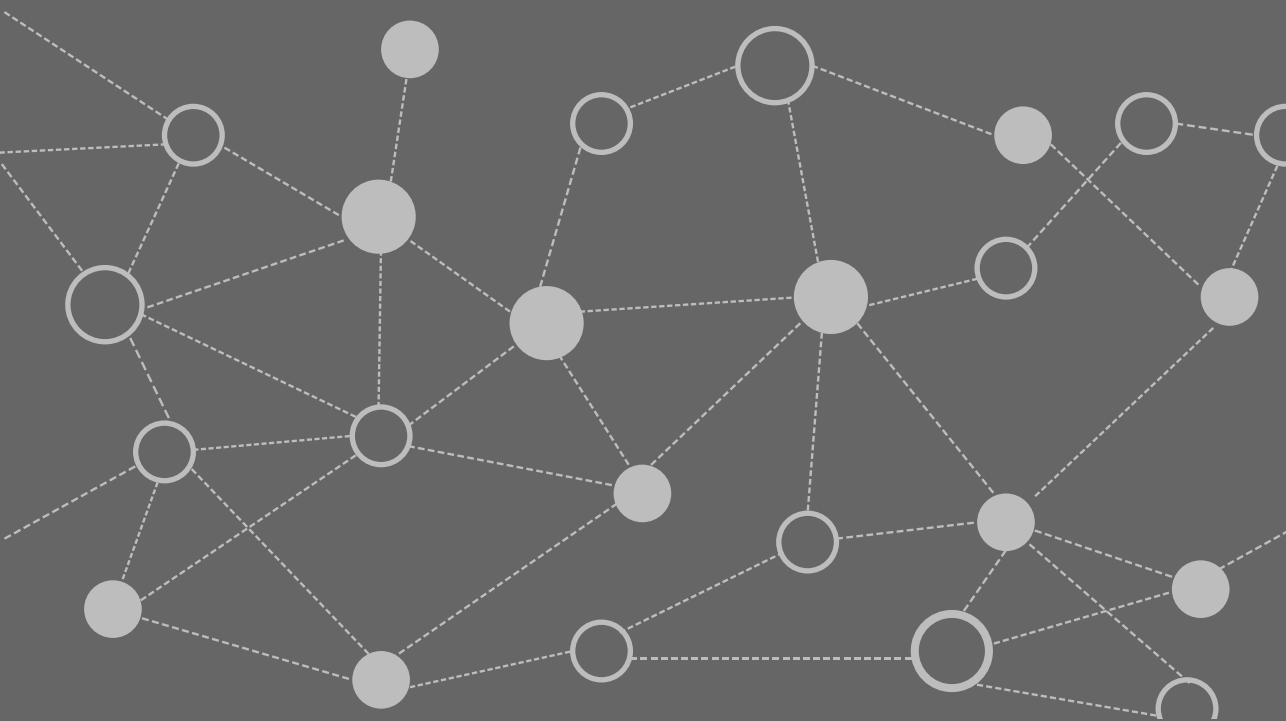
Region	Recruiter works/ studies and lives in same municipality?	Predict. Prob.	CI based on fixed-effects uncertainty ONLY		CI based on FE uncertainty + RE variance	
			2.5%	97.5%	2.5%	97.5%
West-Netherlands	no	0.026	0.006	0.106	0.001	0.576
Middle-Netherlands	no	0.034	0.009	0.121	0.001	0.633
South-Netherlands	no	0.061	0.016	0.209	0.001	0.764
North-Netherlands	no	0.066	0.012	0.290	0.001	0.801
West-Netherlands	yes	0.562	0.361	0.745	0.030	0.982
Middle-Netherlands	yes	0.631	0.439	0.789	0.040	0.986
South-Netherlands	yes	0.760	0.513	0.905	0.066	0.993
North-Netherlands	yes	0.774	0.433	0.939	0.062	0.994

References

1. Newman MEJ. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005;46:323-51.
2. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. BioScience. 2001;51(5):341-52.
3. Hilbe JM. Alternative variance parameterizations: Poisson inverse Gaussian regression. Negative Binomial Regression. 2nd ed. New York: Cambridge University Press; 2011. p. 341-3.
4. Dean C, Lawless JF, Willmot GE. A mixed poisson-inverse-Gaussian regression model. The Canadian Journal of Statistics. 1989;17(2):171-81.
5. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine. 2008;5(3):e74.
6. Van de Kassteele J, Van Eijkeren J, Wallinga J. Efficient estimation of age-specific social contact rates between men and women. In preparation.
7. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M. Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiology and infection. 2008;136(6):813-22.

Part II

Identifying cases with online respondent driven detection



Chapter 5

Enhancing syndromic surveillance with online respondent-driven detection

Mart L. Stein[1,2], Jim E. van Steenbergen[2,3], Vincent Buskens[4], Peter G.M. van der Heijden[4,5], Carl E. Koppeschaar[6], Linus Bengtsson[7,9], Xin Lu[7,8,9], Anna E. Thorson[7], Mirjam E.E. Kretzschmar[1,2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[3] Centre for Infectious Diseases, Leiden University Medical Centre, Leiden, The Netherlands

[4] Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, The Netherlands

[5] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom

[6] Science in Action BV, Amsterdam, The Netherlands.

[7] Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

[8] College of Information System and Management, National University of Defense Technology, Changsha, China

[9] Flowminder Foundation, Stockholm, Sweden

American Journal of Public Health 2015, Vol 105, No. 8

ABSTRACT

Objectives

We investigated the feasibility of combining an online chain recruitment method (respondent-driven detection) and participatory surveillance panels to collect previously undetected information on infectious diseases via social networks of participants.

Methods

In 2014, volunteers from 2 large panels in the Netherlands were invited to complete a survey focusing on symptoms of upper respiratory tract infections and to invite 4 individuals they had met in the preceding 2 weeks to take part in the study. We compared sociodemographic characteristics among panel participants, individuals who volunteered for our survey, and individuals recruited via respondent-driven detection.

Results

Starting from 1015 panel members, the survey spread through all provinces of the Netherlands and all age groups in 83 days. A total of 433 individuals completed the survey via peer recruitment. Participants who reported symptoms were 6.1% (95% confidence interval = 5.4-6.9) more likely to invite contact persons than were participants who did not report symptoms. Participants with symptoms invited more symptomatic recruits to take part than did participants without symptoms.

Conclusions

Our findings suggest that online respondent-driven detection can enhance identification of symptomatic patients by making use of individuals' local social networks.

BACKGROUND

Syndromic surveillance provides information necessary to monitor trends in disease incidence and implement and evaluate response plans[1,2]. To date, most efforts have focused on developing systems based on data from inpatient and ambulatory care health records[3]. In a majority of high-income countries, including the Netherlands, influenza surveillance is based on a combination of reports of influenza-like illness (ILI) collected by sentinel surveillance clinics and additional microbiological testing of subgroups of symptomatic patients[4]. This type of system excludes symptomatic patients who do not visit a general practitioner, and such patients are likely to account for the majority of cases in most influenza outbreaks[5].

Many communicable diseases (e.g., influenza, severe acute respiratory syndrome, measles) spread largely between socially connected individuals, such as household members and schoolchildren, and they often occur in clusters[6,7]. Therefore, cases of infection are expected to cluster in social networks (i.e., contacts of an infected individual are infected at a level of probability higher than that expected if the distribution was random), and clusters can be detected via local social networks of individuals reporting symptoms.

Increased Internet use facilitated the emergence of participatory surveillance (PS) systems, which enable real-time monitoring of diseases through regular submission of syndromic information by volunteers[8,9]. These systems provide information that is not collected in regular surveillance, such as the proportion of symptomatic individuals who actually visit a general practitioner and the proportion who are hospitalized.

To test the feasibility of eliciting information about infections in local networks of symptomatic individuals, we combined a chain recruitment method with existing online PS platforms. Under certain conditions, such a recruitment method permits stepwise and controlled sampling of contacts of contacts, and so forth, in social networks in the general population[10]. We asked PS volunteers to complete a questionnaire and to invite their contacts into the study. In this way, we collected data on chains of contacts to analyze whether other symptomatic individuals could be detected via the local social network of symptomatic respondents. Our aims were to determine whether respondents can be recruited via respondent-driven detection, to report on which individuals are reached, and to assess whether there is clustering of symptomatic patients.

METHODS

Between March and June 2014, we invited volunteers from 2 Internet PS panels in the Netherlands to complete Internet-based surveys focusing on upper respiratory tract infections. We asked them, in completing the surveys, to provide information on their symptoms and to invite 4 individuals with whom they had had face-to-face contact in the preceding 2 weeks to participate in our study (Figure 1). Participants could invite contacts via e-mail, by providing their own e-mail address and receiving 4 invitations for forwarding (indirect invitations), or by providing e-mail addresses of contacts who were then invited via the system (direct invitations). They could also invite friends via a private Facebook message. Individuals were able to opt out and provide reasons for not participating.

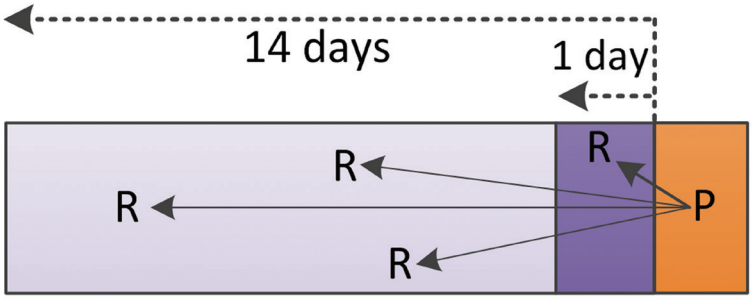


Figure 1. Differences between contact persons and recruits in the sample.
Note. The orange area indicates the day of participation. A participant (P) was asked to report the number of contacts from the preceding day (dark purple area). The light purple area indicates contacts met 2–14 days before the participation day. Participants were asked to invite 4 recruits (Rs). A participant could invite a recruit (R) met either the preceding day or preceding 2–14 days.

We use panel A to refer to a PS system collecting ILI data during the winter seasons[11]. Each week, registered volunteers were reminded via an electronic newsletter to report any symptoms they had experienced since their last log-in. Panel A's 12,957 active volunteers in the Netherlands were invited to participate in the respondent-driven detection survey, and repeated requests (a total of 3) asking them to do so were placed in the weekly newsletter they received[12]. Panel B refers to a comparable system collecting data on pneumonia; the 1691 volunteers in this panel were first invited to participate in the respondent-driven detection survey with a bulletin and then were reminded once as part of their regular newsletter[13]. The majority of panel A members were healthy volunteers of various backgrounds, whereas panel B consisted primarily of patients with asthma and chronic obstructive pulmonary disease.

"Seeds" indicate volunteers from the 2 panels, and "recruits" were invited contacts who completed the questionnaire. Waves denote consecutive subsamples (seeds in wave 0, recruits invited by seeds in wave 1, and so forth). Network trees denote chains of connected

respondents. We refer to seeds enrolled via panels A and B as the ASeed and BSeed groups, respectively; recruits in consecutive waves as the ARec and BRec groups; and the overall samples as ASeedRec and BSeedRec. After completing their questionnaires, participants were referred to a research Web site displaying the latest results (e.g., anonymous network trees). ASeedRec participants who completed the survey had the opportunity to join a raffle for 1 of 10 gift cards of €25. More details on the survey system are provided in supplementary materials A and in Stein et al[10].

Questionnaire

Participants were asked to provide the number of contacts with whom they had interacted (i.e., with whom they had touched or talked within a distance of about one arm's length) during 1 full day ("yesterday"). Contacts needed to be specified by age group and location (contact at home, work, or school; at the house of friends or family; and in other places). Participants were asked to report, from a predefined list, any clinical symptoms they had experienced during the preceding 2 weeks (Table A) Those with symptoms were asked for day of onset, symptom duration, presumed disease, whether they stayed at home, whether they visited a general practitioner, whether they had used any medicines, and whether they knew any people (not restricted to contacts from the preceding day) who had similar symptoms in the past 2 weeks. For each participant, we also collected information on age, gender, education, postal code, household size (including age groups of household members), influenza vaccination status, work or study location, and contact with groups at high risk for influenza infection during a regular workday. If they so chose, participants could complete questionnaires for their children.

Data Analysis

We excluded participants living outside the Netherlands. We compared the sociodemographic characteristics of PS participants, individuals who volunteered in our survey (seeds), and individuals recruited via respondent-driven detection. Also, we conducted the Pearson χ^2 test to assess the samples' representativeness relative to the general population. Data on the demographic characteristics of the general population were obtained from Statistics Netherlands[14]. We computed marginal effects via multilevel logistic regression for the combined samples to assess which sociodemographic characteristics were associated with invitations to contacts (supplementary materials A). To analyze the geographical spread of recruitment, we plotted seeds and total samples separately at the 4-digit postal code level. We used the great circle distance computation (i.e., the shortest distance between 2 points on the surface of a sphere, measured along the surface of the sphere) to assess the distance participants commuted between home (postal code) and their work or study locations. Geocoding was used to convert entered locations into coordinates.

We defined ILI as a combination of fever and at a least headache or muscle pain and at least a cough or sore throat. This definition was similar to the one used for panel A, although that definition also included sudden onset of symptoms and a fever of at least 38° Celsius[12]. A common cold was defined as a runny nose, cough, and sore throat. Participants with ILI or a common cold who reported symptoms that began more than 3 weeks before they participated were excluded. We computed secondary attack rates (i.e., proportion infected among assumed susceptible contacts of an infected participant) in the affected households by assuming that only one household member was a patient with a primary community-acquired case. We analyzed whether symptomatic participants recruited more other symptomatic participants than those who reported no symptoms. We compared symptomatic contacts recruited by symptomatic seeds and symptomatic contacts recruited by asymptomatic seeds with respect to whether they had "at least one symptom" or whether they had a common cold, fever, and ILI. We also compared ILI incidence rates based on regular surveillance by the Dutch Sentinel Network (NIVEL[15]) with estimates obtained from panel A and our survey population. Statistical analyses were performed in R (version 3.1.1)[16].

RESULTS

Overall, 1448 individuals (1177 in ASeedRec and 271 in BSeedRec) completed our questionnaire during March to May 2014 (Figure 2; see also Figure A). We excluded 288 responses for the following reasons: multiple participation (19), residence outside the Netherlands (115), and incomplete information (154; supplementary materials A). Our survey was distributed within all provinces in the Netherlands (Table B). The age distributions of both panels differed from that of the general population in that they were shifted toward older age groups. Ages varied from 3 to 97 years in ASeedRec and 17 to 82 years in BSeedRec (Figure B and Table C). The age distributions of ASeedRec and BSeedRec differed significantly (Kolmogorov–Smirnov $p < .001$). Relative to the general population, women, those with higher levels of education, those with a household composed of 2 adults, and those who had been vaccinated against the flu (self-reported) were strongly overrepresented in ASeedRec, BSeedRec, and panels A and B (Table 1).

Participants reporting symptoms were 6.1% (95% confidence interval [CI] = 5.4%-6.9%) more likely to invite contacts to take part than were participants without symptoms. Those with a bachelor's degree or higher were 5.2% (95% CI = 4.6%-5.8%) more likely to invite contacts than those at lower educational levels. Men were 8.6% (95% CI = 7.6%-9.6%) less likely than women to invite contacts, and participants recruited via panel B were 6.4% (95% CI = 5.7%-7.2%) less likely to do so than those recruited via panel A. The latter difference might have been due to the incentive offered to panel A participants. Participants' age did not seem to influence their probability of inviting contacts to take part.

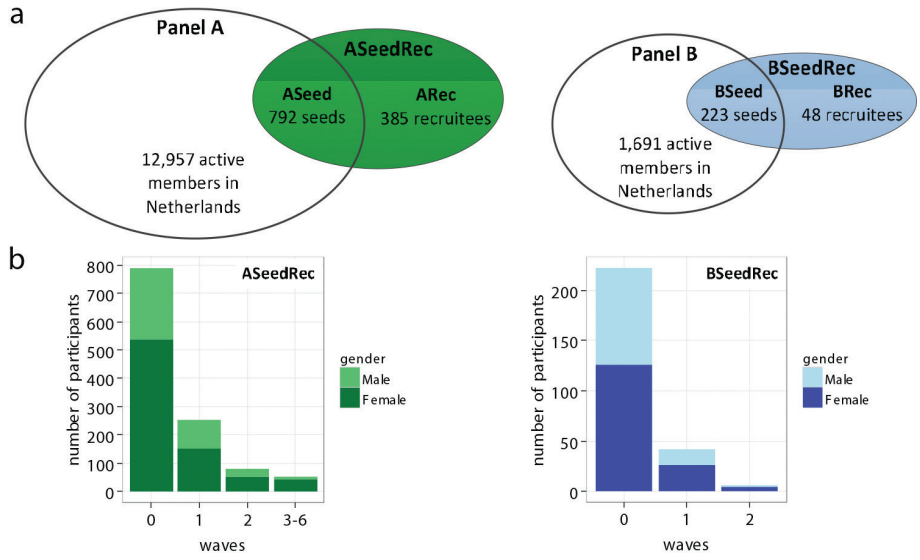


Figure 2. Overview of the sample composition by (a) participatory surveillance panels, seeds, and recruits taking part in the study, and **(b)** number of participants by wave and gender: The Netherlands, 2013–2014 influenza season.

5

Overall, 171 panel volunteers and invited contacts opted out via the link in the invitation or the survey Web site; 109 of these individuals provided one or more reasons, of whom 88.3% indicated that they did not want to invite or provide information about their contacts. Five (4.6%) individuals indicated that they had participated before and had received an invitation from another person in the same network tree.

Recruitment via Panel A

A total of 792 seeds were enrolled via panel A (the ASeed group; 6.1% of panel A). Overall, 385 recruits completed the questionnaire in 165 network trees (the ARec group); 30.9% of these trees had 2 or more waves, and one of the trees reached 6 waves. On average, ARec recruits invited more contacts than individuals in ASeed (Table D). In ASeedRec, a total of 1802 invitations were sent out via the recruitment page (Table E). Because all panel volunteers were invited in batches spread over 1 week, we do not know exactly how fast volunteers responded to invitations; however, 443 (55.9%) seeds participated within the first week after the initial invitation was sent. Recruits in ARec responded, on average, within 2.1 days (SD = 4.1); 43.1% did so on the day of the invitation.

Table 1. Comparisons of Sociodemographic and Health-Related Indicators: General Dutch Population and the Study Samples, 2014

Indicator	Dutch Population,[a] % or Mean (SD)	Panel A[b]		ASeedRec Sample		Panel B		BSeedRec Sample	
		% or Mean (SD)	*p*	% or Mean (SD)	*p*	% or Mean (SD)	*p*	% or Mean (SD)	*p*
Female	50.5	57.4	< .001	66.5	< .001	64.8	< .001	57.6	.023
Age, y	40.2 (23.0)	52.0 (16.2)	. . .	53.0 (14.8)	. . .	55.0 (13.7)	. . .	56.8 (12.6)	. . .
Educational level, bachelor's degree or higher	18.9[c]	56.1	< .001	58.9	< .001	43.2	< .001	50.6	< .001
Single-member household	16.7[d]	19.5	< .001	26.2[e]	< .001	17.8	.238	20.7[e]	.095
Household with only adults	25.7[d]	42.8	< .001	48.3	< .001	51.3	< .001	59.4	< .001
Household with children	24.3[d]	27.9	< .001	25.6	.344	30.9	< .001	19.9	.104
Daily contact with patients	12.2	9.6	< .001	11.2	.345	. . .	. . .	11.4	.788
Vaccinated against seasonal influenza in past 12 month	23.8	36.1	< .001	37.9	< .001	48.4	< .001	56.4	< .001
Asthma or chronic obstructive pulmonary disease	7.4	12.4f	< .001	3.6[g]	< .001	34.4	< .001	20.7[g]	< .001
Allergy	8.5	55.9	< .001	6.6[g]	.024	32.9[h]	< .001	7.7[g]	.738

[a]Based on data provided by Statistics Netherlands.[14]
[b]Based on information from Influenzanet.[12]
[c]With respect to educational level, Statistics Netherlands[14] provides data on the population aged 15–64 years.
[d]StatLine provides information only on the number of children in the household, regardless of age.
[e]Forty ASeedRec participants and 6 BSeedRec participants did not provide information on household members and were assumed to live alone.
[f]Also includes panel A volunteers with lung disease.
[g]We asked participants about these risk conditions only if they reported symptoms.
[h]Panel B volunteers were asked about hay fever and allergy.

Seeds lived in a total of 636 (15.7%) of a possible 4061 postal code areas; we reached another 140 (3.4%) areas through recruitment (Figure C). In the ARec group, 141 (36.6%) members had the same postal code as their recruiter. In ASeedRec, 800 (68.0%) participants provided a school or work location outside their home. On average, these participants commuted within a radius of 12.2 kilometers (range = 0.03-179.8); 73.3% commuted within a 15-kilometer radius, and 86.0% commuted within their province of residence.

Panel A's mean age of 52.0 years was higher than that of the general population; consequently, ASeed members also had a high mean age of 54.9 years. The mean age of the ARec group was 49.0 years, indicating ASeed members' recruitment of individuals from a younger age group. Related to the high mean age was the overrepresentation in panel A and ASeedRec of households with only 2 adults (ARec contained 9.4% more individuals with a household of 4 or more members than ASeed). The overrepresentation of women and highly educated individuals in ASeed relative to the general population decreased through peer recruitment, with 3.5% and 7.2% in ARec, respectively (Table D).

In ASeedRec, 565 (48.0%) participants reported at least one symptom; 59.5% of these participants reported that they knew at least one contact person with similar symptoms. Symptomatic participants mostly identified similar symptoms among household members, family members, and colleagues (Figure 3a). Headache, muscle pain, and common cold symptoms were most frequently identified among contacts (Figure D). ASeedRec had crude attack rates of 7.5% for common colds and 2.5% for ILI. We estimated corresponding secondary attack rates of 29.7% and 43.2% (Table F).

Overall, 17% (p = .001) more contacts with at least one symptom were recruited by ASeed members with symptoms than by members without symptoms (Table 2). Similar results (although less strong) were found for common colds (10.9%; p = .061) and fever (14.3%; p = .015). There were 28 ASeed members with ILI, and only one contact with ILI was recruited by a seed without ILI. As a result of the participation of a relatively high proportion of panel A volunteers with ILI in ASeed, the observed proportion of symptomatic individuals in ASeedRec in March 2014 was higher than the proportion in panel A by a factor of 10.4 and higher than the proportion in NIVEL by a factor of 68.8[17].

Among the 29 ILI patients, 23 reported in total 73 contacts with similar symptoms. Participants in ASeedRec reported a total of 22204 contacts (mean = 18.9 contacts, median = 10.0, SD = 27.2). By taking participants' number of contacts and contacts with similar symptoms into account, we computed the proportions of ASeedRec members and reported contacts with symptoms during the month of March; these proportions were higher by a factor 1.8 relative to panel A and by a factor of 11.8 relative to NIVEL (Figure 3b).

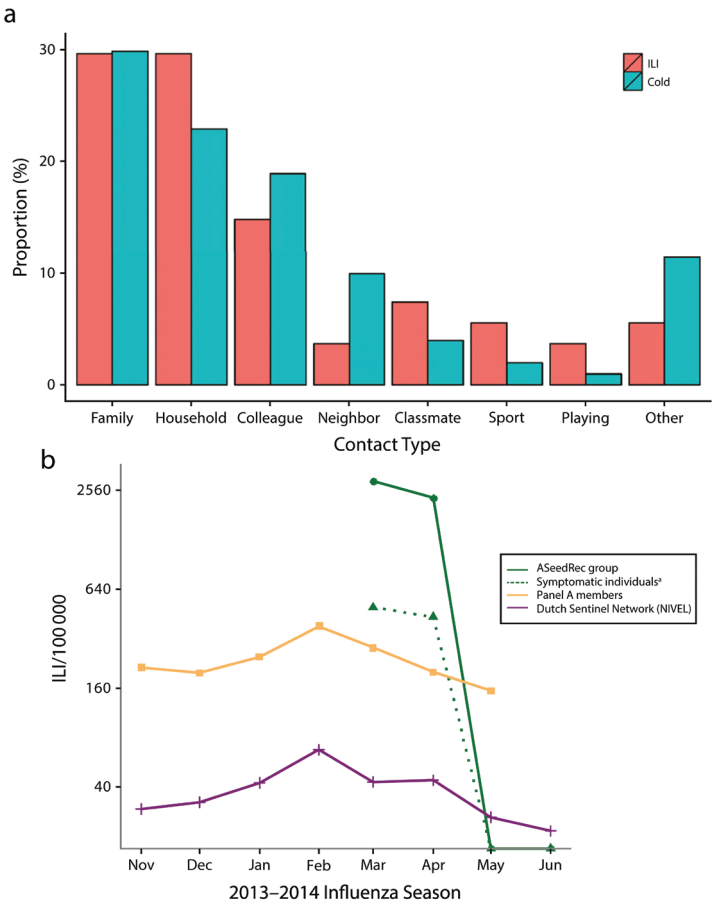


Figure 3. Reported symptoms by influenza-like illness (ILI) and common cold by (a) contacts with similar symptoms identified by symptomatic ASeedRec group members stratified by contact type, and (b) cases of ILI per 100,000 individuals: The Netherlands, 2013–2014 influenza season.

Note. According to national criteria, ILI became epidemic in week 2 of 2014, when the incidence exceeded 51 per 100,000 population. NIVEL defined the incidence between weeks 5 and 8 of 2014 as a mild ILI epidemic.

[a]The dotted green line indicates the proportions of symptomatic individuals in ASeedRec and among reported contacts.

Recruitment via Panel B

A total of 223 seeds were enrolled via panelB (the BSeed group; 13.2% of panel B). In total, 48 recruits completed the questionnaire in 29 network trees (the BRec group); 17.2% of these trees reached 2 waves. On average, recruits in BRec invited more contacts than seeds. A total of 340 invitations were sent in BSeedRec (Table E). The majority of BSeed members (95.5%) participated in the first week after receiving the initial invitation. BRec members responded on average within 2.6 days (SD = 5.3); 41.7% responded on the day of the invitation.

BSeedRec covered 233 (5.7%) postal code areas, indicating that almost every participant lived in a different area (Figure C). Twenty-two BRec recruits (45.8%) had the same postal code as their recruiter. In BSeedRec, 115 (42.4%) participants provided a study or work location outside their place of residence. On average, these participants commuted within a 11.4-kilometer radius (range = 0.1-83.3); 79.1% commuted within a 15-kilometer radius, and 81.7% commuted within their province of residence.

Relative to the general population, women were slightly overrepresented in the BSeed group (56.5%), and, as a result of the recruitment of predominantly women, this percentage increased to 62.5% in BRec. The mean age of panel B was relatively high, and thus BSeed also had a high mean age of 57.8 years. The mean age of BRec was 52.4 years, indicating BSeed's recruitment from a younger age group. Although the mean household size of BSeed was similar to that of the general population, the mean size increased in BRec because a few recruits had relatively large households (Table D).

In BSeedRec, 161 (59.4%) participants reported at least one symptom, and 55.3% of these participants indicated at least one contact person with similar symptoms. Symptoms were mostly identified among household members, family members, colleagues, and school classmates. Headache, muscle pain, and common cold symptoms were again most frequently identified among contacts, although these results were less clear than with the ASeedRec sample owing to the small numbers reported (Figure D). BSeedRec had overall attack rates of 8.9% for common colds and 1.1% for ILI (Table F). The estimated secondary attack rate for common colds was 36.8%. Seeds with ILI did not report any household members with similar symptoms, and they did not recruit any contacts (Table G).

Table 2. Detection of symptoms in network chains of seeds in the ASeedRec sample: The Netherlands, 2014

Symptom category, no. (%)	Seeds[a] (A)		Recruits by Seeds with Symptoms (B)		Recruits by Seeds Without Symptoms (C)		*p*[b]		
	Seeds with symptoms	Seeds w/o symptoms	Recruits with symptoms	Recruits w/o symptoms	Recruits with symptoms	Recruits w/o symptoms	A vs B	A vs C	B vs C
One or more symptoms	71 (43.0)	94 (57.0)	109 (69.0)	49 (31.0)	118 (52.0)	109 (48.0)	< .001	.099	.001
Common cold symptoms	16 (9.7)	149 (90.3)	4 (16.7)	20 (83.3)	21 (5.8)	340 (94.2)	.292	.14	.061
Fever	12 (7.3)	153 (92.7)	4 (18.2)	18 (81.8)	14 (3.9)	349 (96.1)	.101	.143	.015
Influenza-like illness	8 (4.8)	157 (95.2)	0 (0.0)	12 (100.0)	1 (0.3)	372 (99.7)	> .999	< .001	> .999

[a]Only successful seeds (n = 165) were considered in this analysis (seeds who invited a recruit who also completed the survey). Recruits in waves 1–6 were lumped together (n = 385).
[b]We used the 2-sample chi-square test for equality of proportions with continuity correction to estimate p values. The Fisher exact test was used for contingency tables containing small values (n < 10).

DISCUSSION

PS systems collect data in real time and include infected individuals who do not seek health care[9]. Here we have described, for the first time (to our knowledge), the feasibility of sampling volunteers via PS for an online respondent-driven survey. Beginning with PS volunteers, our survey was distributed via peer-driven recruitment in several waves through all Dutch provinces and reached, within a short period, individuals from all age groups, those with a wide range of household compositions, and those at a variety of educational levels. Neither PS panel was representative of the general population in terms of basic demographic characteristics; after recruitment, however, the representativeness of the overall sample in terms of age and gender improved slightly.

Combining online communities with respondent-driven detection might enhance the identification of symptomatic patients not detected via conventional surveillance systems. Such information, combined with data derived from regular surveillance, can improve estimations of severity indices (e.g., probability of hospitalization after development of symptoms), especially for infectious disease outbreaks in which the majority of symptomatic patients do not seek health care[18-20]. Through respondent-driven detection, we increased the geographical coverage of our ASeed group, in that recruits mostly resided in regions other than those of recruiters. Seeds with symptoms recruited more symptomatic contacts than asymptomatic seeds, at least with regard to experiencing general symptoms, common colds, and fever. Symptomatic participants mostly reported similar symptoms among their close contacts. This might have been due to higher transmission rates among close contacts but also higher recall rates with respect to contacts seen most often.

The findings just described indicate that recruitment of peers by symptomatic participants led to higher rates of detection of other symptomatic patients. This is supported by the fact that symptomatic participants seemed more likely to invite contacts than participants who did not report symptoms. Possibly, recent experiences of symptoms motivate individuals to recruit others; similarly, it has been observed that PS participation rates are related to illness status[8]. The majority of the volunteers responded within 1 week after being invited, and many recruits responded the same day they were invited, suggesting that information on symptoms and behaviors can be quickly and efficiently obtained.

According to different criteria, the participation rate of panel volunteers can be judged as either low (considering that 90.7% of panel A members reported information on their health status 3 or more times) or high (considering that no massive communication campaign was implemented). The invitation of panel B members via a special bulletin and the difference in target groups (e.g., volunteers in panel B had more chronic health conditions) could explain

the differences in participation rates between the panels.

Volunteers who enrolled were mostly women and were more highly educated than the overall population; overrepresentation of those groups relative to the general population is common in PS systems[8,9]. Although we provided participants with the option of completing the survey for their children, only a few did so, and children were underrepresented in our samples. In addition, elderly individuals were overrepresented in our samples and PS panels. This high-risk group receives yearly notifications to obtain a flu vaccination and might be more interested than younger groups in influenza-related topics. This, in combination with our samples' relatively high mean ages, might also explain the overrepresentation of vaccinated individuals and households with only adults.

Limitations

Our study involved limitations. Less than half of all seeds in the 2 samples invited a contact person, a proportion not sufficient to generate long recruitment chains. Also, this percentage was lower than during our earlier research, in which seeds were first contacted personally[21]. Although, similar to earlier online respondent-driven surveys[22], we used an incentive in ASeedRec, only a slightly higher recruitment rate was observed in this group than in BSeedRec. Concerns about privacy or not wanting to bother acquaintances with a questionnaire were reported and withheld some participants from sending invitations to contacts. Even though the majority of Internet users share information with each other via social media[23], sending survey invitations specifically to a few contacts is a step many participants did not want to take.

Previously, overall attack rates of 2.5% (95% CI = 2.1%, 3.2%; based on NIVEL[5]) and 29.2% (95% CI = 21.6%, 37.9%; based on PS[24]) were estimated for a typical Dutch influenza season. In ASeedRec, only one recruit reported ILI, and a crude attack rate of 2.5% was observed. Generalization of sample estimates to the general population requires weighting (e.g., for age) to enable a proper comparison with NIVEL data. The proportion of ILI in our sample increased as a result of the participation of a select group of PS volunteers with ILI symptoms. Although we were unable to determine the true enhancement factor in the proportion of symptomatic ILI patients via respondent-driven detection, we did observe a slight enhancement in case detection relative to ILI surveillance when participants' numbers of contacts and contacts with similar symptoms were taken into account.

The probability of identification of disease through respondent-driven detection depends on numerous factors, especially the type of disease being assessed, the incidence of the disease, and methodological aspects such as recruitment of and by symptomatic seeds. Our survey was launched after the peak period of a relatively mild influenza season and during an interval in which the number of active panel A volunteers was declining[12]. Only 8 of the 28 seeds with

ILI actually invited a recruit, and only 12 recruits completed the questionnaire. Motivation and participation might be much higher during the increasing phase of an influenza season and with other perceived threats of emerging infections.

Conclusions

Our findings in this novel combination of respondent-driven detection with a large PS system provide insights into which groups are reached and indicate that an increased number of symptomatic patients can be detected when the underlying connections in a local social network are used. Although online peer recruitment involves challenges, we demonstrated that respondent-driven detection through PS with large geographical coverage is possible and that timely, detailed information about respondents and their contacts can be obtained. Repeating this type of study during the epidemic peak of a more severe influenza season will provide more information on the extent to which respondent-driven detection enhances disease surveillance.

Contributors

M.L. Stein, J.E. van Steenbergen, and M.E.E. Kretzschmar originated and designed the study and developed the questionnaires. M.L. Stein, J.E. van Steenbergen, C.E. Koppeschaar, and M.E.E. Kretzschmar performed the study. M.L. Stein, J.E. van Steenbergen, V. Buskens, P.G.M. van der Heijden, and M.E.E. Kretzschmar analyzed the data. M.L. Stein, L. Bengtsson, and A. Thorson developed the Web-based respondent-driven survey system. M.L. Stein, J.E. van Steenbergen, V. Buskens, P.G.M. van der Heijden, and M.E.E. Kretzschmar wrote the article.

Acknowledgments

This study was conducted within the Utrecht Center for Infection Dynamics. The Swedish Research Council financed the programming of the online respondentdriven survey system. We are grateful to Martin Camitz, Antwan Wiersma, and Ronald Smallenburg for their help with the survey launches. Note. The funders had no role in the study design, data collection and analysis, or the decision to prepare or publish the article.

Human Participant Protection

This study was approved by the Medical Ethical Committee of the University Medical Center Utrecht. Participants provided informed consent before completing the survey.

REFERENCES

1. Henning KJ. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004;53(suppl):5-11.

2. van den Wijngaard CC, van Pelt W, Nagelkerke NJ, Kretzschmar M, Koopmans MP. Evaluation of syndromic surveillance in the Netherlands: its added value and recommendations for implementation. Euro Surveill. 2011;16:9.

3. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J AmMed Inform Assoc. 2004;11(2):141-150.

4. Fourquet F, Drucker J. Communicable disease surveillance: the sentinel network. Lancet. 1997;349: 794-795.

5. McDonald SA, Presanis AM, De Angelis D, et al. An evidence synthesis approach to estimating the incidence of seasonal influenza in the Netherlands. Influenza Other Respir Viruses. 2014;8(1):33-41.

6. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008;5(3):e74.

7. Christakis NA, Fowler JH. Social network sensors for early detection of contagious outbreaks. PLoS One. 2010;5(9):e12948.

8. Wojcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol. 2014;11:7.

9. Paolotti D, Carnahan A, Colizza V, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin Microbiol Infect. 2014;20(1):17-21.

10. Stein ML, van Steenbergen JE, Chanyasanha C, et al. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PLoS One. 2014;9(1): e85256.

11. Marquet RL, Bartelds AI, van Noort SP, et al. Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season. BMC Public Health. 2006;6:242.

12. Influenzanet. Influenzanet home page. Available at: https://www.influenzanet.eu. Accessed April 23, 2015.

13. De GroteLongontstekingMeting. Science in action. Available at: www.degrotelongontstekingmeting.nl. Accessed April 23, 2015.

14. Statistics Netherlands. StatLine. Available at: http://statline.cbs.nl/Statweb. Accessed April 23, 2015.

15. Netherlands Institute for Health Services Research. Primary care database. Available at: http://www.nivel.nl/en/nivel-primary-care-database-Sentinel-Practices%20. Accessed April 23, 2015.

16. R Core Team. R: a language and environment for statistical computing. Available at: http://web.mit.edu/r_v3.0.1/fullrefman.pdf. Accessed April 23, 2015.

17. National Institute for Public Health and the Environment. Jaarrapportage surveillance respiratoire infectieziekten 2013. Available at: http://www.rivm.nl/Documenten_en_publicaties/Wetenschappelijk/Rapporten/2014/september/Jaarrapportage_Surveillance_Respiratoire_Infectieziekten_2013. Accessed April 23, 2015.

18. Barboza P, Vaillant L, Le Strat Y, et al. Factors influencing performance of Internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. PLoS One. 2014;9(3):e90536.

19. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. N Engl J Med. 2009;360(21):2153-2155, 2157.

20. Anema A, Kluberg S, Wilson K, et al. Digital surveillance for enhanced detection and response to outbreaks. Lancet Infect Dis. 2014;14(11):1035-1037.

21. Stein ML, van Steenbergen JE, Buskens V, et al. Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling. PLoS One. 2014;9 (11):e113711.

22. Bengtsson L, Lu X, Nguyen QC, et al. Implementation of Web-based respondent-driven sampling among men who have sex with men in Vietnam. PLoS One. 2012;7(11):e49417.

23. Lehmann BA, Ruiter RA, Kok G. A qualitative study of the coverage of influenza vaccination on Dutch news sites and social media websites. BMC Public Health. 2013;13:547.

24. Patterson-Lomba O, Van Noort S, Cowling BJ, et al. Utilizing syndromic surveillance data for estimating levels of influenza circulation. Am J Epidemiol. 2014;179(11): 1394---1401.

SUPPLEMENTARY MATERIALS

Supplementary text A.

Detailed Description Survey System

Seeds registered on the survey website with their email address to obtain a unique token for participation. An email address was required to confirm the registration and avoid automated registration attempts. Informed consent was obtained via the first webpage, on which inter alia study purposes and benefits, and statements on privacy and confidentiality were displayed. Users were able to accept the informed consent form by clicking the 'Start the Survey' button and to continue to the questionnaire, or to deny by clicking the button "No Thanks" whereby users were automatically logged out of the survey. Each survey page contained a link to more background information about the research project and all pages contained a log-out button that referred users to a search engine. The four invitations sent by recruiters contained each a unique token-based URL that gave recruitees direct access to the survey, without having to register with an email address. Tokens were only valid for a single participation and could not be used on two computers at the same time. The tokens allowed us to anonymously follow the online recruitment process. The system converted IP addresses to a unique anonymous code using a one-way encryption algorithm and original addresses were deleted. All communication between the users and the survey website was encrypted.

In case participants chose to provide the four email addresses of their contact persons, the system automatically sent an invitation to these contact persons anonymously (i.e., without revealing the identity of the participant). Participants were also provided with the possibility to add their own name or nick name to the email invitation, but this was not necessary. This direct email method was most frequently used by the participants (see Table S4).

The survey could be filled in on a mobile device. Webpage dimensions were set to automatically adapt to the screen size of the device used. Facebook Send Dialog is not supported for mobile devices and therefore Facebook private messages were not available when a mobile device was used.

Subgroup analysis of uncompleted surveys

A total of 154 persons entered our survey, but did not complete the questionnaire. The reasons why these participants did not complete the survey are unknown, but there may be several explanations such as technical issues (e.g., slow internet connection) or personal issues (e.g., limited time or don't want to answer certain questions). In ASeed, 75 seeds did not complete the survey, compared to 49 persons in ARec. In BSeed, 22 seeds did not complete the survey, compared to 8 in BRec. These participants were excluded from the data analysis in the manuscript. Here we provide information on these individuals.

Table I displays the characteristics of all individuals who did not complete the survey, as far as this information was received from them. Note that the composition of this group (who did not complete the survey) is very similar to participants who completed the survey (see values in Table 2). The only notable difference is the level of education between those who completed and did not complete the survey. Compared to ASeed and ARec in Table 2, it appears that more participants with another educational level than a Bachelor's degree did not complete the survey.

There were 38 participants in ASeedRec (namely 25 in ASeed and 13 in ARec) who did not complete the questionnaire, but did provide information on whether or not they experienced symptoms in the last two weeks. We investigated the effect of these extra symptomatic and non-symptomatic individuals on our main outcome, by reanalyzing the detection of symptoms in network chains of seeds in the ASeedRec sample. See Table II for the results. The analysis resulted in only minor differences in proportions compared to values displayed in Table 4. For the understandability of the various subgroup comparisons in the manuscript, we therefore decided to include only the completed questionnaires in the main data analysis.

Table I. Characteristics of participants who did not complete survey

Variable	ASeed		ARec		BSeed		BRec	
	n_{info}*		n_{info}*		n_{info}*		n_{info}*	
Mean age (median; SD; range)	60	59.6 (61.5; 14.0; 27.0-90.0)	40	49.4 (51.5; 17.5; 3.0-76.0)	14	56.9 (63.5; 18.4; 17.0-80.0)	5	39.4 (35.0; 13.9; 22.0-56.0)
Female	66	45 (68.2%)	41	27 (65.9%)	16	11 (68.8%)	5	4 (80.0%)
Mean household size (median; SD)	58	2.2 (2.0; 1.3)	34	2.5 (2.0; 1.2)	14	2.6 (2.0; 1.4)	4	2.5 (2.5; 1.3)
Education	65	28 (43.1%)	40	16 (40.0%)	15	9 (60.0%)	5	3 (60.0%)
One-or-more symptoms	25	17 (68.0%)	13	10 (76.9%)	5	5 (100%)	2	2 (100%)
Common cold symptoms	25	2 (15.4%)	13	1 (7.7%)	5	2 (0%)	2	0 (0%)
Fever	25	3 (23.1%)	13	1 (7.7%)	5	0 (0%)	2	0 (0%)
Influenza-like-illness (ILI)	25	1 (4%)	13	0 (0%)	5	0 (0%)	2	0 (0%)

*n_{info}= number of participants that provided information for the particular variable.

Table II. Detection of symptoms in network chains of seeds in ASeedRec sample, with uncompleted surveys included.

	Seeds* **(I)**		Recruits by seeds with symptoms **(II)**		Recruits by seeds without-symptoms **(III)**		*P-value*†		
	with symptoms	without symptoms	with symptoms	without symptoms	with symptoms	without symptoms	I vs II	I vs III	II vs III
One or more symptoms	74 (43.3%)	97 (56.7%)	114 (65.9%)	59 (34.1%)	123 (48.6%)	130 (51.4%)	< .001	.238	<.001
Common cold symptoms	16 (9.4%)	155 (90.6%)	4 (16.0%)	21 (84.0%)	22 (5.5%)	379 (94.5%)	.295	.129	.057
Fever	13 (7.6%)	158 (92.4%)	4 (16.0%)	21 (84.0%)	15 (3.7%)	386 (96.3%)	.242	.080	.015
Influenza-like illness (ILI)	8 (4.7%)	163 (95.3%)	0 (0.0)	13 (100%)	1 (0.2%)	412 (99.8%)	1.00	< .001	1.00

*The total number of seeds was 171, who recruited a total of 426 recruitees.

†We used the two-sample chi-square test for equality of proportions with continuity correction to estimate p-values. Fisher's exact test was used for contingency tables containing small values (n<10).

Multilevel logistic regression analysis: drivers of online recruitment

Each participant was asked to invite (maximum) four contact persons at the end of the questionnaire (see Table III). We investigated which socio-demographic characteristics are associated with inviting contact persons by using multilevel logistic regression analysis for the samples ASeedRec and BSeedRec taken together. Multilevel logistic regression was used (instead of ordinary logistic regression) to correct for clustering of observations within network trees.

We considered invitations sent by a participant ('0' or '1 to 4') as binary dependent variable and as independent variables: sex, age (integer), education level ('higher educated'/'other'), household size (integer), personal network size (integer), reporting of symptoms (yes/no), flu vaccination (yes/no) and panel (A/B). The latter was added to account for differences between panels and in approach (e.g. use of incentive for panel A). 'Higher educated' denotes a bachelor degree or higher.

We calculated average marginal effects (AME) in order to easily interpret the effects of independent variables. The marginal effect of an independent variable measures the impact of change of 1 in an independent variable on the estimated probability in the logistic regression model. First we calculated predicted probabilities for individual respondents in the sample using the linear predictor

$$LP = \text{constant} + ß1*\text{sex} + ß2*\text{age} + [\ldots] + \text{random effect},$$

followed by

$$\pi = e^{LP} / (1+ e^{LP}).$$

The individual marginal effect is then defined by

$$\textit{Individual marginal effect for a variable} = \pi*(1- \pi)*ß$$

The AME of a variable is calculated by taking the average over these individual marginal effects[1]. Confidence intervals of the marginal effects were calculated using standard error of these averages. The multilevel logistic regression analysis was performed using the R package 'lme4'. One participant reported a household size of 310, which was excluded from the analysis.

Table IV displays the regression coefficients. Males had a significantly lower log odds (-0.363, p=0.002 [-0.594—0.132]) of inviting contact persons when adjusted for age, education, household size, experiencing symptoms, vaccinated against flu in the past 12 months, degree and panel. Similar results were found for those recruited via panel B (note that no

significant effect was observed). Higher educated and those reporting at least one symptom were more likely to invite their contact persons.

Table V displays the marginal effects for these variables, as displayed in the manuscript.

Reference

1. Scott Long J. Regression Models for Categorical and Limited Dependent Variables. Indiana University, Bloomington: SAGE Publications, Inc; 1997.

Table IV. Parameter estimates of multilevel logistic regression analysis.

	estimate*	SE	z value	Pr(>|z|)	2.5%	97.5%
Constant	-0.538	0.287	-1.877	0.060	-1.100	0.024
male	-0.363	0.118	-3.078	***0.002***	-0.594	-0.132
age	0.000	0.004	-0.067	0.947	-0.009	0.008
higher educated	0.219	0.111	1.964	***0.049***	0.000	0.437
household size	0.032	0.033	0.980	0.327	-0.032	0.096
symptoms	0.259	0.112	2.314	***0.021***	0.040	0.478
vaccinated	0.107	0.123	0.873	0.382	-0.133	0.348
degree	0.000	0.002	0.061	0.952	-0.003	0.004
panel B	-0.271	0.146	-1.853	0.064	-0.557	0.016

*Deviance: 1943.3; logLik: -971.6; df.resid: 1438; AIC: 1963.3; BIC: 2016.0; random effect (intercept) variance 0.097 (Std. Dev. 0.311).

Table V. Average marginal effects.

	dF/dx	95% CI
males	-0.086	-0.096 – -0.076
age	-0.000	-0.000 – -0.000
higher educated	0.052	0.045 – 0.058
household size	0.007	-0.007 – 0.008
symptoms	0.061	0.054 – 0.069
vaccinated	0.025	-0.022 – 0.028
degree	-0.000	-0.000 – 0.000
panel B	-0.064	-0.072 – -0.057

Table III. Correlations between any two linked individuals in the same network tree.

Number of invitations sent	Sex	Education	Symptoms	Vaccinated	Via panel A	Via panel B
sent no invitations (0)	Female	Other education	0	0	53	11
				1	35	12
			1	0	62	17
				1	39	14
		Higher educated	0	0	83	6
				1	43	10
			1	0	80	13
				1	24	12
	Male	Other education	0	0	36	6
				1	28	4
			1	0	19	3
				1	15	21
		Higher educated	0	0	51	10
				1	38	17
			1	0	35	5
				1	26	11
sent 1 to 4 invitations	Female	Other education	0	0	35	4
				1	29	7
			1	0	59	12
				1	25	9
		Higher educated	0	0	68	1
				1	31	2
			1	0	75	18
				1	42	8
	Male	Other education	0	0	15	5
				1	15	3
			1	0	7	1
				1	12	5
		Higher educated	0	0	25	2
				1	27	10
			1	0	28	4
				1	17	8
					1177	**271**

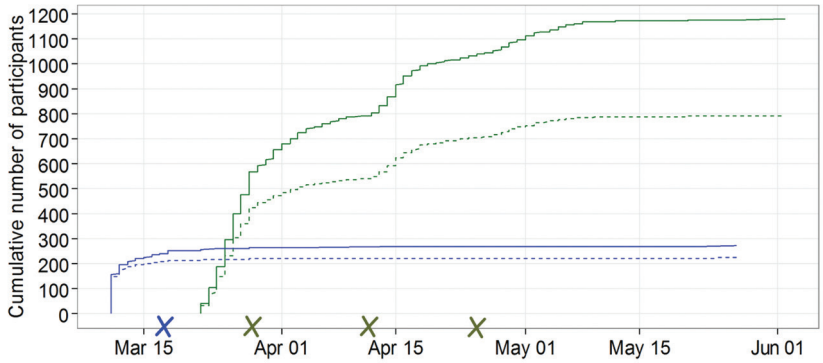


Figure A. Participation by number of days the survey was active. Solid line indicates cumulative percentage of participants and dotted lines the seeds. The green lines indicate the ASeedRec sample and the blue lines the BSeedRec sample. The survey was launched via panel A on March 22 and via panel B on March 11. The days on which the panels were reminded are shown with crosses, in green for ASeedRec (3 times) and in blue for BSeedRec (1 time): The Netherlands, 2013-2014 influenza season.

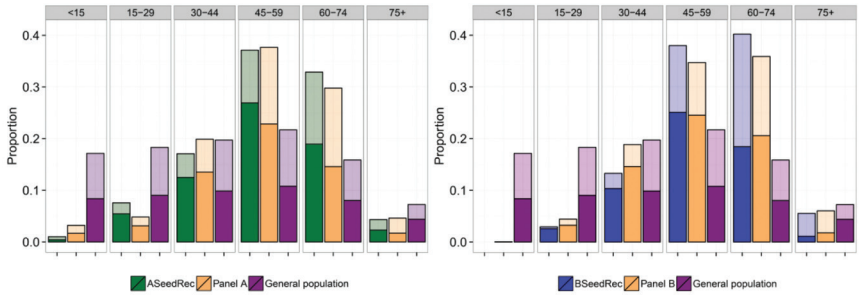


Figure B. Comparison of age groups stratified by sex. Left: age distribution stratified by sex of ASeedRec, panel A and the general population. Right: age distribution of BSeedRec, panel B and the general population: The Netherlands, 2013-2014 influenza season.

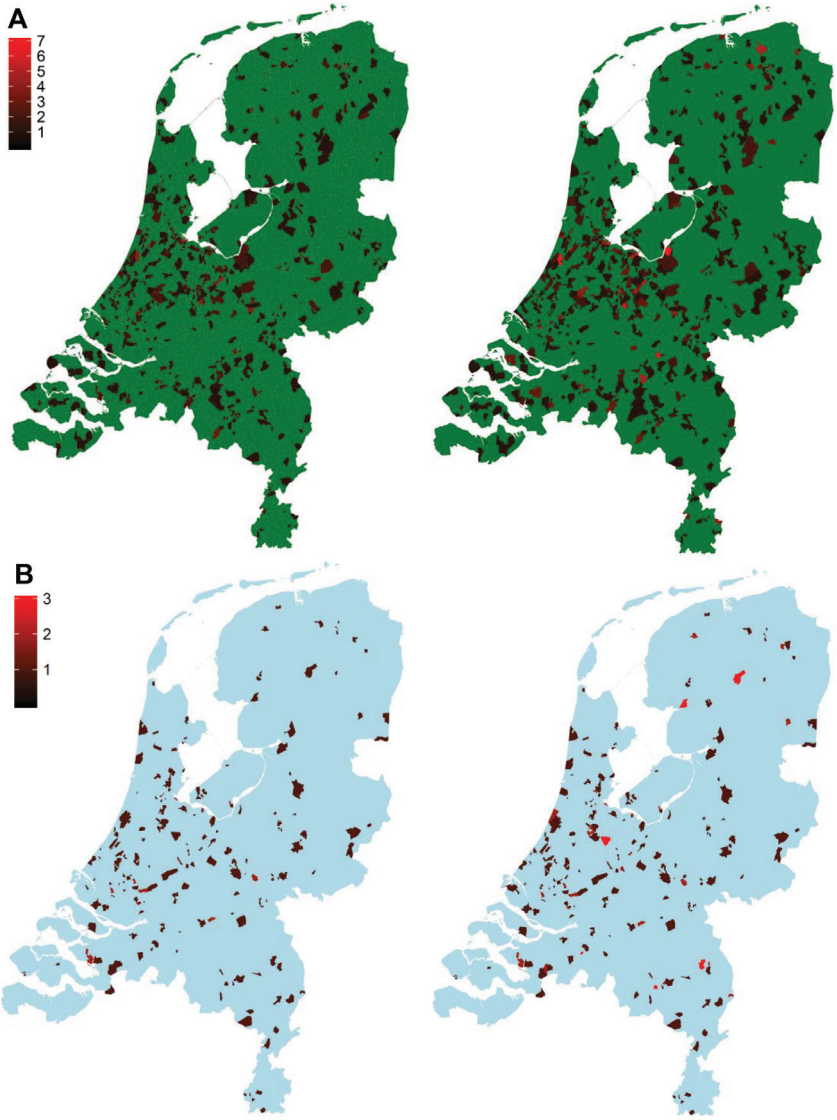


Figure C. Geographical distribution of recruitment. (A) Geographical location of ASeed (left) and the total sample ASeedRec (right) on a 4-digit postal code level. **(B)** Geographical location of BSeed (left) and the total sample BSeedRec (right). Red colored areas indicate a higher density of participants: The Netherlands, 2013-2014 influenza season.

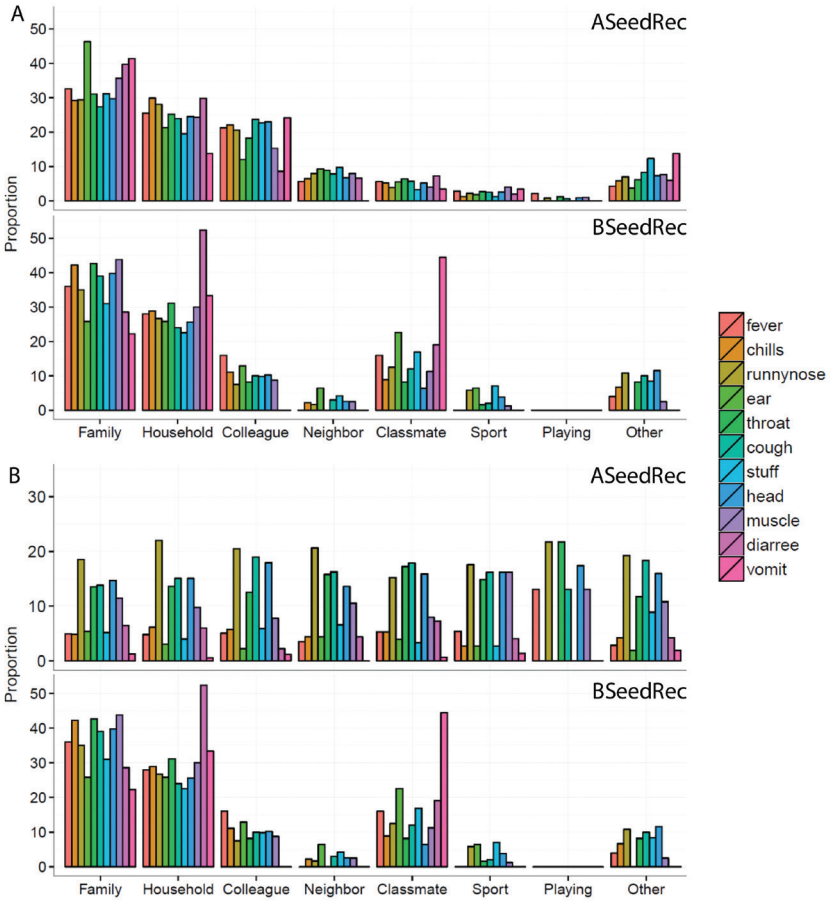


Figure D. Contacts with similar symptoms as symptomatic participants, stratified by symptoms displayed in questionnaire. (**A**) Proportions show type of contact persons for whom symptomatic participants recognized similar symptom(s). (**B**) Proportions show type of symptom recognized (by symptomatic participants) per type of contact person. Contact persons were specified as household members, family members not living in the household, colleagues, neighbors, classmates at day care, school or university, contact persons met during sports, children with whom your child plays regularly, and 'other contact persons': The Netherlands, 2013-2014 influenza season

Table A. List of symptoms as displayed in questionnaire.

No complaints
Fever
Chills
Runny or blocked nose
Earache
Sore throat
Cough
Stuffiness
Headache
Muscle/joint pain
Diarrhea
Vomit
Other, namely

Table B. Geographical spread of recruitment on Dutch province level.

Provinces	Census data* (%)	Panel A (%)	*P*	ASeed (%)	*P*	ASeedRec sample	*P*	Panel B	*P*	BSeed (%)	*P*	BSeedRec sample	*P*
Groningen	581705 (3.5)	514 (4.4)	<.001	36 (4.6)	.116	58 (4.9)	.007	53 (3.1)	.504	8 (3.6)	.854	11 (4.1)	.714
Friesland	646862 (3.9)	391 (3.4)	.005	36 (4.6)	.355	46 (3.9)	.980	57 (3.4)	.338	6 (2.7)	.485	12 (4.4)	.740
Drenthe	489918 (2.9)	344 (3.0)	.864	16 (2.0)	.164	26 (2.2)	.175	46 (2.7)	.687	4 (1.8)	.607	6 (2.2)	.716
Gelderland	2015791 (12.0)	1360 (11.7)	.255	88 (11.1)	.475	131 (11.1)	.381	203 (12.0)	.970	26 (11.7)	.952	31 (11.4)	.844
Oversijssel	1139350 (6.8)	676 (5.8)	<.001	39 (4.9)	.045	56 (4.8)	.007	98 (5.8)	.119	7 (3.1)	.032	7 (2.6)	.003
Flevoland	398441 (2.4)	332 (2.8)	<.001	25 (3.2)	.182	38 (3.2)	.067	55 (3.3)	.021	10 (4.5)	.047	11 (4.1)	.105
North Holland	2724300 (16.2)	2258 (19.4)	<.001	157 (19.8)	.007	234 (19.9)	<.001	312 (18.5)	.013	36 (16.1)	1	41 (15.1)	.681
South Holland	3563935 (21.2)	2450 (21.0)	.562	149 (18.8)	.108	204 (17.3)	.001	408 (24.2)	.004	60 (26.9)	.047	67 (24.7)	.184
Utrecht	1245294 (7.4)	1093 (9.4)	<.001	98 (12.4)	<.001	164 (13.9)	<.001	162 (9.6)	<.001	22 (9.9)	.206	28 (10.3)	.087
Zeeland	381077 (2.3)	326 (2.8)	<.001	29 (3.7)	.012	36 (3.1)	.085	35 (2.1)	.643	5 (2.2)	1	5 (1.8)	.838
North Brabant	2471011 (14.7)	1411 (12.1)	<.001	96 (12.1)	.045	152 (12.9)	.089	187 (11.1)	<.001	25 (11.2)	.166	33 (12.2)	.272
Limburg	1121891 (6.7)	503 (4.3)	<.001	22 (2.8)	<.001	31 (2.6)	<.001	72 (4.3)	<.001	14 (6.3)	.913	19 (7.0)	.926
Total	16779575	11658		791†		1176[b]		1688		223		271	

*Census data from 2013, used as reference population.
†One participant provided an invalid postal code.

Table C. Comparison by sex and age of panels A-B and samples ASeedRec-BSeedRec with the Dutch population as reference population.

	Age	Dutch population (%)	Panel A (%)	*P*	ASeedRec sample (%)	*P*	Panel B (%)	*P*	BSeedRec sample (%)	*P*
F	<15	16.6	2.9	<.001	0.6	<.001	0.1	<.001	0	<.001
	15-29	17.9	5.4	<.001	8.2	<.001	5.0	<.001	4.5	<.001
	30-44	19.5	23.6	<.001	18.8	.629	22.5	.013	17.9	.693
	45-59	21.4	39.8	<.001	40.5	<.001	37.9	<.001	43.6	<.001
	60-74	15.9	25.4	<.001	28.5	<.001	31.8	<.001	32.1	.002
	≥75+	8.7	3.0	<.001	3.4	<.001	2.7	<.001	1.9	<.001
	Total	**50.5**	**57.4**	**<.001**	**66.5**	**<.001**	**64.8**	**<.001**	**57.6**	**.023**
M	<15	17.7	3.6	<.001	1.8	<.001	0	<.001	0	<.001
	15-29	18.7	4.0	<.001	6.3	<.001	3.3	<.001	8.7	<.001
	30-44	19.9	14.9	<.001	13.7	<.001	12.1	<.001	7.0	<.001
	45-59	22.0	34.9	<.001	30.5	<.001	28.9	<.001	30.4	.039
	60-74	15.8	35.7	<.001	41.6	<.001	43.5	<.001	51.3	<.001
	≥75+	5.7	6.8	.001	6.1	.847	12.1	<.001	10.4	.049
	Total	**49.5**	**42.6**	**<.001**	**33.5**	**<.001**	**35.2**	**<.001**	**42.4**	**.023**

Table D. Changes in basic demographic characteristics through recruitment.

	ASeedRec sample)		BSeedRec sample	
	ASeed	ARec	BSeed	BRec
Number of participants	792*	385†	223	48‡
Mean age (median; SD)	54.9 (56.0; 13.1)	49.0 (52.0; 17.1)	57.8 (59.0; 12.1)	52.4 (56.0; 13.9)
Women	536 (67.7%)	247 (64.2%)	126 (56.5%)	30 (62.5%)
Mean household size (median; SD)	2.4 (2.0; 1.7)	2.6 (2.0; 1.7)	2.2 (2.0; 1.0)	3.1 (2.0; 4.2)
Education (Bachelor or higher)	485 (61.2%)	208 (54.0%)	91 (49.8%)	46 (54.2%)
Sent at least one invitation to contact persons	329 (41.5%)	182 (47.3%)	78 (35.0%)	21 (43.8%)
Recruited at least one recruitee§	165 (20.8%)	93 (24.2%)	29 (10.3%)	5 (10.4%)

*Five parents in ASeed and seven parents in ASeedRec completed the questionnaire for their child.
†ASeedRec reached to a maximum of 6 waves of recruitees and for ARec waves 1 to 6 were lumped together.
‡BSeedRec had a maximum of 2 waves, for BRec waves 1 to 2 were lumped together.
§Recruitment of recruitee means that an invited contact person completed the survey.

Table E. Recruitment by invitation options.

	ASeedRec sample	**BSeedRec sample**
Total number of participants	1177	271
Used a recruitment option (%)	510 (43.3%)	99 (36.5%)
Recruited at least one recruitee (%)	258 (24.2%)	34 (12.5%)
Total invitations sent out via recruitment page	1802	340
Invitations sent by indirect email method* (%)	680 (37.7%)	148 (43.5%)
Invitations sent by direct email method (%)	954 (52.9%)	165 (48.5%)
Invitations sent via Facebook (%)	167 (9.3%)	27 (8.0%)
URLs shared† (%)	1 (0.1%)	0 (0%)
Total number of recruitees	385	48
Recruitees invited by indirect email method (%)	85 (22.1%)	11 (22.9%)
Recruitees invited by direct email method (%)	262 (68.1%)	33 (68.8%)
Recruitees invited via Facebook (%)	37 (9.6%)	4 (8.3%)
Recruitees invited via sharing of URL (%)	1 (0.2%)	0 (0%)

* As it is unknown how many emails were forwarded by participants after they received four invitation emails, we assumed here that all four 'indirect invitations' were sent out to recruitees.
† Four unique URLs were displayed on the recruitment page, but sharing these URLs could not be registered by the system. One participant recruited one recruitee without using the three other invitation options and therefore we expect that this participant shared a URL.

Table F. Self-reported symptoms in total samples.

	Definition	ASeedRec sample	BSeedRec sample
Common cold	Crude attack rate	7.5% (88)	8.9% (24)
	% cases stayed at home due to symptoms	40.9% (36)	45.8% (11)
	% cases visited general practitioner	11.4% (10)	16.7% (4)
	Secondary attack rate	29.7% (46 of 243 household members)	36.8% (14 of 62 household members)
	Attack rate according to perception participant	17.6% (207)	19.2% (52)
Influenza-like-illness (ILI)	Crude attack rate*	2.5% (29)	1.1% (3)
	% cases vaccinated against flu	27.6% (8)	66.7% (2)
	% cases stayed at least one day at home due to symptoms	82.6% (24)	66.7% (2)
	% cases visited general practitioner	10.3% (3)	66.7% (2)
	Secondary attack rate	43.2% (16 of 66 household members	0%
	Attack rate according to perception participant	8.8% (103)	8.5% (23)

* Two participants reported ILI symptoms that started more than 3 weeks before participation and were therefore excluded.

Table G. Detection of symptoms in network chains of seeds in BSeedRec sample.

	Seeds*		Recruitees by seeds-with-symptoms		Recruitees by seeds-without-symptoms	
	with symptoms	without symptoms	with symptoms	without symptoms	with symptoms	without symptoms
One or more symptoms	20 (69.0%)	9 (31.0%)	17 (48.6%)	18 (51.4%)	7 (53.8%)	6 (46.2%)
Common cold symptoms	2 (6.9%)	27 (93.1%)	1 (20.0%)	4 (80.0%)	1 (2.3%)	42 (97.7%)
Fever	2 (6.9%)	27 (93.1%)	0 (0%)	5 (100%)	1 (2.3%)	42 (97.7%)
Influenza-like illness (ILI)	0 (0%)	29 (100%)	-	-	0 (0%)	48 (100%)

*Only successful seeds (n=29) were considered for the analysis, so only seeds who invited a recruitee who also completed the survey. Recruitees in waves 1 to 2 were lumped together (n=48).

Chapter 6

Social networking sites as a tool for contact tracing: urge for ethical framework for normative guidance

Mart L. Stein[1,2*], Babette O. Rump[3], Mirjam E.E. Kretzschmar[1,2], Jim E. van Steenbergen[1,4]

[1] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[2] Utrecht Centre for Infection Dynamics, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands

[3] Municipal Health Service GGD Midden-Nederland, Zeist, The Netherlands

[4] Centre for Infectious Diseases, Leiden University Medical Centre, Leiden, The Netherlands

Public Health Ethics; Case Discussion 2014, 7(1): 57-60.

The growing popularity of social networking sites (SNS) and its increasing accessibility due to the advent of new technologies (such as smartphones and tablets, enabling users to be online more frequently) provide public health agencies with new opportunities for health promotion, prevention, disease control and research. However, the use of SNS by public health agencies raises ethical concerns. Mandeville and colleagues addressed a number of these concerns by means of an interesting case of contact tracing through Facebook around a patient with meningococcal septicaemia[1]. Although the described situation is especially complicated due to the patient's incapability to provide informed consent, the unexpected interference of a proactive contact and the use of the patients' public Facebook 'wall', it clearly outlines the ethical issues surrounding the use of social media for contact tracing. Fortunately, most situations are less complex in daily practice of communicable disease control.

SNS are here to stay. Although methods on how to use SNS most effectively in public health are understudied, evidence is growing that strategies including SNS can be more effective than traditional communication methods[2-5]. In the area of communicable disease control, particularly for sexually transmitted infections, there are ample examples where SNS, and Internet in general, proved to be valuable tools for surveillance[6, 7], reaching and involving target populations[8-10], or for learning more about (risk) perceptions and behaviors within these populations[11].

However, the use of SNS for contact tracing, a primary means of control of infectious diseases with low prevalence, is relatively new. Since the effectiveness of contact tracing depends mainly on timeliness and completeness of access to, and dissemination of (correct) information in an individual's social network[12, 13], applying SNS in this context also has promising prospects. During a syphilis outbreak among men having sex with men, the use of communication through the Internet enabled public health officials to educate and inform a large number of at-risk persons, offering opportunities for enhanced control of sexually transmitted infections[14]. Also, the same research showed that meeting sexual partners through the Internet was associated with acquisition of syphilis among gay men, indicating that virtual preventive public health activities may in some situations be warranted. More recently, Ladbury et al. investigated the use of Facebook as a recruitment tool for tracing a measles outbreak after a party of a local youth club[15]. After receiving a low response using traditional means of recruitment, they contacted members directly through the Facebook Event group (especially set up for the party). Consequently, response rates increased significantly, sufficient to conduct outbreak investigation.

Nevertheless, SNS push the boundaries of confidentiality, autonomy and consent. Mandeville et al. express a number of concerns related to the use of SNS specifically for contact tracing. First of all, the authors raise the issue of the so-called viral dissemination of messages posted

in public areas. Even when carried out carefully, the use of SNS has the potential to disseminate information beyond the original target population. Connected to this is the fear that when scattered information is also incorrect it could have potentially adverse consequences, such as unnecessary unrest in the population and also might breach confidentiality. One could argue, however, that this is also the case when using traditional methods. Besides SNS, there are many other communication methods such as mobile communication applications (e.g., instant messaging) that are widely used and also permit the dissemination of (incorrect) information to other individuals outside the target group, and traditionally the same applies for newspapers and local radio broadcasts. It is always difficult for public health professionals to control the spread of personal information in social networks, especially during small local outbreaks, even without the use of SNS. It all comes down to the judgment on whether it is reasonable to reveal information about a patient's identity to contacts at risk for infection (to prevent further spread). Professionals in communicable disease control frequently rely on 'the harm principle' as normative guidance for this decision. The harm principle roughly entails that the state may limit the liberty of some individuals to prevent harm to others[16]. This principle offers basic support in decisions on privacy issues, although when it comes to the question whether contact tracing is effective and in proportion we are in need for more normative guidance. This issue is, however, not specifically related to SNS but complicates the overall field of communicable disease control. Once confidentiality is breached the professional can hardly be held accountable for the further dissemination of personal (or incorrect) information in social networks, neither with traditional nor with modern media.

As discussed by Mandevile et al., the application of SNS for tracing outbreaks could potentially improve the targeting of messages limiting the spread of information to only those who really are at increased risk (e.g., the close contacts). The majority of SNS, including the largest SNS such as Facebook, MySpace and LinkedIn, provide members with the option to restrict the visibility of their profile for others. For example, Facebook has the option to create closed groups, prohibited to be accessed or searched for by persons who are not group members[10]. In addition, SNS often have a private messaging feature similar to webmail[17]. These features allow an active two-way communication strategy tailored to the target audience, likely to be more effective for certain age groups than traditional means, while retaining confidentiality. In addition, avoiding communication via public sections of SNS (such as the Facebook 'wall') could prevent unjustified use of personal information for commercial gains by companies owning or associated with these SNS (at least when privacy policies are respected). Unfortunately, unauthorized access (e.g., by 'hackers') to closed groups or private communication cannot be fully prevented, however, this is an issue for computer mediated communication in general. In the complicated case presented by Mandeville et al., it was indeed a close friend that posted the first message on the patients' open wall, later necessitating the public health agency to correct the wording while the patient was still in coma.

The ethical challenges encountered when using SNS for real time contact tracing, especially with regard to autonomy and (absence of) informed consent, were rightly pointed out by Mandeville et al. These challenges mostly relate to the overall question of how to weigh individual interests and needs against those of the public at large. These questions are not new and have always been important issues in source and contact tracing[18], SNS only give these existing concerns a new sense of urgency. With the increasing use of SNS among the general public and the potentially important contribution to daily practice of communicable disease control, it is not a question of whether, but how SNS should be used in communicable disease control. That is, of course, if the initial decision to undertake any form of action can be justified. This first question has, in our opinion, always been the most challenging part in the usual risk-analysis around individual cases of endemic transmissible diseases.

In conclusion, we agree with Mandeville et al. that public health officials are in need of guidelines for the use of SNS for tracing contacts and identifying outbreaks early. However, in our opinion, this need can be further extended to the complete field of communicable disease control (including research), when it comes to setting professional boundaries for which methods are prohibited, obligatory or permissible. There is a need for an ethical framework that offers normative guidance. Such a framework should help professionals in weighing individual interests against those of the general public. Once the decision to undertake action has been made, it seems obvious that professionals choose an effective strategy, and this may well include the use of SNS. In this context, we would like to rephrase the claim of Mandeville et al. and argue that the use of SNS requires an equally careful assessment of risks and benefits as the use of traditional communication ways.

REFERENCES

1. Mandeville KT, Harris M, Thomas HL, Chow Y, Seng C: Using social networking sites for communicable disease control: innovative contact tracing or breach of confidentiality? Public Health Ethics 2013:1-4.

2. Thackeray R, Neiger BL, Smith AK, Van Wagenen SB: Adoption and use of social media among public health departments. BMC public health 2012, 12:242.

3. Gold J, Pedrana AE, Sacks-Davis R, Hellard ME, Chang S, Howard S, Keogh L, Hocking JS, Stoove MA: A systematic examination of the use of online social networking sites for sexual health promotion. BMC public health 2011, 11:583.

4. Bennett GG, Glasgow RE: The delivery of public health interventions via the Internet: actualizing their potential. Annual review of public health 2009, 30:273-292.

5. Bauermeister JA, Zimmerman MA, Johns MM, Glowacki P, Stoddard S, Volz E: Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (webRDS) strategy. Journal of studies on alcohol and drugs 2012, 73(5):834-838.

6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L: Detecting influenza epidemics using search engine query data. Nature 2009, 457(7232):1012-1014.

7. Signorini A, Segre AM, Polgreen PM: The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PloS one 2011, 6(5):e19467.

8. Bull SS, Lloyd L, Rietmeijer C, McFarlane M: Recruitment and retention of an online sample for an HIV prevention intervention targeting men who have sex with men: the Smart Sex Quest Project. AIDS care 2004, 16(8):931-943.

9. Young SD: Social Media Technologies for HIV Prevention Study Retention Among Minority Men Who Have Sex with Men (MSM). AIDS and behavior 2013.

10. Young SD, Cumberland WG, Lee SJ, Jaganath D, Szekeres G, Coates T: Social Networking Technologies as an Emerging Tool for HIV Prevention: A Cluster Randomized Trial. Annals of internal medicine 2013, 159(5):318-324.

11. Love B, Himelboim I, Holton A, Stewart K: Twitter as a source of vaccination information: content drivers and what they are saying. American journal of infection control 2013, 41(6):568-570.

12. Klinkenberg D, Fraser C, Heesterbeek H: The effectiveness of contact tracing in emerging epidemics. PloS one 2006, 1:e12.

13. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, Abu-Raddad LJ, Ho LM, Thach TQ, Chau P et al: Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 2003, 361(9371):1761-1766.

14. Klausner JD, Wolf W, Fischer-Ponce L, Zolt I, Katz MH: Tracing a syphilis outbreak through cyberspace. JAMA : the journal of the American Medical Association 2000, 284(4):447-449.

15. Ladbury G, Ostendorf S, Waegemaekers T, Hahne S: "Liking" Social Networking Sites – Use of Facebook as a Recruitment Tool in an Outbreak Investigation, The Netherlands, 2012. Epidemiology: Open Access 2013, 3(123).

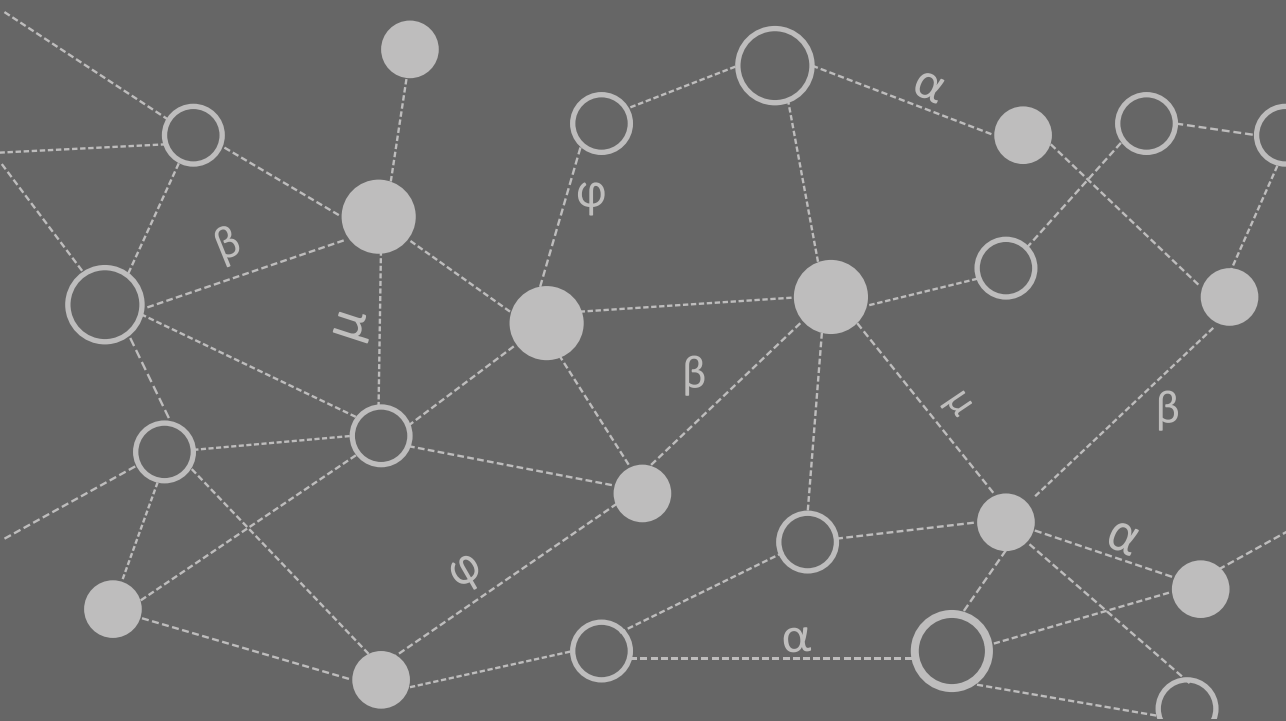
16. Krom A: The harm principle as a mid-level principle?: three problems from the context of infectious disease control. Bioethics 2011, 25(8):437-444.

17. Boyd DM, Ellison NB: Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication 2008, 13.

18. Thomas JC, Sage M, Dillenberg J, Guillory VJ: A code of ethics for public health. American journal of public health 2002, 92(7):1057-1059.

Part III

Factors driving online peer recruitment



Chapter 7

Drivers of respondent-driven detection

Mart L. Stein[1,2], Vincent Buskens[3], Peter G.M. van der Heijden[3,4], Jim E. van Steenbergen[2], Mirjam E.E. Kretzschmar[1,2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[2] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[3] Faculty of Social and Behavioural Sciences, University Utrecht, Utrecht, The Netherlands

[4] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom

Submitted

ABSTRACT

Respondent-driven detection is a chain recruitment method that can be used for sampling of contact persons of infected persons to enhance case finding. It starts with initial individuals, so-called seeds, who are invited for participation. Afterwards, seeds receive a fixed number of invitation coupons for inviting individuals with whom they had contact during a specific time period. Recruitees are then asked to do the same, resulting in successive waves of contact persons who are connected in one recruitment tree. However, often the majority of participants fail to invite others, or invitees do not accept an invitation, and in this situation recruitment stops after a few waves. A simulation model can help to analyse the influence of various factors on peer recruitment and to understand under which circumstances sustainable recruitment is possible. We implemented a stochastic simulation model, where parameters were suggested by empirical data, to determine the thresholds for obtaining large recruitment trees and the number of waves needed to reach a steady state in the sample composition for individual characteristics. Our main finding is that a situation where participants send out any number of coupons between one and the maximum number is more effective in reaching large recruitment trees, compared to a situation where the majority of participants does not send out any invitations and a smaller group sends out the maximum number of invitations. We also examined the relationship between mean and variance of the number of invitations sent out by participants and the probability of obtaining a large recruitment tree. The presented model is a helpful tool that can assist public health professionals in preparing research and contact tracing using respondent-driven detection, in particular, by providing input on the required mean number of successfully sent invitations to reach large recruitment trees, a certain sample composition or certain number of waves.

INTRODUCTION

Many infectious diseases are transmitted via close or intimate contact between individuals. By sampling contact persons of infected persons in a network, it is possible to study the possible transmission routes of these pathogens within networks and to detect hitherto hidden cases. Such information is important for an effective control of disease outbreaks. Respondent-driven detection (RDD), a method derived from snowball sampling, is a chain recruitment method that allows for sampling of contact persons of participants[1, 2]. With RDD, initial individuals (or 'seeds') are invited for participation, which includes filling in a questionnaire. At the end of the questionnaire, seeds receive a fixed number of invitation coupons (usually four, according to standardised methodology of respondent-driven sampling (RDS)[3]), and are asked to invite a number of contact persons whom they have met during a specific time period. Recruitees (i.e., invitees who participate) are then asked to do the same, resulting in successive 'waves' of contact persons. Unlike with snowball sampling, each coupon contains a personal code with which peer recruitment is followed. A set of participants connected via recruitment links to one seed is referred to as a 'recruitment tree'.

With a respondent-driven survey, the composition of the sample consisting of all participants who completed the survey is a function of the probabilities with which individuals with certain characteristics are recruited. Thus, individuals with a high probability of being recruited will likely be overrepresented in the sample. For RDS, statistical techniques are available for estimating population characteristics from the respondent-driven sample[4-6]. Most of these techniques are based on the assumption of a first-order Markov chain process, i.e., correlations found between recruiters and recruitees are only dependent on the direct recruiter[7]. Given that seeds are non-randomly chosen, and that people tend to have contact with similar others[8], the characteristics of the participants in the first few recruitment waves tend to be similar to those of the seeds[9]. If peer recruitment proceeds through a sufficiently large number of waves, recruitees with different characteristics enter the sample. With increasing number of waves, the composition of the sample will converge to a stable distribution and become independent of the (often) non-randomly selected seeds[3, 10, 11]. Although the primary objective of RDD is to detect cases, rather than to estimate population proportions from samples, we were interested in how fast equilibrium is reached under different mixing conditions, e.g., random mixing versus mixing with a preference.

Previously, we conducted surveys to investigate the feasibility of online RDD to study contact patterns that are relevant for the transmission of respiratory pathogens and to exploit the use of the network of cases to detect other cases[1, 2]. However, in those studies, the numbers of recruitees via online peer recruitment were low. The majority of participants did not fully comply or did not comply at all with sending invitations, and a limited number of invitees

actually participated. Therefore, peer recruitment stopped after a few waves. These issues are common with respondent-driven studies[12]. To improve case finding with RDD, or obtain a required number of recruitees or a required sample composition, it is key to find ways to increase rates of peer recruitment, and to understand under which circumstances sustainable recruitment is possible.

In the literature, there are many examples of models used to study recruitment behaviour with RDS or snowball sampling (see e.g., [13-18]). Recently, Malmros et al. (2014) used a configuration model, without empirical data, to analyse the influence of the number of coupons and probability of coupon transfer between a recruiter and recruitee on the recruitment process[13]. To our knowledge, there is one simulator of the recruitment process available in the public domain; this model allows for only a limited number of factors influencing recruitment, and only reports aggregated numbers of recruitees[19].

Here, we used a simulation model that included more parameters to increase our understanding of the dynamics of a respondent driven recruitment process. Simulating the recruitment process enabled us to analyse the influence of various factors on peer recruitment, the final sample size and sample composition. We used empirical data to suggest parameter choices in the simulation model. Data were taken from a sample of recruitment trees collected with RDD in the Netherlands during the winter season of 2013-2014[1, 2]. We aimed to determine which factors are important for the success of recruitment, to identify thresholds for reaching sustainable recruitment in recruitment trees, and to determine how fast a steady state is reached in the sample composition for individual characteristics. Successful recruitment can be defined at two levels. At the level of a recruiter, successful recruitment means that an invitee also completes the questionnaire and becomes a recruitee. At the level of a recruitment tree, successful recruitment indicates that a tree continues to grow due to continued peer recruitment. The results of our analysis can be used to inform future online RDD surveys, e.g., for determining the minimum mean number of successfully sent invitations required to reach a specified number of recruitees.

METHODS

A stochastic simulation model

We considered respondent-driven recruitment in a heterogeneous population. Individuals (i.e., recruiters and recruitees) were characterized by three categorical variables, namely sex, age groups, and education level. Starting from a seed, the recruitment process was modelled as a multi-type discrete time branching process. Recruitees of a recruiter in wave W depended on that recruiter's characteristics, on the number of invitations, and the number of invitees who

accepted the invitation, for all *W*. Each recruiter could send a maximum number of invitations *c*.

We assumed that the number of invitations per recruiter had a beta-binomial distribution, with shape parameters α and β depending on the characteristics of the recruiter. The reason to choose a beta-binomial distribution was that it can reproduce bimodal distributions of numbers of invitations as observed in the data, with a peak at zero and another peak at the maximum value *c*. We then assumed that given a number of invitations, the number of accepted invitations had a binomial distribution where the probability of acceptance *p* depended on the characteristics of the recruiter. The characteristics of recruitees were dependent on the characteristics of the recruiter to reflect correlations between recruiter – recruitee pairs in their characteristics, but assuming independence of the three characteristics. We used the mean number of invitations sent out ($\bar{x}$) and the proportion accepted invitations (*p*) from our sample to calculate the mean number of successfully sent out invitations ($\bar{x} * p$).

Simulation model

Each model run started with one seed in wave 0. For each wave *W*, the model identified the characteristics of all potential recruiters. Then, for each recruiter, the number of invitations sent out and the number of invitations accepted by invitees was determined based on random draws from probability distributions. New recruitees then formed wave *W*+1. Each recruitee in the model received a unique identifier (a numerical string) that linked it to its recruiter. A model run stopped when no invitations were sent out, when no invitations were accepted, or when a maximum total number *N* of recruitees was reached. In our simulations, the maximum number of recruitees was set to *N*=1000. If a recruitment tree became larger than 1000 recruitees, the tree showed ongoing recruitment and continued to grow if the model run was not stopped manually. The final recruitment tree consisted of all recruitees recruited during the simulation run. The model was implemented in R (R Foundation for Statistical Computing, Vienna, Austria) version 3.2.2. For model formulation and details, we refer the reader to supplementary text S1.

Outcome measures

To quantify the success of recruitment, we kept track of the number of recruitees in each recruitment tree starting with one seed (i.e., the size of recruitment trees). If the size of a recruitment tree was larger or equal than *N* = 1000, we labelled it as 'large'. We then computed the proportion of runs with large recruitment trees for each set of parameters. The probability of obtaining large recruitment trees depends on the specific parameter combination. If participants successfully invite on average less than 1 new recruitee, only small recruitment trees occur, while above this threshold small and large recruitment trees occur. The proportion of large recruitment trees increases with increasing mean number of successfully sent invitations. We recorded for which parameter combinations at least 5% of runs resulted in a

large recruitment trees. We also recorded the maximum wave reached by each tree to quantify the length of recruitment trees.

To investigate the sample composition at equilibrium, we calculated for each simulation run the composition of each wave with respect to sex, age group and educational level. We used the standard criterion from the RDS literature to define equilibrium[3, 9]. This criterion states that sample proportions are viewed as stable when the difference between proportions in subsequent waves is less than two percent. To quantify the convergence of the composition in our simulations, we investigated at which wave equilibrium according to the above criterion was reached. Proportions were calculated per wave and averaged over simulation runs.

By way of illustration of how the model can provide guidance for future surveys, we also included influenza vaccine beliefs in the model. Beliefs about a vaccine can strongly affect individual vaccination decisions. In particular, negative vaccine beliefs can lead to low vaccination rates, which in turn can lead to higher likelihood of a disease outbreak[20].

Model parameters

Study population

Model parameters were suggested by respondent data collected during a RDD study performed in the Netherlands during the winter season of 2013-2014[1]. Participants were enrolled via a large web based participatory surveillance panel. After filling out a questionnaire, each participant received four unique electronic coupons to invite contact persons whom they had met in the previous two weeks. A total of 1015 volunteers entered the RDD survey as seeds, and 433 recruitees were successfully recruited. Recruitment reached up to 6 waves of recruitees. For each participant, sociodemographic variables were recorded, including sex (females and males), age (three age groups), education level (two categories) and vaccine belief (two categories: 'positive' and 'negative or undecided'). Participants were allowed to fill in the survey for one of their children. Overall, participants had a mean age of 53.7 years (standard deviation (SD): 14.5 years; range: 3 – 97 years), 64.8% was female, 57.3% had an academic education and 53.5% had a negative vaccine belief (see Table 1).

Recruitment behaviour

Of the 1448 participants (i.e., seeds and recruitees) who completed the survey, 609 (42.1%) sent out invitations. The overall mean number of invitations sent out was 1.36 (s^2: 3.21). The distribution of the number of invitations sent out was bimodal, with a peak at zero and another peak at four. We fitted a beta-binomial distribution to the observed frequency distribution of invitations sent out by participants, using the R package "VGAM" (function "betabinomialff"). The mean number of invitations sent out varied for different subgroups of participants depending on sex, age and educational level between 0.64 (s^2: 2.02) and 1.67 (s^2: 3.48) (see

supplementary Tables S1). Furthermore, a difference in mean numbers of sent invitations was observed between seeds and recruitees in waves 1 to 6 (see Figure S1). Seeds sent on average slightly less invitations compared to recruitees in waves 1 to 6, 1.33 (s^2: 3.24) and 1.47 (s^2: 3.23) respectively (see Table 1 and supplementary Table S2). Female seeds, with an academic education in the age group 60 years and older, were most active with sending invitations to others.

We based the probability of an invitation being accepted in the model on the number of recruitees divided by total number of invitations sent out by recruiters with specific characteristics in the data set. The acceptance probability therefore depended on the characteristics of the recruiter, and not on the characteristics of the recruitee. Overall, 19.2% (range: 10.0% - 27.2% for different recruiter characteristics) of all invitations sent were accepted by recruitees. The mean proportion of acceptance only slightly differed between seeds (18.9%) and recruitees (20.1%). See supplementary Table S2 for an overview of the fitted beta binomial distributions and proportions of invitations accepted stratified by recruiter characteristics, and by seeds and recruitees in waves 1 to 6.

Table 1. Participants' characteristics and recruitment behaviour as observed in data set.

Variables	Seeds in wave 0*	Recruitees in waves 1 to 6*	Overall (seeds + recruitees)
Number of female participants (%)	662 (65.2%)	277 (34.8%)	939 (64.8%)
Age in years of participants (SD; range)	55.5 (13.0; 4–97)	49.4 (16.8; 3–82)	53.7 (14.5; 3–97)
Number of participants in age group 0-39, a1 (%)	127 (12.5%)	121 (27.9%)	248 (17.1%)
Number of participants in age group 40-59, a2 (%)	465 (45.8%)	173 (40.0%)	638 (44.1%)
Number of participants in age group 60+, a3 (%)	423 (41.7%)	139 (32.1%)	562 (38.8%)
Number of participants with an academic education, B	596 (58.7%)	234 (54.0%)	830 (57.3%)
Number of participants with a negative vaccine belief	482 (47.5%)	293 (67.7%)	775 (53.5%)
Mean number of invitations sent out, $\bar{x}$ (s^2)	1.33 (3.24)	1.47 (3.23)	1.36 (3.31)
Mean proportion invitation was accepted, *p* (%, range)	18.9 (0–26.9)	20.1 (7.7–31.6)	19.2 (10.0–27.2)

* A stratification by recruiter characteristics can be found in supplementary Table S1.

Mixing parameters

The characteristics of the recruitees were determined based on random draws from probability distributions for sex, age and educational level, where the probability distributions depended on characteristics of the recruiter and were assumed to be independent from each other. Overall, the data showed that participants tended to recruit recruitees with similar characteristics. This is reflected by correlation coefficients for recruiter-recruitee pairs by characteristic[21]. A positive correlation (>0.10) reflects assortative mixing (i.e., inviting recruitees with the same characteristics), while a negative correlation (< -0.10) reflects disassortative mixing

(i.e., inviting recruitees with other or the opposite characteristics). A correlation between -0.10 and 0.10 is usually interpreted as random mixing, which indicates that participants do not have a certain tendency to recruit recruitees with specific characteristics. Participants recruited mainly recruitees of a similar age group (r_{rank}: 0.37 (0.28-0.44)) and with a similar education (r_{φ}: 0.31 (0.22-0.39). We based the probability distributions for sex, age and educational level on the proportions observed in the data set, e.g., the proportion of female participants who invited female recruitees. As the mixing behaviour varied over waves (e.g., recruiters in wave 1 invited more females and recruitees with an academic education, as compared to seeds in wave 0), we calculated both overall mixing proportions and mixing proportions stratified by waves (see Table 2).

We assumed that a vaccine belief is determined by individual characteristics and does not influence the recruitment process. We used a logistic regression model to estimate the probabilities of having a positive or negative belief about the influenza vaccine depending on age, sex, and education. Vaccine beliefs of individuals were determined based on random draws from the estimated probability distributions (see supplementary text S1). The observed time between sending of invitations by a recruiter and the moment of acceptance by their recruitees was not related to the characteristics of the recruiter, and therefore not included in the scenario analyses.

Scenarios

In total, we defined and investigated 18 scenarios (see Table 3), which were distinct in their parameter choices. For each parameter combination, starting with one seed, we performed 100 simulation runs. In scenario S1, we explored the relation between mean (μ) and variance (σ^2) of the beta-binomial distribution for the number of invitations, and the probability of obtaining large recruitment trees. We randomly drew 9000 combinations of μ and σ^2, ran the simulations and recorded the proportion of large trees per parameter combination. We assumed a probability of acceptance of invitations of 1.

In scenarios S2 to S4, we ran 1000 simulations with 1015 seeds each, with characteristics (sex, age group and education level) and recruitment patterns as observed in the data. One set of seeds together with their recruitment trees was interpreted as one simulated data set, which could be compared to the observed data set. In scenario S2, the mean number of sent invitations and probability of acceptance were stratified by recruiter characteristics, with overall mixing proportions as observed in the data. In scenario S3, probabilities of sending and acceptation were stratified by recruiter characteristics, and differed between seeds (wave 0) and recruitees in waves 1 to 6. In addition, in scenario S4, we used mixing proportions stratified by waves (Table 2).

Table 2. Heterogeneity in mixing patterns over waves as observed in data set.

Mixing	Waves 0-1 (n: 295)	correlation	Waves 1-2 (n: 86)	correlation	Waves 2 to 6 (n: 52)	correlation	Mean overall proportions (n: 433)	Overall correlations (n: 433)
Female – Female	0.61	0.02 [-0.09–0.14]	0.73	0.16 [-0.05–0.36]	0.83	0.23 [-0.05–0.47]	0.66	0.08 [-0.01–0.17]
Female - Male	0.39		0.27		0.17		0.33	
Male – Male	0.41		0.43		0.50		0.42	
Male – Female	0.59		0.57		0.50		0.58	
Age 0-39 (a1) – Age 0-39 (a1)	0.56	0.23 [0.12–0.34]	0.57	0.52 [0.34–0.66]	0.86	0.65 [0.46–0.78]	0.60	0.33 [0.25–0.42]
Age 0-39 (a1) – Age 40-59 (a2)	0.12		0.29		0.14		0.18	
Age 0-39 (a1) – Age 60+ (a3)	0.31		.014		0.00		0.22	
Age 40-59 (a2) – Age 0-39 (a1)	0.28		0.19		0.20		0.26	
Age 40-59 (a2) – Age 40-59 (a2)	0.51		0.67		0.67		0.56	
Age 40-59 (a2) – Age 60+ (a3)	0.21		0.14		0.13		0.19	
Age 60+ (a3) – Age 0-39 (a1)	0.23		0.10		0.07		0.19	
Age 60+ (a3) – Age 40-59 (a2)	0.29		0.21		0.20		0.26	
Age 60+ (a3) – Age 60+ (a3)	0.49		0.69		0.73		0.55	
Lower than academic education (A) - Lower than academic education (A)	0.65	0.29 [0.18–0.39]	0.67	0.41 [0.22–0.57]	0.68	0.21 [-0.06–0.46]	0.65	0.31 [0.22–0.39]
Lower than academic education (A) - Academic education (B)	0.35		0.33		0.31		0.35	
Academic education (B) - Academic education (B)	0.65		0.76		0.53		0.66	
Academic education (B) - Lower than academic education (A)	0.35		0.24		0.47		0.34	

Table 3. Scenarios.

			Beta-binomial distribution			accept	Mixing by sex, proportion		Mixing by age groups, proportion						Mixing by education	
Scenarios		Seeds	c	μ	σ^2	p	F-F	M-M	a1-a1	a1-a2	a2-a1	a2-a2	a3-a1	a3-a2	A-A	B-B
Randomly draw 9000 combinations of μ and σ^2	S1	1	4	[0.1-3.9]	[0.1-3.9]	1	Successful sending was not dependent on characteristics recruiter, thus no influence of mixing behaviour on number of recruitees in recruitment trees.									
μ and p depend on recruiter's characteristics	S2	1015 from data	4	Table S1	Table S1	Table S1	0.66	0.42	0.60	0.18	0.26	0.56	0.19	0.26	0.65	0.66
μ and p depend on recruiter's characteristics + whether seed or recruitee	S3	1015 from data	4	Table S2	Table S2	Table S2	0.66	0.42	0.66	0.18	0.26	0.56	0.19	0.26	0.65	0.66
μ and p depend on recruiter's characteristics + whether seed or recruitee, assuming heterogeneity in mixing patterns over waves	S4	1015 from data	4	Table S2	Table S2	Table S2	Assumed heterogeneity in mixing behaviour for different waves, as observed in the data. Values taken from Table 2.									
Assuming a beta-binomial distribution with **highest** possible σ^2 (as observed in data)	S5	1	4	0.1, 0.2,..., 3.9	0.1-3.4*	0.19	Successful sending was not dependent on characteristics recruiter, thus no influence of mixing behaviour on number of recruitees and wave reached in recruitment trees.									
	S6	1	4	0.1, 0.2,..., 3.9	0.1-3.4*	0.40										
	S7	1	4	0.1, 0.2,..., 3.9	0.1-3.4*	0.60										
	S8	1	4	0.1, 0.2,..., 3.9	0.1-3.4*	0.80										
	S9	1	4	0.1, 0.2,..., 3.9	0.1-3.4*	1										

**High σ^2 chosen, similar to observed in data. See supplementary Figure S2A for the different shapes of the beta-binomial distribution.

Table 3. Scenarios (continued)

			Beta-binomial distribution			accept	Mixing by sex, proportion		Mixing by age groups, proportion						Mixing by education	
Scenarios		Seeds	c	μ	σ^2	p	F-F	M-M	a1-a1	a1-a2	a2-a1	a2-a2	a3-a1	a3-a2	A-A	B-B
Assuming a beta-binomial distribution with **lowest** possible σ^2	S10	1	4	0.1, 0.2,..., 3.9	0.1-1.1[€]	0.19										
	S11	1	4	0.1, 0.2,..., 3.9	0.1-1.1[€]	0.40	Successful sending was not dependent on characteristics recruiter, thus no influence of mixing behaviour on number of recruitees and wave reached in recruitment trees.									
	S12	1	4	0.1, 0.2,..., 3.9	0.1-1.1[€]	0.60										
	S13	1	4	0.1, 0.2,..., 3.9	0.1-1.1[€]	0.80										
	S14	1	4	0.1, 0.2,..., 3.9	0.1-1.1[€]	1										
Differences in mixing behaviour	S15	100 random seeds from data	4	Table S3 (+0.6)	Table S3 (-0.6)	1	Assumed heterogeneity in mixing behaviour for different waves, as observed in the data. Values taken from Table 2.									
	S16	100 random seeds from data	4	Table S3 (+0.6)	Table S3 (-0.6)	1	0.50	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.50	0.50
	S17	100 random seeds from data	4	Table S3 (+0.6)	Table S3 (-0.6)	1	1	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.50	1
Active seeds	S18	100 highly active seeds	4	Table S3 (+0.6)	Table S3 (-0.6)	1	1	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.50	1

[€] Lowest possible σ^2 with respect to chosen μ. See supplementary Figure S2B for the different shapes of the beta-binomial distribution.

Next, we defined 10 scenarios (S5 to S14) to study conditions for obtaining large recruitment trees, and the influence of increased recruitment on the total number of recruitees and maximum wave reached by recruitment. Per scenario, we varied the mean number of invitations sent out by individuals (μ) between 0.1 and 3.9, and we varied the probability of acceptance of an invitation between 0.19 (as observed in the data), 0.40, 0.60, 0.80 and 1. In total this led to 10 scenarios with different recruitment probabilities, where recruitment probability was not stratified by characteristics of the recruiter. In scenarios S5 to S9, we defined σ^2 based on the sample variance of the number of sent invitations observed in the data; for higher values of μ, the observed sample variance was not compatible with the beta-binomial distribution; we then used the maximum σ^2 possible. We used a high σ^2 to maintain in each set of simulations a bimodal distribution of the number of invitations sent out, as observed in the data. To investigate the influence of other than bimodal distributions of number of invitations sent out, in scenarios S10 to S14 the lowest possible σ^2 compatible with a given value of μ was used. The shapes of the beta-binomial distributions used in scenarios S5-S14 are shown in Figure S2.

To investigate how differences between recruiters and asymmetric recruitment can be exploited to increase the number of recruitees and waves, and how this affects the sample composition, we performed simulations for an additional four scenarios (S15 to S18). For each scenario, we assumed that each type of recruiter sent out on average $\bar{x}$ + 0.6 invitations. This value was chosen to ensure that the mean number of successfully sent invitations was above 1 for all types of recruiters (see Table S3). The variance of number of invitations was set to s^2 - 0.6, to ensure compatibility with the beta-binomial distribution. The probability of acceptance was set to 1. In scenario S15, we used mixing proportions stratified by waves as observed in the data. We then compared random mixing by characteristics (S16) with a situation where females recruit only females and academics only academics (S17). In scenarios S15 to S17, we used 100 seeds, whose characteristics were matched to 100 randomly drawn seeds from the observed data set. Again, we interpreted one set of 100 seeds together with simulated recruitment trees as one simulated data set. In the last scenario (S18), we used the mixing proportions described in scenario S17, but started with 100 seeds, whose characteristics matched seeds from the data set who had the highest observed mean numbers of sent invitations.

RESULTS

Probability of a large recruitment tree

Figure 1 shows the proportion of large trees for scenario S1. The domain in the parameter region, for which results are plotted, is determined by the range of possible combinations of μ and σ^2 of the beta-binomial distribution. In general, the probability of getting a large recruitment tree increases with increasing μ. There is a parameter region, however, where increase of μ leads to a decrease in the probability of large trees. This is the case for μ around

3 and σ^2 equal to 2. This suggests that, for some parameter combinations it is not necessary to achieve the highest possible number of invitations for maximal success of peer recruitment for a fixed variance. In this parameter region, increasing μ may result in a lower probability of getting a large recruitment tree. For a fixed variance, increasing the mean is only possible if more weight of the distribution is moved to the extremes, i.e., the probability of sending either zero or c invitations increases while probabilities of the values in between decreases. The extinction probability in such a situation is larger than when probability weights are distributed more evenly over the possible values of number of sent out invitations (see Figure S2 for different shapes of the beta-binomial distribution).

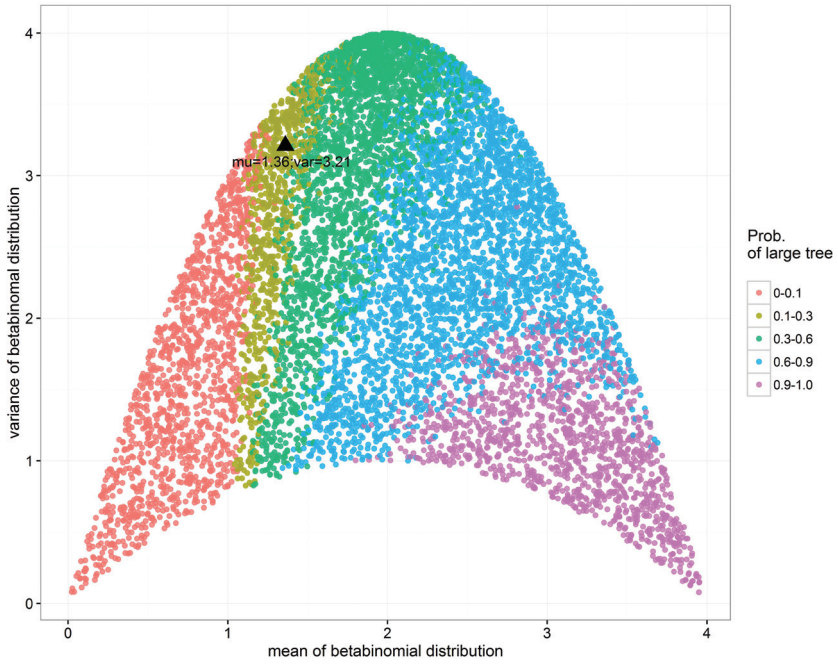


Figure 1. Relation between numbers of invitations and proportion of large recruitment trees. For 9000 randomly chosen combinations or μ and σ^2 for the beta-binomial distribution for number of invitations we ran 100 simulations (scenario S1). We then calculated the proportion of simulations runs that resulted in a large recruitment tree (N≥1000 recruitees) under the assumption of an acceptance probability of 1. The black triangle indicates the $\overline{x}$ and s^2 of numbers of invitations sent out by participants in the data set collected in the Netherlands.

Simulation runs based on empirical data

In scenarios S2 to S4 we compared the simulations with the observed data, to explore how well the model with the chosen parameter values could reproduce observed data. Our model slightly underestimated the number of recruitment trees with zero recruitees, and overestimated the number of trees with one and two recruitees, and one and two waves (see Figures 2A and 2B). By taking into account that seeds sent out a lower mean number of invitations than recruitees

in consecutive waves (scenario S4), the model estimations slightly improved as compared to only taking recruiters' characteristics into account (scenario S3). This suggests that seeds recruited recruitees who were more motivated than themselves to participate and to invite others. The remaining discrepancy between the simulations and the observed data, is most likely due to a practical issue during the data collection. The online software system used for sending invitations was only able to register the number of invitation letters that participants requested for further use via the survey web site. However, no information is available on whether participants actually used those requested invitation letters. If participants did not use all the invitation letters they requested, the mean number of invitations actually sent out during the data collection is lower than the mean number of requested invitation letters. The actual mean number of invitations sent out is therefore lower than estimated from the data. Using a lower value of $\bar{x}$ in the simulations would reduce the number of seeds in the model who successfully recruit recruitees and would thus improve the agreement of model results with data.

Simulated proportions of characteristics agreed with observed data in the first four waves with respect to all three variables (see Figures 2C-2E). For the remaining waves, the numbers were too small to obtain good estimates. For the first few waves, and for females in waves 0 to 4, the simulated mean proportions were in better agreement with the observed data when using mixing proportions stratified by waves (scenario S4), instead of overall mixing proportions (scenarios S2 and S3). This illustrates the influence of heterogeneity in mixing behaviour over different waves on the sample composition.

Increasing successfully sent invitations

In general, an increased mean number of invitations in combination with an increased probability of acceptance led to a higher probability of obtaining large recruitment trees (scenarios S5 to S9). The proportion of 0.19 that accepted an invitation as observed in the data set (used in scenario S5), was too low to reach a value of successfully sent invitations above 1, even when all recruiters sent out the maximum number of 4 invitations. For probabilities of acceptance between 0.40 and 1, the probability of a large tree was below 5% when the mean number of successfully sent invitations was approximately 1, but above for larger mean values. When assuming a bimodal distribution for the numbers of invitations sent out (as assumed in scenarios S5 to S9), with peaks at zero and four, the probability of acceptance seems to be of less importance for obtaining large recruitment trees, especially for values for μ of 2.4 and higher with acceptance probabilities of 0.8 and 1, as the (solid) lines overlap for larger p (see Figure 3).

We further investigated a situation where 60% of all invitations sent out are accepted (scenario S7), to explore the influence of increased successful recruitment on the distribution of the

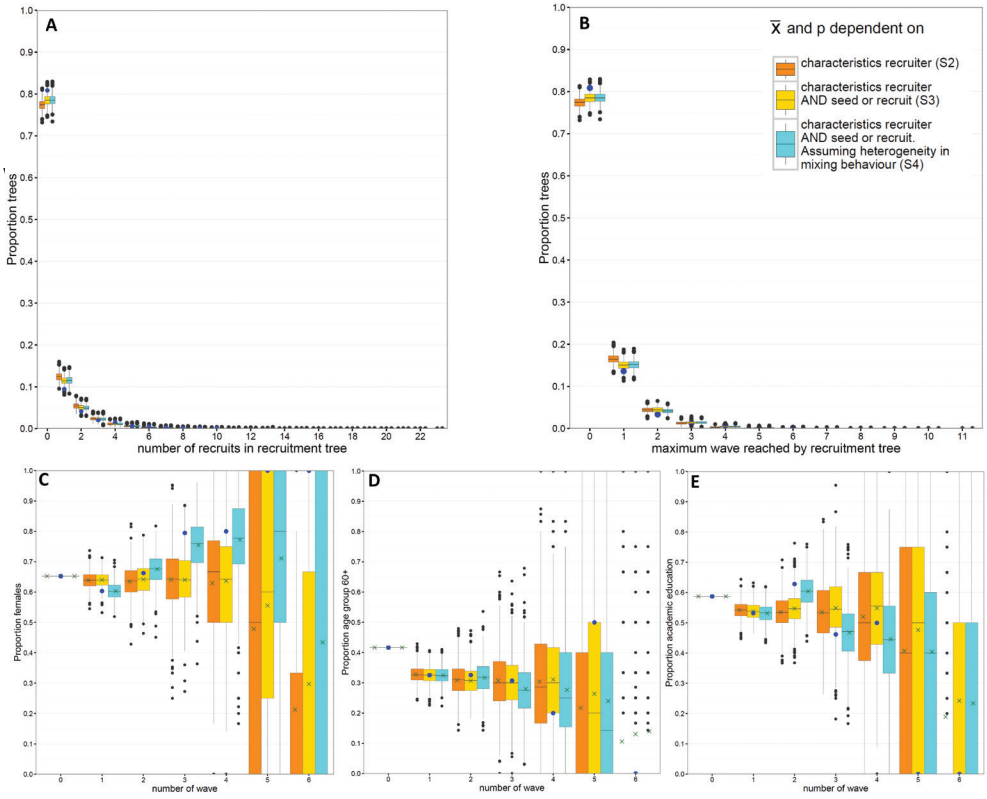


Figure 2. Simulation runs based on empirical data. Starting with a set of 1015 seeds with characteristics as in the empirical data, we performed 1000 simulation runs. One set of seeds together with their recruitment trees was interpreted as one simulated data set, which could be compared to the observed data set. The boxplots show the variability among simulated data sets, each boxplot consists of 1000 simulated points. The legend refers to scenarios S2 to S4 described in Table 3. The blue dots indicate the observed data. (**A**) Number of recruitees in recruitment trees in one simulated data set; (**B**) wave reached by recruitment trees in one simulated data set. Plots C to E show, for each wave, the variability in composition of each simulated data set, with (**C**) proportion of females, (**D**) age group 60+ and (**E**) having an academic education. A green cross indicates, for each wave, the mean proportion over all simulated data sets.

number of waves reached in recruitment trees. Around a mean number of successfully sent invitations of 1, the number of waves in a tree at the end of the simulation ranged from 0 to 35. Increasing the number of successfully sent invitations resulted in lower numbers of waves reached by recruitment trees (Figure S3). This can be explained by the termination of recruitment when a maximum number of recruitees was reached. If more recruitees enter the sample in each wave, the recruitment tree becomes wider and reaches this maximum in a smaller number of waves.

Remarkably, when we assumed that more participants send out 2 or 3 invitations than 0 or 4, i.e., if the beta-binomial distribution was unimodal (scenarios S10 to S14), the mean number of invitations can be lower to reach the same probability of a large recruitment tree (Figure 3, and supplementary Figure S2B). This holds for simulations where the probability of acceptance of an invitation was at least 0.4. Figure 3 also shows, in situations with a unimodal beta-binomial

distribution, that the probability of obtaining a large tree can approach 1 for situations with a high mean μ and a probability of invitation acceptance of at least 0.8. Furthermore, the probability of acceptance seems to influence the probability of obtaining a large tree for all values of μ (i.e., the dashed lines do not overlap in Figure 3). Our simulations suggest that for a given mean number of invitations it is more effective to motivate participants to send out at least one invitation, compared to a situation where most participants do nothing and a smaller group sends out *c* invitations.

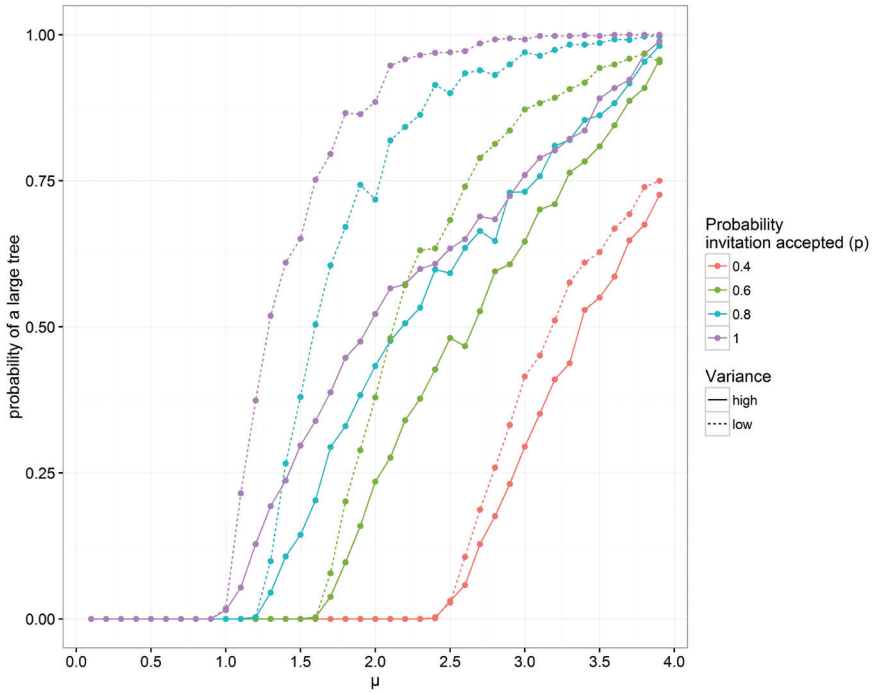


Figure 3. Increased number of invitations sent and probability of acceptance. Solid lines show probability of a large tree for different values of μ and a high σ^2, as observed in the data (scenarios S5 to S9). The dashed lines show the same, but now assuming the lowest possible σ^2 for each μ (scenarios S10 to S14).

Exploiting asymmetric differences in recruitment

In scenarios S15 to S18, we increased the mean number of successfully sent invitations above 1, to explore how differences in who mixes with whom can be used to reach a higher proportion of large recruitment trees, and how this affects the sample composition. Increasing the probability that females only recruit females and academics only recruit academics (scenario S17) slightly increased the number of recruitees, and reached large recruitment trees within a lower number of waves, compared to simulations based on the data (scenario S15) and random mixing parameters (scenario S16). Large recruitment trees were even reached faster when simulations were additionally started with 100 active seeds (scenario S18; see Figures 4A-B).

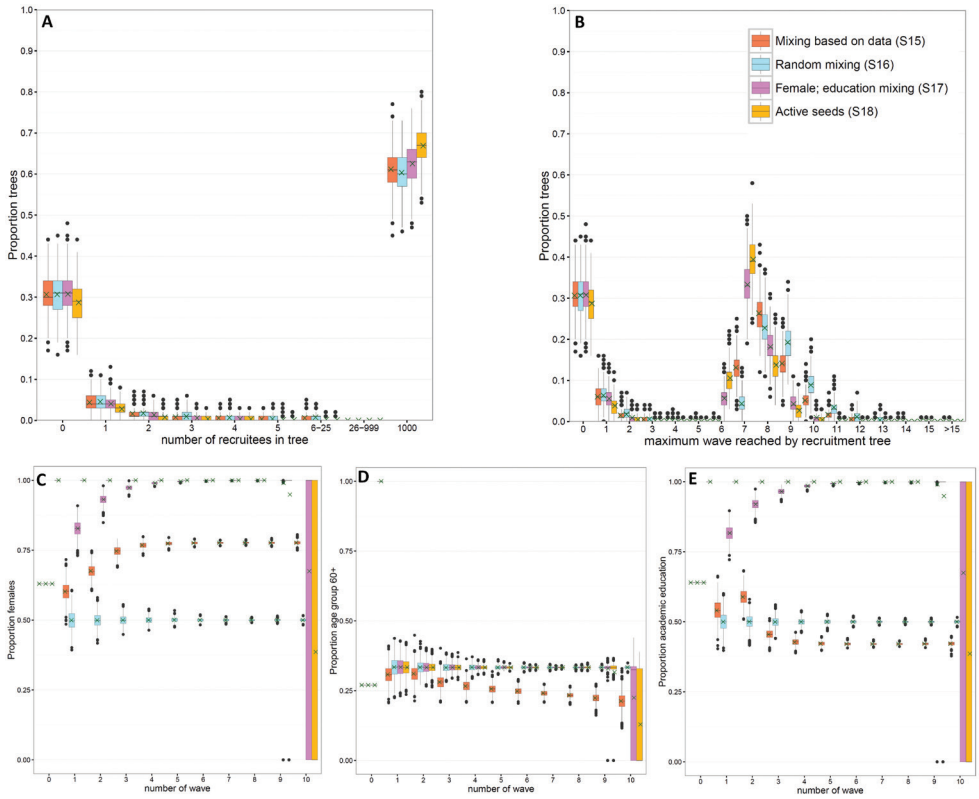


Figure 4. Scenario analyses of mixing behaviour. We ran simulations starting with 100 seeds with characteristics randomly drawn from the observed data (scenarios S15 to S17) and 100 active seeds (scenario S18), 1000 runs per seed. The boxplots show the variability between runs. The legend refers to scenarios S15 to S18 described in Table 3. (**A**) Number of recruitees in recruitment trees in one simulated data set; (**B**) wave reached by recruitment trees in one simulated data set. Plots C to E show, for each wave of large recruitment trees, the variability in composition of each simulated data set, with (**C**) proportion of females, (**D**) age group 60+ and (**E**) having an academic education. A green cross indicates, for each wave, the mean proportion over all simulated data sets.

With random mixing (scenario S16), an equilibrium was reached with a lower number of waves than when mixing proportions were stratified by waves as observed in the data (scenario S15), or mixing where participants had an increased preference for recruiting recruitees with the same characteristics (scenario S17; see Figures 4C-E). With random mixing (scenario S16), the mean proportions of the three characteristics, over all simulated data sets, did not change more than 2% after wave 1. For mixing behaviour based on the data (scenario S15), and mixing behaviour where participants have a strong preference to invite similar others (scenario S17), equilibrium was attained at wave 4 for the proportions female and academic education. However, in scenario S15, no equilibrium was reached within 10 waves for the age group of 60 years and older. If simulations started with a random sample of seeds (scenario S16), strong assortative mixing of females and participants with an academic education eventually led to

proportions of 1 in the sample (i.e., recruiters only invited recruitees with the same characteristics). If simulations started with seeds with these characteristics (scenario S18, with mixing parameters of 1 for females and academic education), the sample only included participants with the same characteristics.

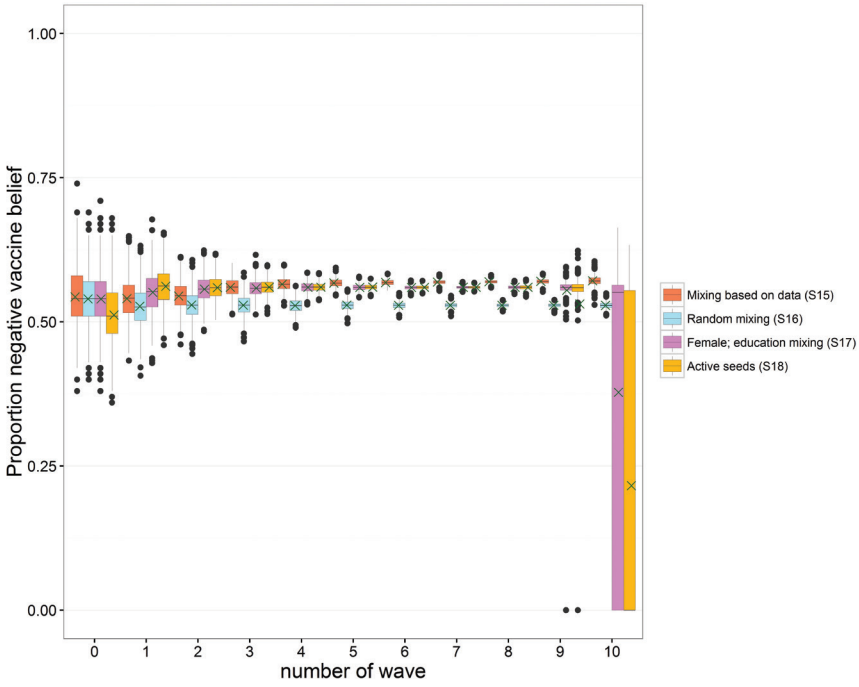


Figure 5. Composition per wave for negative vaccine beliefs. We ran simulations starting with 100 seeds with characteristics randomly drawn from the observed data (scenarios S15 to S17) and 100 active seeds (scenario S18), 1000 runs per seed. The boxplots show the variability between runs. The legend refers to scenarios S15 to S18 described in Table 3. A green cross indicates, for each wave, the mean proportion over all simulated data sets.

The mean proportion of individuals with a negative vaccine belief in each wave did not change more than 2% after wave 1 for simulations based on mixing behaviour as observed in the data (scenario S15; see Figure 5). With random mixing (scenario S16), preferred mixing (scenario S17), and with simulations additionally started with active seeds (scenario S18), equilibrium was attained at wave 1. The logistic regression analysis showed significant influence of sex and age on having a positive or negative vaccine belief (see supplementary Table S6). However, the changes over waves in the combined characteristics of individuals (e.g., the number of females in the age group 60 years and older with an academic education) were small, and therefore did not lead to large differences over waves in the proportion of individuals with a negative vaccine belief.

DISCUSSION

This is the first study, to our knowledge, where a simulation model and empirical data are used to analyse factors specifically important for the success of respondent-driven recruitment. We formulated the recruitment process as a simulation model and used empirical data to quantify parameters. We included heterogeneous recruitment behaviour into our model that depended on individual characteristics. By analysing the impact of changes in model parameters on recruitment, we were able to investigate thresholds for successful peer recruitment and obtain evidence based guidance for future implementation of RDD.

One main finding is that for some parameter combinations, it is more effective if participants send out any number of invitations between 1 and 4 for reaching large recruitment trees, than a situation where the majority of participants does nothing and a low proportion sends out all four invitations. Also, in the former situation, the probability of acceptance appeared to be of relevance for obtaining large recruitment trees for all values of the mean number of invitations. In the latter situation, the probability of acceptance seemed to be of less importance, especially for high values of the mean. In the observed data, the probability of invitation acceptance by invitees (0.19) was too low, and even with increasing the mean number of invitations to a maximum value of 4, large recruitment trees could not be obtained. Here, the average number of successful invitations stayed below 1, and therefore only small recruitment trees occurred. We also examined the relationship between mean and variance of the number of invitations sent out by participants and the probability of reaching large recruitment trees. This allowed us to identify parameter combinations where it is not necessary to achieve the highest possible number of invitation for maximal success of peer recruitment for a fixed variance.

We explored the influence of different mixing behaviour on the recruitment process, by choosing a combination of parameter values that led to mean numbers of successful invitations above the threshold. With assortative mixing, and by starting with seeds active with sending invitations, large recruitment trees are reached faster, within a lower number of waves, compared to random mixing. Nevertheless, assortative mixing (e.g., females only invite other females) led to samples with an overrepresentation of participants with specific characteristics, and equilibrium is reached slower, compared to random mixing. In practice, random recruitment (i.e., participants invite persons randomly from their total pool of contact persons) is difficult to ensure[16, 22], but not necessary in case RDD is used for finding other cases[2]. The observed data showed correlations for the three characteristics that increased over consecutive waves. This indicates a tendency of recruitees to 'copy' the recruitment behaviour of their recruiters. When using RDS estimators, which are often based on a Markov chain model, to estimate proportions of population characteristics, recruitees with certain characteristics who have a higher probability of being recruited receive a larger weighting

factor. However, a violation of the first-order Markov process not only influences the point at which equilibrium is reached (as our simulations showed), it can also influence RDS estimations of the population composition[15].

Our simulations were based on a data set collected during an online RDD survey. During this study, participants were asked to invite contact persons they had physically met in the past two weeks. Assuming that participants invited contact persons whom they actually met, the simulation results are relevant for online RDD studies that sample contact networks relevant for the transmission of respiratory pathogens. A future survey requires a combined approach to reach ongoing peer recruitment and large recruitment trees. When the mean number of invitations sent out by participants is limited, it is important that the acceptation of invitations is high. More information is needed on how to improve peer recruitment in practice. One way to learn more about ways to stimulate peer recruitment is to ask participants directly about the reasons why they do or do not invite others, e.g., as part of a questionnaire or in a follow-up study. However, this will result in information from those participants who are inclined to participate and invite others, and is therefore biased. It would be particularly important to obtain information about individuals who do not participate, but this is obviously much more difficult as they do not take part in the first place. A similar approach is to conduct proper formative research before the actual data collection, to choose the best way to invite others and the most appealing (monetary) incentives, but also by sending reminders and using technical innovations to make participation as easy as possible[23, 24].

It should be kept in mind that models always represent a simplification of reality. We assumed that the probabilities of accepting an invitation by invitees were dependent on the characteristics of the recruiter, not on those of the invitee. This adds uncertainty to our simulations, and may be one the factors why the simulation results are not fully in agreement with the observed data. We require more information on the invitees who did not participate. Such information can be collected by asking the participants more details about the persons they invite. The other limitation concerns the data itself, these were just one realization of the recruitment process and therefore provide limited information about the process in reality. We did perform two other pilot studies using RDD[25, 26], but both samples consisted of small numbers of participants and involved mainly university students of one age group.

In our model, the maximum number of invitations was kept constant (c=4). A larger number of invitations per recruiter will generally lead to a larger number of recruitees[13], but the extent to which this happens is unknown. The marginal benefits of setting a higher number of invitations are likely to be decreasing, as the number of invitations is also dependent on the number of family members, friends and acquaintances that a recruiter is able to invite. Some participants may have few close contact persons who they can invite, while others have many.

Although the branching process describing recruitment may also be studied using an analytic approach, we chose to use a simulation model to be able to include population heterogeneity. Although a simulation model does not provide exact analytic results, it is more flexible for incorporating heterogeneity and correlations between model variables. To illustrate model applicability, we included vaccine belief in the model to investigate the influence of RDD on the proportion of individuals with negative beliefs in each wave. Any other individual characteristic could be added to the model in a similar way, in order to understand the influence of different recruitment behaviour on the sample composition. In a next step, we plan to add more complexity to the model by considering, among others, more covariates of participants (e.g., number of contact persons, infection status, and the behaviour of participants towards prevention programs).

By combining a simulation model with empirical data, we were able to explore the conditions for obtaining large recruitment trees and to investigate how the size and structure of recruitment trees are influenced by heterogeneous recruitment behaviour of participants. The presented model is a helpful tool that can assist public health professionals with preparing research or contact tracing using online RDD. In particular, the simulation model can provide input on the required mean number of successfully sent invitations to reach large recruitment trees, a certain sample composition or a certain number of waves.

Acknowledgements
We are grateful to Martin Bootsma for fruitful discussions on the model implementation and to Albert Wong for helpful input on the R model and formulation.

REFERENCES

1. Stein ML, van der Heijden PG, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, et al. Tracking social contact networks with online respondent-driven detection: who recruits whom? BMC infectious diseases. 2015;15:522.

2. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Koppeschaar CE, Bengtsson L, et al. Enhancing Syndromic Surveillance With Online Respondent-Driven Detection. American journal of public health. 2015;105(8):e90-7.

3. Heckathorn D. Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems. 1997;44:174-99.

4. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol. 2004;34:193-240.

5. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. J Off Stat. 2008;24:79-97.

6. Gile KJ. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. J Am Stat Assoc. 2011;106.

7. Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. Stat Med. 2009;28(17):2202-29.

8. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in social networks. Annual Review of Sociology. 2001;27:411-44.

9. Tyldum G, Johnston L. Applying Respondent Driven Sampling to Migrant Populations: Lessons from the Field: Palgrave Macmillan UK; 2014. 126 p.

10. Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using Respondent-Driven Sampling data. Electron J Stat. 2014;8(1):1491-521.

11. Heckathorn DD. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations Social Problems. 2002;49(1):11-34.

12. Malekinejad M, Johnston LG, Kendall C, Kerr LR, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. AIDS and behavior. 2008;12(4 Suppl):S105-30.

13. Malmros J, Liljeros F, Britton T. Respondent-driven sampling and an unusual epidemic. 2014.

14. Martin-Lof A. Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. J Appl Prob. 1986;23(265-282).

15. Poon AF, Brouwer KC, Strathdee SA, Firestone-Cruz M, Lozada RM, Pond SL, et al. Parsing social network survey data from hidden populations using stochastic context-free grammars. PloS one. 2009;4(9):e6777.

16. Gile KJ, Johnston LG, Salganik MJ. Diagnostics for Respondent-driven Sampling. J R Stat Soc Ser A Stat Soc. 2015;178(1):241-69.

17. Gile KJ, Handcock MS. Respondent-Driven Sampling: An Assessment of Current Methodology. Sociological methodology. 2010;40(1):285-327.

18. Crawford FW. The graphical structure of respondent-driven sampling. Sociological methodology. 2016:1-25.

19. RDS Sim [Internet]. 2010 [cited 16 August 2016]. Available from: http://www.erikvolz.info/rdssimulator.

20. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS computational biology. 2011;7(10):e1002199.

21. Newman ME. Assortative mixing in networks. Phys Rev Lett. 2002;89(20):208701.

22. Heckathorn DD. Extensions of respondent driven sampling: analyzing continuous variables and controlling for differential recruitment. Sociological methodology. 2007;37(1):151-207.

23. Johnston LG, Whitehead S, Simic-Lawson M, Kendall C. Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: lessons learned from four case studies. AIDS care. 2010;22(6):784-92.

24. Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, Nguyen TA, et al. Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. PloS one. 2012;7(11):e49417.

25. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Chanyasanha C, Tipayamongkholgul M, et al. Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and The Netherlands using respondent-driven sampling. PloS one. 2014;9(11):e113711.

26. Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PG, et al. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PloS one. 2014;9(1):e85256.

SUPPLEMENTARY MATERIALS

Text S1. Model formulation

We considered a respondent-driven recruitment process, and modelled it as a multi-type discrete time branching process. The process can be described as follows. Let W denote the wave, i.e., the number of steps the branching process has completed up until that moment. At each wave, there are recruiters that will invite new individuals. By definition, we let the process start from W=0. In this initial wave, we sample individuals from the population, and use them as seeds to initiate the recruitment process (i.e., seeds are the initial recruiters). Each seed can invite a number of individuals to join the study, after which then each invitee may decide to accept the invitation or not. Those that accept, the so-called 'recruitees', may then proceed to invite new individuals in wave W=1. Those new recruitees will end up in wave W=2. This recruitment process is repeated until wave $W=w_{max}$. In our model, w_{max} was determined by the total number of recruitees N; set to 1000 recruitees. If a recruitment tree became larger than 1000 recruitees, the tree showed ongoing recruitment and continued to grow if the model run was not stopped manually.

The branching process depended on the following factors:

- ***The characteristics R of the recruiters***. The recruiters and their subsequent recruitees come from a heterogeneous population, i.e., a population containing individuals with different characteristics. In general, a participant i is characterized by a vector $\boldsymbol{R}_i=(\boldsymbol{R}_{i1},\ldots,\boldsymbol{R}_{iq})$, where $\boldsymbol{R}_{i1},\ldots,\boldsymbol{R}_{iq}$ are q covariates for participant i. In our simulation model, we consider covariates in the form of sex, age groups and education level (as categorical variables). Sex included two categories (females, males), age groups three categories (0-39 years, 40-59 years and 60 years and older) and educational level two categories (lower than academic education and academic education). It is likely that certain strata of sex, age and education are more likely to send invitations.
- ***The maximum number of invitations per (potential) recruiter c (c=0,1,…)***. The researcher may opt to cap the number of invitations per recruiter; a larger number of invitations will in general lead to a larger number of recruitees, but the extent to which this happens is unknown. c is a constant that is equal for all recruiters and for all waves. The marginal benefits of setting a higher number of invitations are likely to be decreasing, as the number of invitations is also dependent on the number of family members, friends and acquaintances of the recruiter.
- ***The number of invitations sent out per (potential) recruiter J (j=0,1,…, c).*** The actual number of invitations is a random variable that takes values between 0 and the maximum c. A higher number of invitations will generally lead to higher number of accepted invitations.

- ***The number of accepted invitations per (potential) recruiter M (m=0,1,..., j)***. The actual number of accepted invitations is also a random variable that takes values between 0 and the number of invitations *j*.
- ***The characteristics R of the recruitees***. The recruitees at wave *W* will become recruiters in wave *W+1*, so the branching process is dependent on the characteristics of the recruitees who are recruited by the recruiter in wave *W*. The characteristics of the recruitees may in turn be correlated with the characteristics of the recruiter. For instance, highly educated women may tend to invite other highly educated women more frequently.

In this study, we assume that the actual number of invitations for recruiter *i*, the accepted number of invitations for recruiter *i* and the characteristics of the recruitees follow probability distributions that depend solely on the recruiter characteristics $\mathbf{R}_i$ (and not on those of the recruitees). In practice, the number of accepted invitations likely also depends on the characteristics of those invited. However, no data on *all invitees* were currently available in the data set[1, 2], so we made the simplifying assumption that the accepted number of invitations depends on the recruiter characteristics only.

The number of invitations successfully sent out

We assumed that the actual number of invitations J_i for recruiter *i* follows a beta-binomial distribution with parameters α and β that depend on $\mathbf{R}_i$ of recruiter *i*, and that can be estimated from the data:

$$J_i \sim Betabin\left(\alpha_{\boldsymbol{R_i}}, \beta_{\boldsymbol{R_i}}\right). \qquad (1)$$

The rationale for a beta-binomial distribution is that it is capable of reproducing bimodal distributions that were observed in the data: the most frequently observed number of invitations occurred at zero and at the maximum value *c*, with any number of invitations in between occurring much less often. The corresponding probability density function is given by:

$$P\left(J_i = j \mid c, \alpha_{\boldsymbol{R_i}}, \beta_{\boldsymbol{R_i}}\right) = \binom{c}{j} \frac{B(j + \alpha_{\boldsymbol{R_i}}, c - j + \beta_{\boldsymbol{R_i}})}{B(\alpha_{\boldsymbol{R_i}}, \beta_{\boldsymbol{R_i}})}, \qquad (2)$$

where *B* is the beta function. The expected value of J_i (which is comparable to the reproduction number of the branching process) is:

$$E[J_i] = \frac{c\alpha_{\boldsymbol{R_i}}}{\alpha_{\boldsymbol{R_i}} + \beta_{\boldsymbol{R_i}}}. \qquad (3)$$

The number of invitations accepted

Given that j invitations were sent, the number of invitations accepted M_i for recruiter i is assumed to follow a binomial distribution:

$$M_i \sim Bin\left(j, p_{\mathbf{R}_i}\right), \tag{4}$$

where the parameter $p_{\mathbf{R}i}$ that describes the probability of acceptation, is dependent on the characteristics of the recruiter $\mathbf{R}_i$, and can be estimated from the data. The probability density function and expected value of M_i respectively are given by:

$$P\left(M_i = m \mid j, p_{\mathbf{R}_i}\right) = \binom{J}{m} p_{\mathbf{R}_i}^m (1 - p_{\mathbf{R}_i})^{j-m}, \tag{5}$$

$$E[M_i] = j p_{\mathbf{R}_i}. \tag{6}$$

The probability that recruiter i recruited m individuals into the survey is given by:

$$P(X_i = m \mid \boldsymbol{R}_i) = \sum_{j=m}^{c} P\left(J_i = j \mid c, \alpha_{\mathbf{R}_i}, \beta_{\mathbf{R}_i}\right) P\left(M_i = m \mid j, p_{\mathbf{R}_i}\right). \tag{7}$$

The expected number of recruitees recruited by a recruiter with characteristic $\mathbf{R}$ is then given by:

$$E[X_i \mid \boldsymbol{R}_i] = \sum_{m=0}^{c} m\, P(X_i = m \mid \boldsymbol{R}_i). \tag{8}$$

Let N_W be the number of recruiters in wave W. The expected number of recruitees in wave W+1, N_{W+1}, is given by:

$$E[N_{W+1} \mid N_w] = \sum_{i=1}^{N_w} E[X_i \mid \boldsymbol{R}_i]\,. \tag{9}$$

The characteristics of the recruitees

The characteristics of $\mathbf{R}_{ik}$ of a recruitee k belonging to recruiter i are assumed to be dependent on the characteristics of i. Ideally, we would use the joint probability distribution $P(R_{ik1}=r_{ik1},\ldots,R_{ikq}=r_{ikq} \mid \mathbf{R}_i)$ to characterize this relationship. The available data were however not enough to estimate this joint distribution with sufficient precision; many covariate combinations $(R_{ik1}=r_{ik1},\ldots,R_{ikq}=r_{ikq})$ had too few, or even no, observations. To construct a credible

joint probability distribution, we assume that the probability distribution of each covariate are mutually independent:

$$P(R_{ik1} = r_{ik1}, \dots, R_{ikq} = r_{ikq} \mid \boldsymbol{R_i}) = \prod_{t=1}^{q} P(R_{ikt} = r_{ikt} \mid \boldsymbol{R_i}), \qquad (10)$$

where $P(R_{ikt}=r_{ikt} \mid \mathbf{R}_i)$ is the probability distribution of covariate R_{ikt}, and can be estimated straightforwardly from the data. Since we only have three covariates in our study, this equation reduces to:

$$\begin{aligned} &P(Sex_{ik} = s_{ik}, Age_{ik} = a_{ik}, Education_{ik} = e_{ik} \mid \boldsymbol{R_i}) = \\ &P(Sex_{ik} = s_{ik} \mid \boldsymbol{R_i}) \times \\ &P(Age_{ik} = a_{ik} \mid \boldsymbol{R_i}) \times P(Education_{ik} = e_{ik} \mid \boldsymbol{R_i}) \end{aligned} \qquad (11)$$

Influenza vaccine belief

As an illustration of the applicability of the simulation model, we added influenza vaccine belief to the model. Hereby we assumed that vaccine beliefs are determined by individual characteristics and do not influence the recruitment process, i.e., recruiters do not invite recruitees with a specific vaccine belief. A logistic regression model was used to estimate the probability of an individual, with a certain sex, age and educational level, of having a positive or negative belief about the influenza vaccine (see supplementary Tables S4, S5 and S6). Vaccine beliefs of individuals were determined based on random draws from the estimated probability distributions, where the probability distributions depended on the sex, age and educational level of individuals.

References

1. Stein ML, van der Heijden PG, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, et al. Tracking social contact networks with online respondent-driven detection: who recruits whom? BMC infectious diseases. 2015;15:522.
2. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Koppeschaar CE, Bengtsson L, et al. Enhancing Syndromic Surveillance With Online Respondent-Driven Detection. American journal of public health. 2015;105(8):e90-7.

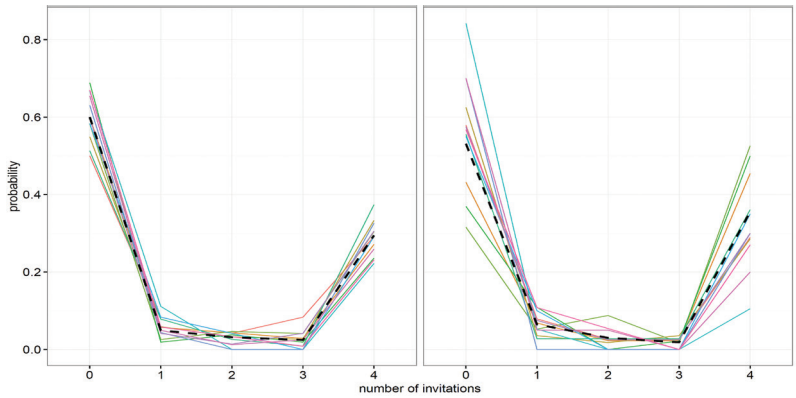


Figure S1. Beta binomial distributions stratified by seeds (left figure) and recruitees (right figure), and by recruiters' characteristics. The coloured lines indicate the different types of recruiters; the means of seeds and recruitees are indicated with the dashed lines. See Table S2 for the corresponding values observed in the data.

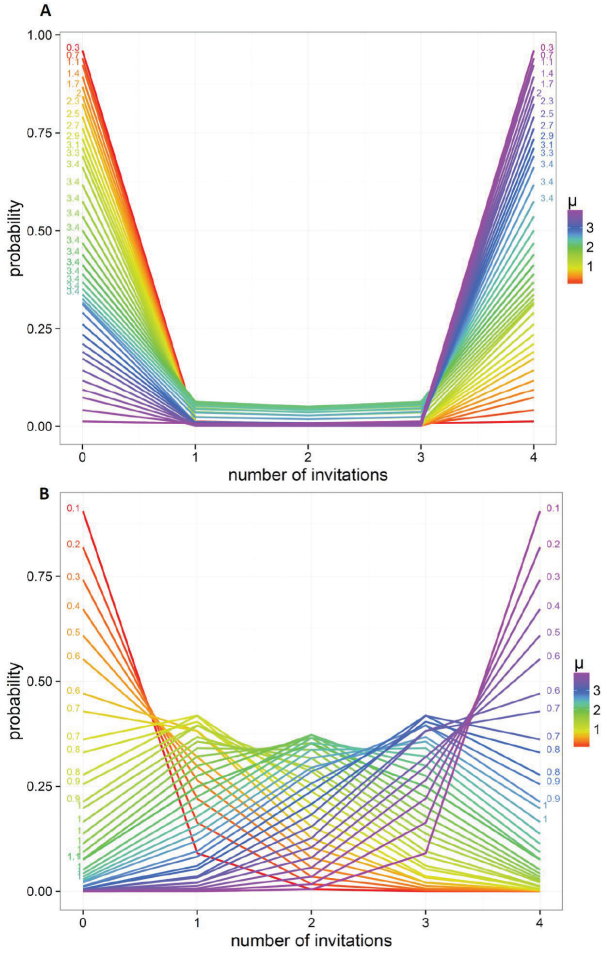


Figure S2. Shapes of the beta-binomial distributions.
(**A**) Beta-binomial distributions for different μ and with the maximum possible σ^2.
(**B**) Beta-binomial distributions for different μ and with the minimum possible σ^2. The values in the plot show for each line the σ^2.

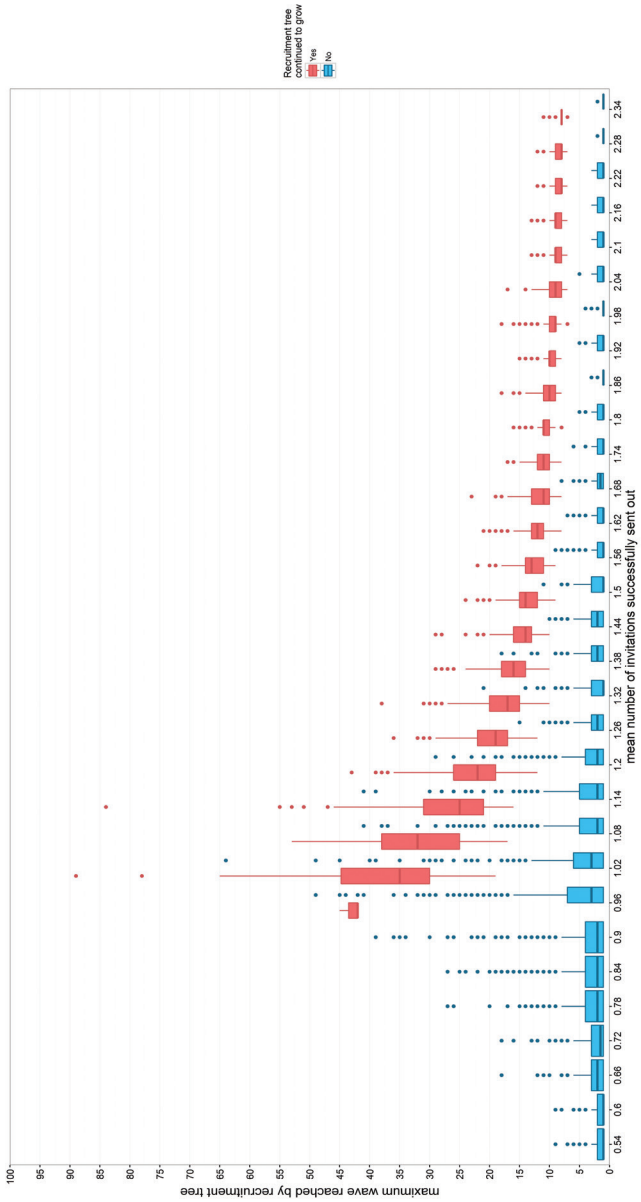


Figure S3. Influence on the maximum wave reached by recruitment trees. We looked closer at the green solid line in Figure 3 that represents a 60% invitation acceptance probability (scenario S7), by varying the mean number of successfully sent invitations (μ^*p). The boxplots show the variability in maximum wave reached by recruitment trees for different mean numbers of successfully sent invitations.

Table S1. Mean number of successfully sent invitations, stratified by recruiter's characteristics, as observed in the data set.

Recruiter			Proportions that sent 0-4 invitations					Beta-binomial distribution					
Sex	Age group	Educational level	0	1	2	3	4	$\bar{x}$	s^2	α	β	p	$\bar{x}$ * p
F	A1	A	0.50	0.06	0.04	0.05	0.36	1.70	3.49	0.08	0.11	0.18	0.31
F	A1	B	0.59	0.05	0.04	0.04	0.28	1.36	3.17	0.07	0.13	0.23	0.31
F	A2	A	0.59	0.06	0.04	0.02	0.30	1.37	3.25	0.06	0.10	0.21	0.29
F	A2	B	0.55	0.04	0.04	0.03	0.35	1.59	3.49	0.06	0.08	0.24	0.39
F	A3	A	0.74	0.00	0.03	0.03	0.21	0.97	2.78	0.02	0.07	0.27	0.26
F	A3	B	0.56	0.05	0.04	0.00	0.35	1.51	3.51	0.05	0.07	0.14	0.22
M	A1	A	0.46	0.12	0.00	0.00	0.42	1.81	3.76	0.07	0.08	0.13	0.23
M	A1	B	0.67	0.07	0.02	0.00	0.24	1.07	2.91	0.04	0.10	0.24	0.26
M	A2	A	0.72	0.02	0.00	0.02	0.24	1.03	3.02	0.01	0.04	0.25	0.26
M	A2	B	0.56	0.06	0.02	0.03	0.33	1.52	3.45	0.06	0.09	0.16	0.24
M	A3	A	0.65	0.06	0.02	0.02	0.25	1.15	3.03	0.04	0.10	0.22	0.25
M	A3	B	0.70	0.05	0.04	0.01	0.20	0.98	2.66	0.04	0.13	0.25	0.24
						mean values		1.36	3.21			0.19	0.26

Table S2. Successful sending observed in the data set, stratified by seeds and recruitees, and by recruiter's characteristics.

	Recruiter			Proportions that sent 0-4 invitations					Beta-binomial					
seed / recruitees	Sex	Age group	Edu	0	1	2	3	4	$\overline{x}$	s^2	α	β	p	$\overline{x} * p$
seed	F	A1	A	0.50	0.08	0.04	0.08	0.29	1.58	3.30	0.12	0.19	0.13	0.21
seed	F	A1	B	0.60	0.06	0.04	0.03	0.27	1.31	3.15	0.06	0.13	0.21	0.27
seed	F	A2	A	0.55	0.06	0.03	0.03	0.33	1.54	3.42	0.06	0.10	0.20	0.31
seed	F	A2	B	0.58	0.03	0.05	0.04	0.30	1.46	3.34	0.06	0.10	0.27	0.39
seed	F	A3	A	0.69	0.02	0.04	0.02	0.24	1.09	2.94	0.03	0.08	0.25	0.27
seed	F	A3	B	0.51	0.08	0.03	0.01	0.37	1.65	3.54	0.06	0.08	0.16	0.27
seed	M	A1	A	0.67	0.11	0.00	0.00	0.22	1.00	3.00	0.05	0.14	0.00	0.00
seed	M	A1	B	0.58	0.08	0.04	0.00	0.29	1.33	3.28	0.06	0.12	0.25	0.33
seed	M	A2	A	0.63	0.04	0.00	0.00	0.33	1.35	3.52	0.02	0.04	0.18	0.24
seed	M	A2	B	0.60	0.04	0.01	0.04	0.31	1.42	3.40	0.05	0.08	0.17	0.24
seed	M	A3	A	0.65	0.05	0.01	0.02	0.26	1.19	3.10	0.04	0.09	0.24	0.28
seed	M	A3	B	0.67	0.06	0.03	0.01	0.23	1.07	2.84	0.04	0.12	0.22	0.23
recruitees	F	A1	A	0.58	0.08	0.03	0.03	0.29	1.37	3.27	0.07	0.12	0.08	0.11
recruitees	F	A1	B	0.43	0.07	0.02	0.02	0.45	2.00	3.72	0.07	0.07	0.19	0.39
recruitees	F	A2	A	0.63	0.04	0.02	0.04	0.29	1.32	3.31	0.04	0.08	0.19	0.25
recruitees	F	A2	B	0.32	0.05	0.09	0.02	0.53	2.39	3.35	0.13	0.09	0.28	0.67
recruitees	F	A3	A	0.37	0.11	0.00	0.02	0.50	2.17	3.66	0.09	0.07	0.17	0.37
recruitees	F	A3	B	0.56	0.03	0.03	0.03	0.36	1.61	3.67	0.04	0.06	0.16	0.25
recruitees	M	A1	A	0.84	0.05	0.00	0.00	0.11	0.47	1.60	0.02	0.14	0.22	0.11
recruitees	M	A1	B	0.55	0.10	0.00	0.00	0.35	1.50	3.63	0.05	0.08	0.17	0.25
recruitees	M	A2	A	0.70	0.00	0.00	0.00	0.30	1.20	3.54	0.00	0.01	0.25	0.30
recruitees	M	A2	B	0.58	0.08	0.03	0.03	0.30	1.40	3.32	0.06	0.11	0.14	0.20
recruitees	M	A3	A	0.70	0.05	0.05	0.00	0.20	0.95	2.68	0.04	0.14	0.32	0.30
recruitees	M	A3	B	0.57	0.11	0.05	0.00	0.27	1.30	3.05	0.09	0.17	0.25	0.32

Table S3. Parameter values for scenarios S15 to S18.

	Recruiter			Proportions that sent 0-4 invitations							Beta-binomial			
seed / recruitees	Sex	Age group	Edu	0	1	2	3	4	$\bar{x}$+0.6	s^2-0.6	α	β	p	$\bar{x}$* p
seed	F	A1	A	0.27	0.13	0.12	0.14	0.35	2.18	2.70	0.41	0.34	1.00	2.18
seed	F	A1	B	0.30	0.15	0.13	0.15	0.26	1.91	2.55	0.45	0.49	1.00	1.91
seed	F	A2	A	0.29	0.12	0.10	0.13	0.36	2.14	2.82	0.34	0.30	1.00	2.14
seed	F	A2	B	0.30	0.13	0.12	0.13	0.32	2.06	2.74	0.37	0.35	1.00	2.06
seed	F	A3	A	0.33	0.18	0.15	0.15	0.19	1.69	2.34	0.48	0.66	1.00	1.69
seed	F	A3	B	0.29	0.10	0.09	0.11	0.41	2.25	2.94	0.28	0.22	1.00	2.25
seed	M	A1	A	0.37	0.17	0.13	0.13	0.19	1.60	2.40	0.40	0.60	1.00	1.60
seed	M	A1	B	0.32	0.14	0.12	0.14	0.28	1.93	2.68	0.38	0.41	1.00	1.93
seed	M	A2	A	0.35	0.12	0.10	0.11	0.32	1.95	2.92	0.27	0.29	1.00	1.95
seed	M	A2	B	0.31	0.13	0.11	0.13	0.32	2.02	2.80	0.34	0.33	1.00	2.02
seed	M	A3	A	0.33	0.16	0.14	0.14	0.23	1.79	2.50	0.43	0.53	1.00	1.79
seed	M	A3	B	0.32	0.19	0.16	0.15	0.18	1.67	2.24	0.55	0.76	1.00	1.67
recruitees	F	A1	A	0.31	0.14	0.12	0.14	0.29	1.97	2.67	0.39	0.41	1.00	1.97
recruitees	F	A1	B	0.27	0.05	0.04	0.06	0.57	2.60	3.12	0.15	0.08	1.00	2.60
recruitees	F	A2	A	0.32	0.14	0.12	0.13	0.29	1.92	2.71	0.36	0.39	1.00	1.92
recruitees	F	A2	B	0.21	0.03	0.02	0.03	0.71	2.99	2.75	0.10	0.03	1.00	2.99
recruitees	F	A3	A	0.26	0.04	0.03	0.04	0.64	2.77	3.06	0.11	0.05	1.00	2.77
recruitees	F	A3	B	0.31	0.09	0.08	0.10	0.42	2.21	3.07	0.23	0.19	1.00	2.21
recruitees	M	A1	A	0.34	0.35	0.21	0.08	0.02	1.07	1.00	2.69	7.36	1.00	1.07
recruitees	M	A1	B	0.33	0.10	0.09	0.10	0.38	2.10	3.03	0.25	0.22	1.00	2.10
recruitees	M	A2	A	0.39	0.11	0.09	0.11	0.30	1.80	2.94	0.24	0.29	1.00	1.80
recruitees	M	A2	B	0.31	0.13	0.12	0.13	0.31	2.00	2.72	0.37	0.37	1.00	2.00
recruitees	M	A3	A	0.34	0.20	0.17	0.15	0.14	1.55	2.08	0.59	0.93	1.00	1.55
recruitees	M	A3	B	0.29	0.16	0.15	0.16	0.24	1.90	2.45	0.50	0.56	1.00	1.90

Table S4. Individuals with positive or negative beliefs as observed in Dutch sample.

Sex	Age group	Edu	Positive vaccine belief	Negative vaccine belief
F	A1	A	35.5%	64.5%
F	A1	B	39.5%	60.5%
F	A2	A	37.8%	62.2%
F	A2	B	48.2%	51.8%
F	A3	A	43.4%	56.6%
F	A3	B	41.7%	58.3%
M	A1	A	46.4%	53.6%
M	A1	B	56.8%	43.2%
M	A2	A	45.4%	54.6%
M	A2	B	42.9%	57.1%
M	A3	A	59.4%	40.6%
M	A3	B	63.9%	36.1%

Table S5. Output logistic regression for vaccine beliefs in Dutch sample.

	estimate[a]	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-0.555	0.150	-3.709	0.000
Rec_age_A2	0.068	0.152	0.448	0.654
Rec_age_A3	0.319	0.156	2.044	0.041
Rec_gender_M	0.434	0.113	3.835	0.000
Rec_edu_B	0.184	0.108	1.702	0.089

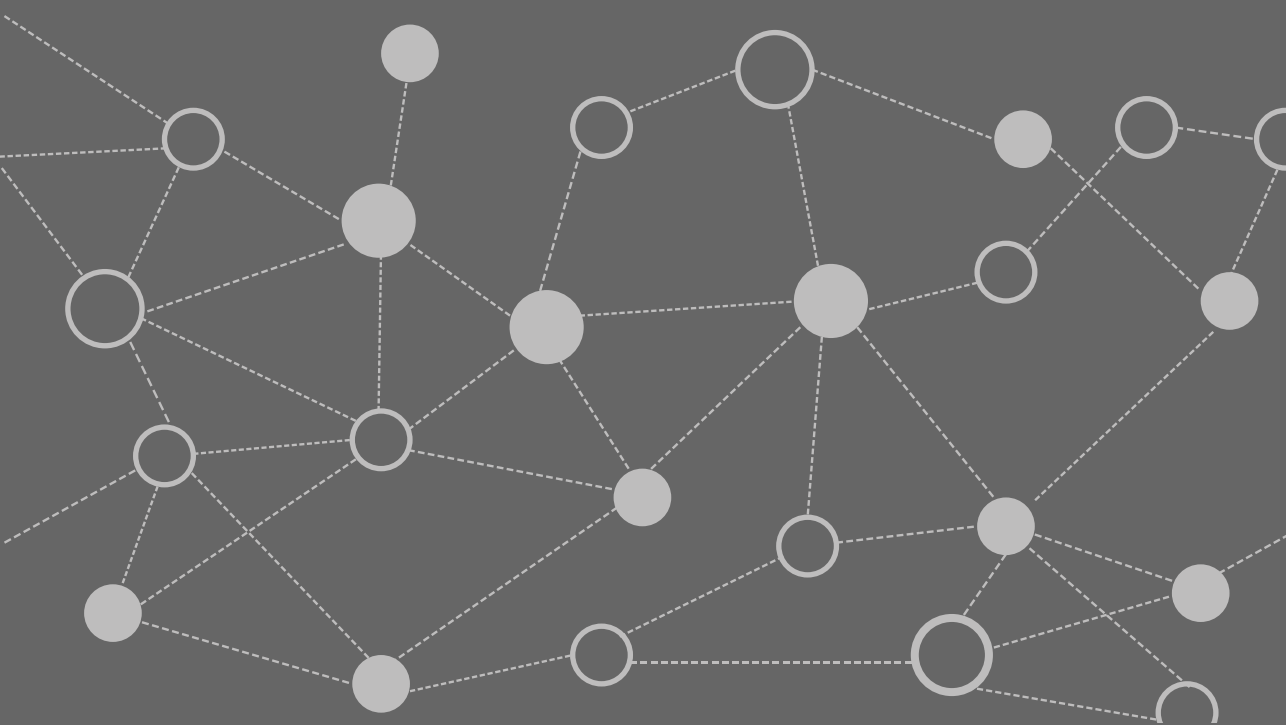
[a]Null deviance: 2000.2 (df: 1447); residual variance: 1971.7 (df: 1443) and AIC: 1981.7.

Table S6. Estimated probability of an individual having a positive or negative vaccine belief.

Sex	Age group	Edu	Positive vaccine belief	Negative vaccine belief
F	A1	A	0.36	0.64
F	A1	B	0.41	0.59
F	A2	A	0.38	0.62
F	A2	B	0.42	0.58
F	A3	A	0.44	0.56
F	A3	B	0.49	0.51
M	A1	A	0.47	0.53
M	A1	B	0.52	0.48
M	A2	A	0.49	0.51
M	A2	B	0.53	0.47
M	A3	A	0.55	0.45
M	A3	B	0.59	0.41

Part IV

In perspective



Chapter 8

General discussion

Online respondent-driven detection for case finding and public health interventions

Submitted

BACKGROUND

Monitoring the spread, preparing for and responding to outbreaks are key public health functions in infectious disease control[1]. Control measures are most effective when based on knowledge on disease transmission and incidence. Such knowledge is, among others, obtained by surveillance, i.e., the continuous and systematic collection, analysis and interpretation of data on cases[2]. A case is an individual with an infection, either with or without symptoms and either being or not being infectious. Surveillance can be defined as 'passive' when it is based on the mandatory reporting of diagnosed cases by physicians or laboratories to public health authorities, or 'active' when public health professionals search for cases, e.g., as in outbreak investigations or screening of individuals at-risk[3]. In this paper, we define both of these ways of identification of cases by public health authorities as 'case finding', providing information for public health action[4].

From the perspective of infectious disease control, the finding of infectious cases is an essential public health task for taking measures aiming to prevent further spread in the population and individual health consequences[1]. Case finding is most useful for mounting an effective response when it is comprehensive and timely[5, 6]. Also, for designing effective strategies for preparedness and intervention, empirical data are required on (potential) pathogen transmission between individuals in order to inform mathematical models (see Figure 1)[7].

However, data collected with case finding is for practically all infectious diseases incomplete and biased[8-10]. For many infectious diseases an infection can occur without symptoms, but asymptomatic cases can still be infectious to others. For some infections that lead to symptoms, the majority of cases only experience mild symptoms and do not seek health care and are therefore not identified. Even if they do seek health care the infection may not be diagnosed, be misdiagnosed or be diagnosed but not reported[3]. Consequently, it is difficult to estimate the true incidence of an infection based on case finding data. Also, timely notification of cases is essential to effective public health action, but cases are often noted after further transmission already occurred, and therefore identified too late[11-14].

Infectious diseases spread from human to human by direct transmission (i.e., direct contact, such as shaking hands, hugging, kissing or sexual intercourse), indirect transmission (i.e., indirect contact, such as via contaminated inanimate materials or objects) or airborne transmission (i.e., suspension of particles of microorganisms that remain in the air for long periods of time, and which may be dispersed over long distances). The spread of respiratory pathogens by droplet sprays, generally over short distances, through coughing, spitting, sneezing or talking, are typically considered as direct transmission[15]. For influenza, measles, mumps, Ebola, and severe acute respiratory syndrome (SARS), a face-to-face conversation within a certain

proximity or physical contact like shaking hands or hugging is likely to be sufficient for pathogen transmission from one person to another[16, 17]. For some of these pathogens, e.g., influenza and rhinovirus, there is limited evidence that they are also transmitted via indirect contact[18, 19]. Therefore, these pathogens do not spread randomly through a population, but follow the structure of contact networks.

Transmission tends to occur mainly between socially connected individuals who have frequently and repeatedly contact such as household members, colleagues and school children, or, in the situation of sexually transmitted infections (STIs), between sexual partners[20-24]. Therefore, cases tend to cluster in time and space (e.g., by setting). This entails that contact persons of cases are at higher risk of acquiring infection than randomly selected persons in a population [25].

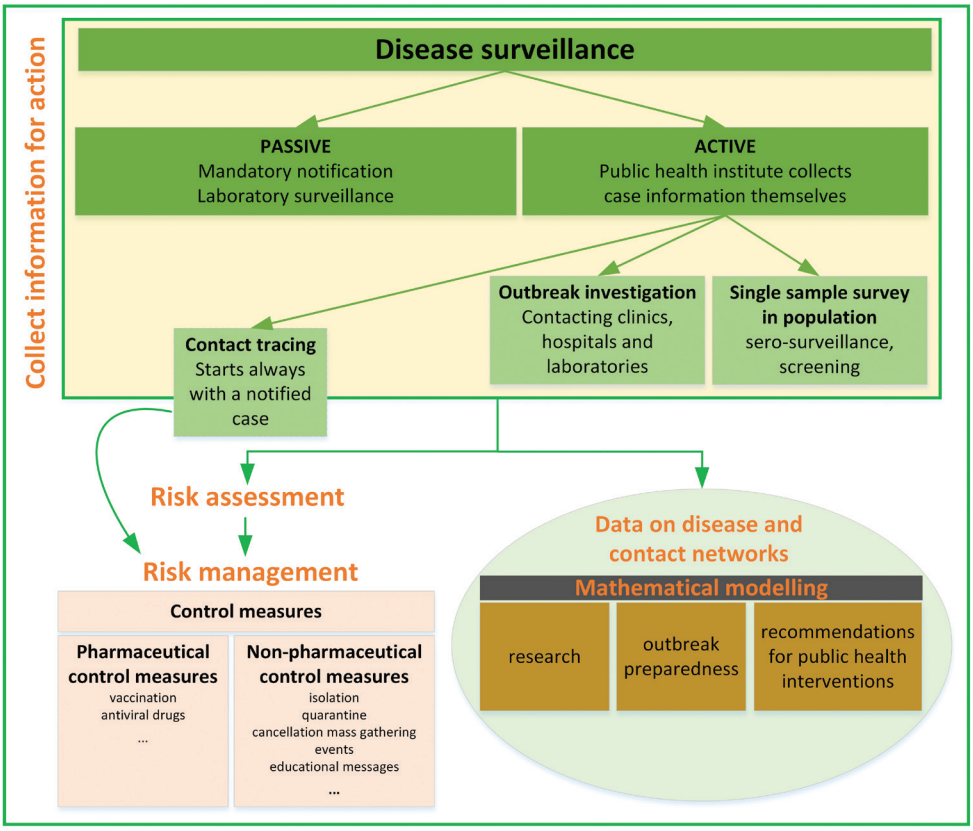


Figure 1. Disease surveillance and control framework. Information for action is collected by passive and active disease surveillance. Public health institutes use this information for risk assessment and management to decide on appropriate control measures. This information also provides input for mathematical models used for research, outbreak preparedness or advising on public health control measures.

Here we introduce a new concept for case finding based on social network structure in a population. We call this method respondent-driven detection (RDD). The method utilises contacts between individuals in a network to find cases. A case is reached through contact with a known case, similar to pathogens spreading through these contact relationships. As individuals are connected in networks, tracking who has contact with whom in a network of infected individuals may help to efficiently identify hitherto 'hidden' cases that go unnoticed by passive surveillance. Contact tracing in active surveillance is based on precisely this principle. However, such systematic tracking of contact persons of cases during outbreaks causes a heavy work burden for public health professionals as most work is done manually, mainly via telephone interviews and house visits.

With RDD cases are asked to recruit contact persons from their network themselves for participation in surveys and diagnostic testing. These contact persons are then asked to do the same, resulting in successive waves of contact persons. Most importantly, with such a peer referral method individuals tend to invite contact persons who are similar to themselves in terms of social demographics such as age, educational level and behaviour, and who may also experience similar symptoms. RDD may therefore enhance case finding, providing further insight in the extent of outbreaks, in a quick and less laborious manner for public health professionals.

In section I, we provide a general introduction to network-based approaches and explain where RDD originated from. In sections II and III, we discuss how RDD can be used to collect epidemiological information on contact networks and increase knowledge on the spread of pathogens within these networks. In section IV, we explain how RDD utilizes the network of cases to identify other cases. In section V, we discuss the prerequisites for online RDD and highlight challenges. Finally in section VI, we discuss its future value for the ultimate goal of case finding, namely to intervene during outbreaks and to reduce transmission in epidemic and endemic situations (i.e., to eliminate sources by diagnosis and treatment, isolation and quarantine, and behavioural hygienic advice; to protect contact persons by vaccination or early treatment), and compare practical use of RDD with traditional methods of case finding.

Section I. A network-based approach for case finding

In the past decades, contact patterns of humans and their implications for transmission of infectious agents have received much interest in infectious disease epidemiology[26]. Patterns of relations in a population can be represented, albeit simplified, as a network where individuals are depicted as nodes, and relations as links or 'edges'. In a social network, these links represent a form of social relation, like friendships or co-worker relationships[27]. A social network is, among others, relevant for social influence and information diffusion. In this paper,

a distinction is made between social networks, and contact networks, the latter being mainly relevant for the transmission of infectious diseases. There may be considerable overlap between these two types of networks, e.g., when an individual's decision for vaccination is influenced by the choices of his social contact group, clusters of susceptible individuals may arise. Disease outbreaks can occur within such clusters, even in a country with a high vaccination coverage[28, 29].

The experiments by Milgram et al. in 1969 suggested that society is a small-world network, with many local and only a few global links. These experiments were designed to measure the average path length of social networks in the United States[30]. They demonstrated that a person can reach any other person in just six link steps on average. This later became known as the 'six degrees of separation'. Although these experiments were not related to infectious disease transmission, they clearly illustrated that individuals are interconnected, and that only a few global links are sufficient to generate a small-world network[31]. The average number of link steps between any two individuals in the world is likely reduced due to high mobility of contemporary society, e.g., through air travel, which facilitates the speed, spread and persistence of infectious diseases[32].

During outbreaks of respiratory pathogens, interventions such as closing of schools or workplaces are often considered hoping to reduce the number of local links (e.g., by avoiding direct contact between schoolmates or colleagues) to slow down pathogen transmission. However, these interventions may be less effective if there are alternative links between individuals in a network. For example, if children play after school frequently with neighbourhood friends or relatives, school transmission might be overestimated and measures like closing schools might be less effective[33]. The density of network links beyond the first wave of contact persons influences, among others, the speed at which a respiratory pathogen spreads through a population[34]. If interventions are applied to specific social nodes (i.e., key individuals in infection dynamics), they may be as effective as the presently applied measures on large populations (e.g., measures that reduce all human contacts by refraining from using public transport, or prohibiting mass gatherings). However, in practice, it is often difficult to identify these key individuals to be able to target interventions to them.

Snowball sampling

To map a complete network of links within a population a sociometric study is required, i.e., a study that includes all individuals and the people they have contact with within a community. As sociometric studies are extremely labour-intensive and infeasible for large populations, less labour-intensive ego-centered studies are often chosen as second best. In the latter, egos (i.e., respondents) and their immediate contact persons are included, or egos are asked to provide information on their immediate contact persons. Such an ego-centered study provides partial network information[31].

Snowball sampling is a well-known technique that was introduced to sample individuals in so-called hidden populations that are difficult to reach with random sampling methods, such as injecting drug users or men who have sex with men. Here, an initial member of the target population is recruited into the study. Then either participants are asked to recruit their contact persons or provide names of these persons for recruitment by researchers. This process is repeated through a succession of waves of sampling[31].

A snowball method by definition leads to a sample where respondents are not independent. Participants tend to recruit contact persons whom they know well and with whom they share characteristics[35]. The obtained sample may therefore be biased and not representative of the population. Another reason for sampling bias is that individuals with a high degree, i.e., a high number of contact persons, have a higher probability of being recruited into the sample than individuals with a low degree[36]. Respondent-driven sampling (RDS), a specific type of snowball sampling, was introduced to address these issues[37]. It combines snowball sampling with a statistical model to correct population estimates, e.g., of disease prevalence or risk factors for biases introduced during sampling, or to estimate the size of populations[36, 38, 39]. In contrast to snowball sampling, with RDS researchers keep track of who recruited whom and their numbers of social contact persons. Enrolment is achieved by providing participants with paper coupons to hand out randomly to persons in their social network. The number of coupons per participant is often limited to three or four to promote compliance among participants and to ensure longer recruitment chains, and therefore to allow the sample to diversify from typically non-randomly selected seeds. Each issued coupon contains a unique token that allows researchers to anonymously track peer recruitment without participants having to identify their contact persons. RDS normally relies on a double incentive structure, where participants receive one incentive (often monetary) for participation and additional incentives for successfully recruiting their contact persons[40].

In this paper, we distinguish RDD from RDS. Although both methods are used to recruit individuals in a target population, the aim is different. The main objective of RDD is to detect cases or clusters of cases within contact networks, rather than to estimate population proportions from samples. Similar to RDS, RDD starts with initial cases, i.e., invitees called 'seeds' who are asked for participation, and subsequently invite contact persons whom they have met face-to-face during a specified time period. However, with RDD, cases are the so-called index cases in traditional communicable disease control, i.e., individuals with a diagnosed or confirmed infectious disease. Recruited contact persons are asked to do the same resulting in recruitment waves. Therefore, RDD allows for analysing contact persons of contact persons, and beyond, and the links between them. This provides more information about the contact network structure as compared to studies that sample participants randomly and independently of one another. In fact, this sampling method to find hidden cases is the basis of traditional contact tracing in public health.

Online recruitment

In the past decade, communication is increasingly digitised due to a substantial increase of Internet and mobile phone usage[41]. This trend is also increasingly taken into account by national governments, e.g., Dutch citizens are invited to use online services for tax declarations or similar administrative activities[42]. The digitisation of society has led to the phenomenon of 'digital epidemiology', which is 'the use of epidemiological knowledge and digital technologies to enable disease surveillance and epidemiological research'[2]. This includes the collection of epidemiological data by interviewing individuals via Internet or mobile devices, and extracting data from social media (e.g., Facebook and Twitter) and online databases, e.g., based on search queries (Google) and consumer buying behaviour[43]. The use of the digital technologies provides a wide range of new opportunities for public health, including near to real-time insights in influenza activity levels based on people's search activities[44], and analysis of sentiments towards pandemic vaccines[29]. There are, however, some doubts about prospective accuracy of this tool[45, 46]. In addition, in the domains of outbreak investigation and case finding, the Internet is increasingly explored as a tool for targeted recruitment of cases and contact persons at-risk for infection[47, 48]. Increased Internet use also facilitated the emergence of participatory surveillance, enabling real-time monitoring of diseases through the regular submission of health-related information by volunteers[49].

Most RDS studies, as most existing contact-tracing strategies, are 'offline', they use time-consuming and logistically demanding paper questionnaires or face-to-face interviews. This requires local research sites where participants collect their reward and receive paper coupons for inviting contact persons. RDD focuses solely on face-to-face ('offline') contacts that permit the transmission of pathogens, and, therefore, it would seem appropriate to use offline recruitment. However, in communicable disease control, timely and comprehensive case finding is important for effective treatment and control measures to prevent further pathogen transmission. Via the Internet, e.g., via e-mail, coupons containing unique identifiers may be distributed more efficiently and faster within social networks[50]. The popularity of social media[51, 52], and instant messenger applications (e.g., WhatsApp) create novel possibilities to make it even easier for participants to directly invite contact persons in an informal manner. Furthermore, persons can be tracked and reached in a large geographical area[53-55]. For example, an RDS study focusing on men-who-have-sex-with-men (MSM) in Estonia compared an offline and online approach, and observed that via the Internet a more diverse group of MSM was reached, i.e., cases that otherwise using an offline approach would remain hidden[56]. These are promising arguments for using an online approach for case finding.

Section II. Online RDD to study contact networks

Mathematical disease models can help to understand the spread and impact of new or re-emerging infectious diseases, and for determining the efficacy of intervention strategies, such as school closure, vaccination and cancellation of mass gatherings[7]. Simple mathematical models assumed homogenous mixing within populations. This means that any pair of individuals is equally likely to have contact with each other during a given time interval. In reality, people tend to mix preferentially with certain types of individuals and contact patterns are thus heterogeneous. Therefore, predictions on the probable impact of pathogens and interventions may be more accurate if models incorporate contact mixing patterns[26]. However, the parametrisation of such models requires real-life data on contact networks.

Collecting empirical data on contact patterns can be challenging, mainly because it is difficult to measure the occurrence of contacts and to define what constitutes a contact[26]. Furthermore, some pathogens have more than one mode of transmission. STIs are transmitted from one person to another via sexual contact, but, e.g., hepatitis A and shigellosis are well known STIs among MSM, but can also be transmitted faecal orally or via contaminated food. The contribution of close (less than 1.5 meters) contact or the contribution of physical contact (e.g., shaking hands) in the overall spread of respiratory pathogens, is still in debate[26]. This may complicate the ascertainment of the route via which a pathogen is transmitted, and consequently the determination of effective control measures during an outbreak.

In the past decades, multiple studies applied various methods and techniques to collect empirical data on contact patterns, mainly those relevant for the transmission of respiratory and sexually transmitted pathogens[26, 57-59]. For small populations in closed settings such as day-care centres, schools or health care facilities, direct observation via video cameras can be an accurate method to measure detailed information about type and duration of contact. However, direct observation is limited to geographically confined locations and may underlie ethical constraints[26]. A similar automated method is the application of proximity sensors. Here, sensors measure spatial proximity between participants carrying electronic tags[60]. Although the burden for participants is low, the contact network data is limited to those wearing a device and the sensors may record contacts that are not relevant for pathogen transmission[26]. In certain situations, it may also be possible to extract information from available data sources to infer where and how often individuals interact. For example, Bengtsson and colleagues used data on mobile phone usage to observe the spatiotemporal movements of people during a cholera outbreak following the Haiti 2010 earthquake[61].

Another method that is often used to collect data on contact patterns is to ask people about their contacts, either using a direct interview or a contact diary. This requires a rigorous

definition of 'contact' between individuals. Participants are usually asked to record themselves certain physical and conversational contacts made within a certain time period at different settings. A commonly used definition of a contact relevant for the transmission of respiratory pathogens is physical contact (e.g., kiss or handshake) or a two-way face-to-face conversation with three or more words[62]. This can be difficult for participants if they have to report over a long time period or if a participants had many different contacts at different settings, even when it only concerns the day before the day of participation. Contact diaries are therefore subject to recall and reporting bias[63, 64]. Nevertheless, a contact diary has important benefits over all other methods. In particular, they are easy to distribute, and able to collect data in a wide range of situations and settings[26]. Contact diaries have been applied in specific settings (see, e.g., [65, 66]), and in general populations (see, e.g., [62, 67]).

The majority of diary studies used an egocentric design to collect data on contact patterns. These studies usually collected information on the type of contact (physical versus non-physical), duration and frequency of contact, and the setting where contact occurred[26]. These studies identified contact patterns that may have important implications for disease transmission, such as the existence of population subgroups with very different contact behaviours[62, 65]. Furthermore, one study also identified mixing at specific settings, which differed between workdays and weekends. This study also highlighted the importance of household contacts as a bridge to other settings[68].

However, an egocentric design cannot provide information about the network structure beyond the contact persons reported by participants. Furthermore, egocentric studies often only provide limited information on the participant's contact persons, e.g., the socio-demographic information on contact persons is often limited to age and gender[26]. Therefore, it is not possible to analyse links between contact persons of contact persons within the same network, nor to analyse mixing patterns by other characteristics that may play an important role in pathogen transmission, e.g., mixing by degree or vaccination status[69, 70].

In [71, 72], we reported on pilot studies in Thailand and the Netherlands in which the contact diary design was combined with online RDD (**Chapters 2 and 3**). Participants were asked to fill in an online questionnaire and report, among others, their numbers of contact persons in different settings. For respiratory transmission a contact person was defined as a person standing or sitting within reach of an arm's length to the participant for 30 seconds or longer. This definition was not limited to physical contact or a conversation, but included a broader range of contacts that may also be relevant for pathogen transmission, such as contact with persons in a crowded bus or train.

In both samples, overdispersed degree distributions (i.e., variance of degree was larger than

the mean degree) were observed. This means that most individuals reported only few daily contact persons and some a substantially higher numbers of contact persons[31]. In theory, a person with a high degree has a higher probability of becoming infected and to infect others. Such a person is more likely to act as a bridge between communities and thus important to target with public health control strategies[73]. The focus of these pilot studies was to sample the underlying contact network relevant for directly transmissible pathogens. Therefore, after completing the questionnaire, participants were asked to invite four contact persons whom they had met in the preceding week. Here, not only an individual was sampled and interviewed, but also his or her contact persons, contact persons of contact persons, and so on. This enabled researchers to analyse larger connected parts of contact networks[71, 72].

Sampling more than one wave of contact persons provides insights in mixing patterns within networks, therefore broadening our understanding of network structure beyond information on degree distributions[74]. In the studies conducted in Thailand and the Netherlands, random mixing by degree (thus not assortative, nor disassortative) was observed. This supports the use of the proportionate mixing assumption (i.e., random mixing weighted by the contribution of a subgroup to the total number of contacts in a population[75]) in mathematical models[71]. Despite strong socio-cultural differences between the countries, participants showed strong assortative mixing patterns by demographic variables such as age, gender and education in both samples[71]. This indicates that participants have a tendency to connect to contact persons with similar characteristics. When individuals have contact primarily with peers who are alike, infections are likely to spread faster within those subgroups than between subgroups. Thus, using these data, more advanced mathematical models that rely on more detailed network data can be informed and calibrated to better predict patterns of disease spread.

Section III. Studying the spread of pathogens through networks

Although the amount of data on contact patterns considerably increased in the past decades, little is still known on how mixing patterns are related to pathogen transmission. Wallinga and colleagues demonstrated that contact data collected with a diary design can capture age mixing behaviour that can explain age-specific patterns of mumps seroprevalence in a population[76]. Similar studies showed this also for parvovirus[77] and pertussis[78], and for the incidence of 2009 H1N1v influenza[79]. These results provided indirect proof that contact diaries can indeed capture social encounters that are relevant for infection transmission.

Empirically linking contact data to the transmission of infection is challenging. Ideally, information is required on who is infected (and who is not) and when within a population, but also on the number, type, duration and setting of encounters between cases and contact persons, and on prior immunity[26]. Transmission links can be ascertained by combining

epidemiological and genomic data, e.g., by pathogen typing methods, from broad (e.g., phenotyping) to highly specific (e.g., whole-genome sequencing)[80, 81]. Nevertheless, it is impossible to fully determine the transmission of a pathogen from one person to another, even when the epidemiological data or genetic data seem to point to a possible transmission link[82].

Only a few small studies have attempted to quantify transmission risk by simultaneously collecting contact data and microbiological samples to confirm infection[83]. Special investigations that use contact tracing may also uncover information that relates network structure to transmission risk. Detailed contact tracing during the SARS outbreak in 2003 in Singapore provided important disease related information (such as the incubation period distribution and serial interval distribution) and allowed for linking cases in a transmission tree (i.e., a graph that describes possible transmission events between infected hosts: an estimation of who infected whom). These trees revealed social heterogeneity among contact persons of cases (e.g., health care workers, family, and visitors of a health-care facility) and identified so-called super-spreaders, i.e., cases who infect many more secondary cases than average [84]. However, contact tracing is labour-intensive, expensive, and often difficult to perform. In practice, contact persons are only traced when an index person tests positive for a disease, and they cannot always be identified by health care professionals, which leads to biased samples of contact persons.

Using self-reported data on disease symptoms can be a convenient alternative if microbiological testing, is infeasible, too expensive, or non-discriminatory (e.g., with measles). A clinical diagnosis based on self-reported symptoms can be a quick approach to obtain an indication on symptomatic cases in a population. Bates et al. associated the risk of acquisition of self-reported diarrhoeal infection with self-reported numbers of food-sharing contact persons[84]. During the winter season 2013–2014, we combined online RDD with a large participatory surveillance panel to study contact networks and the spread of respiratory infections in the general population in the Netherlands and Belgium (**Chapter 4**)[86]. In this cross-sectional study, participants were asked to report their contact persons from the previous day at specific settings, as well as any influenza-like-illness (ILI) symptoms that they experienced in the past two weeks. Afterwards, each participant was asked to recruit four persons with whom they had contact in the preceding two weeks.

Compared to earlier studies, we were able to analyse a large number of recruiter-recruit pairs (n=488), and their spatial distribution over a large geographical area. This study reaffirmed that people mix with contact persons similar to themselves in terms of demographic characteristics such as age and education; these relationships were influenced by the geographic distance between two persons. In this study, participants were also asked whether they had received

influenza vaccination in the past 12 months (as a proxy for immune status), and whether they believed the vaccine protected them against influenza. Assortative recruitment was observed by vaccination status and sentiments towards the vaccine. This suggests clustering of vaccine-induced immunity in a population. Such clustered patterns of similar health behaviour have been described before[29, 87]. Clustering of unvaccinated people or people with a negative attitude towards vaccination leads to clusters of susceptible persons, which can increase the likelihood of outbreaks[29].

Most importantly, in our cross-sectional study assortative recruitment by self-reported symptoms was observed. This means that symptomatic participants had a tendency to recruit symptomatic contact persons, possibly because they were more frequently present in their direct contact networks. It confirmed that RDD can be used to sample the underlying contact networks of individuals that are relevant for the spread of infectious diseases that transmit via close contact. As all participants, with and without symptoms, were asked to invite close contact persons, we obtained a good insight into the local network of individuals, and probably also captured asymptomatic transmitters[86]. Following potential transmission paths with RDD, preferably combined with microbiological testing, may further improve our understanding of infectious disease dynamics by providing information on important parameters such as transmission risk and infectivity.

Section IV. Using the network to find cases

Passive surveillance is used to monitor time trends in incidence and distribution of cases, i.e., temporal, spatial and by socio-demographics, in a population, and of possible exposure and infection determinants[3]. Commonly it provides aggregated data of confirmed reports of disease or condition occurrence that may lead to action[88, 89]. By combining various inpatient and ambulatory health care records that (often) precede case diagnosis, it is possible (for some infectious diseases, such as influenza) to obtain further insights into the unreported disease incidence in populations[3, 4]. Without confirmation, syndromic surveillance has an important utility in developing countries, as it is relatively inexpensive and faster than systems that require laboratory confirmation, but also in developed countries syndromic surveillance is used[90]. Although no scientific evidence is available on the effectiveness, syndromic surveillance is believed to improve the early detection of outbreaks and real-time monitoring of disease trends. To achieve that, syndromic surveillance relies on the existence and real-time accessibility of data sources, which incompleteness unfortunately poses the biggest challenge [43, 91].

In many high-income countries, influenza surveillance is a combination of reports of ILI that is collected by a sentinel system of general practitioners (GPs), and microbiological testing

of samples of symptomatic cases[92]. However, this represents only the tip of the iceberg of all influenza cases, as the majority of symptomatic cases suffer only mild symptoms and do not visit a GP, nor require treatment in a hospital[93]. A substantial part of all cases therefore is not captured by these elaborate surveillance systems. With the little information on the rate of seeking health care with ILI, estimates of the actual disease incidence have a broad range.

A web-based participatory surveillance system was initiated in the Netherlands in 2003, where volunteers submit ILI related information on a regular basis. This method facilitates near to real-time monitoring of ILI in the community and provides data on the rate of attendance of GPs[94]. This information is essential for estimating influenza incidence. This participatory system provides information on the proportion of symptomatic individuals who actually visit a GP, which then allows estimation of the proportion that is hospitalised. Moreover, data is collected longitudinally and on an individual level, providing insight in disease specific characteristics (e.g., duration of symptoms and severity of symptoms) and making it possible to study influenza vaccine efficacy[95].

One could further enhance case finding if the contact network of identified cases can be used to detect other cases. In [96], we were interested in whether the contact network of individuals reporting certain symptoms can be used in practice to detect others with similar symptoms. To test this hypothesis, we compared volunteers of the participatory surveillance system, who participated in our questionnaire, with contact persons recruited via RDD (**Chapter 5**). Participants with symptoms recruited more symptomatic contact persons than did participants without symptoms[96]. This suggests that online RDD can indeed enhance the identification of symptomatic cases or clusters of cases by making use of a case's contact network. Furthermore, the questionnaire was distributed via peer-driven recruitment in several waves through all Dutch provinces and reached, within a short period, individuals from all age groups, those with a wide range of household compositions, and those at a variety of educational levels. This demonstrated that RDD through web based participatory panels has a large geographical coverage and that timely, detailed information about participants and their contact persons can be obtained[96]. Not essential for influenza, but for other infectious diseases finding cases rapidly for early treatment and prevention of transmission can be crucial in controlling the outbreak[14].

Besides recruiting contact persons into the sample, participants can also provide information about others in their local environment, thereby acting as a sentinel in their network. In the RDD questionnaire, symptomatic participants were asked to report on contact persons who experienced similar symptoms in the preceding two weeks. More than half of symptomatic participants reported occurrence of similar symptoms among close contact persons, such as family, housemates and colleagues. This shows that RDD can potentially be used to detect

cases via the contact network of cases.

In standard use of RDD it would not be necessary to ask all volunteers of participatory panels to regularly recruit their contact persons. Instead, RDD may only be implemented for those volunteers who report a certain combination of symptoms, who may then recruit other symptomatic individuals in their social environment. RDD is likely most suitable for active surveillance (rather than for continuous, passive surveillance) to find as many cases as possible. Here selective recruitment by participants is of less importance and even a prerequisite of penetrating the clusters of cases, e.g., when participants invite selectively contact persons whom they know are symptomatic. However, this objective may interfere with the objective of studying contact networks of individuals (including of those not infected), and wherein it is important that participants randomly invite contact persons from their network, for enabling appropriate use of statistics to correct for biases.

Section V. Limitations and determinants of successful online peer recruitment

Limitations

Online RDD has the same advantages as online RDS, such as anonymous recruitment and quick recruitment of contact persons, but it also faces similar challenges[50, 96]. In particular, cases need to be motivated to recruit their contact persons to participate in the study or to report individuals with symptoms. If they fail to recruit their contact persons, it is difficult to obtain long recruitment chains as recruitment stops after just a few waves. In all our studies we observed low proportions of peer-recruitment[71, 72, 86, 96]. On average, less than half of all seeds invited a contact person, a proportion not sufficient to generate long recruitment chains. Participants expressed concerns about privacy and not wanting to bother contact persons with a questionnaire, withholding them from sending invitations to contact persons. Although many Internet-users share information with each other via social media[97], sending an invitation for a questionnaire specifically to a number of contact persons is a step that many participants did not take[96].

From RDS studies, we learned that a double incentive structure, namely receiving a monetary incentive both for filling the questionnaire and for successfully recruiting a contact person, can be motivational in some target populations[53, 54]. However, using monetary incentives can have a down side as they might be associated with cheating (e.g., think of cheaters who recruit themselves to collect a reward) or with moral objections to earn money on other person's disease. With RDS participants receive a limited number of coupons (usually three or four) for recruitment of contact persons. The idea is that this will lead to higher participation over all waves and thus to longer recruitment chains, and therefore deeper penetration into the (possibly hidden) population. Theoretically, a minimal number of recruitment waves is needed

for reaching a stable distribution of population characteristics in the sample and therefore 'forget' about the non-randomness of seeds. With RDD, participants should be able to invite all relevant contact persons, both for the purpose of analysing contact networks, as for the purpose of identifying new cases. However, there is a trade-off between the number of coupons given to participants and their compliance to invite contact persons. Displaying a long list of empty coupons for sending might be discouraging for participants, limit participation, and hence be counterproductive. A solution could be to allow participants to choose any number of contact persons they want to invite. However, when more contact persons are invited, and many contact persons are also contact persons of one another (i.e., contact transitivity), some persons may receive multiple invitations. Records of contact transitivity, and subsequent clustering of survey invitations, provide important information on network structures. But measuring this with online RDD has proven difficult. Repeated digital invitations (e.g., e-mail invitations) are often ignored by recipients, even if they only have to click one button to identify themselves as an earlier participant. On the other side, repeated invitations might provide additional peer pressure to persuade individuals to participate. The experience with online RDD is presently too limited to draw conclusions on how to deal with this problem.

When studying contact networks, an important question is how well recruitment patterns reflect the contact patterns of participants. The representativeness of the network links depends on the extent to which participants randomly invite persons from their 'pool of contact persons' and how closely this choice reflects contacts that are relevant for pathogen transmission. In our pilot studies, we observed a tendency of participants to recruit contact persons with similar traits; this is a requirement for case finding, but might also be a limitation of peer recruitment when it comes to using RDD for studying contact network structures. From RDS theory, it is known that recruitment may be biased towards individuals with a higher degree[35]. There can also be numerous practical reasons why participants invite specific contact persons, especially during an online survey. For example, a participant may not have digital connections with, or phone numbers of the contact persons they are asked to invite according to the recruitment criteria. With online RDS, as with online RDD, it is difficult to determine which contact persons participants invite from their pool of contact persons, which contact persons end up in the questionnaire, and importantly, who decides not to participate. However, a comparison of our recruitment patterns with contact patterns collected during a large egocentric diary study (POLYMOD[62]) showed strong similarities. The recruitment patterns by age resembled the contact patterns by age collected from randomly sampled individuals[86]. To further improve our understanding of who recruited whom and which of the invited contact persons eventually participated, participants may be asked to define the relation (e.g., type and frequency of contact) with their recruiter and indicate whom they invited (e.g., by providing the age group and gender of the invited contact persons).

In our study described in [96], ILI was defined as a combination of fever, and at least headache or muscle pain, and at least a cough or sore throat. Participants who reported symptoms that began more than 3 weeks before they participated were excluded. Due to low recruitment of contact persons by participants who met this ILI definition, we were unable to estimate by how much RDD improved ILI case finding, compared to the participatory surveillance panel and traditional ILI surveillance. The probability of case detection with RDD depends on numerous factors, such as the incidence and type of disease (e.g., incubation period, recognisability of symptoms, laboratory diagnostics). Probably even more important, RDD efficacy depends on methodological factors, such as the extent to which symptomatic individuals are willing to fill in a questionnaire and recruit their contact persons[96].

In non-research settings with perceived threats of other emerging infections, the motivation and participation of individuals may be considerably higher than during our pilot studies on ILI. During a large measles outbreak in the Netherlands, we presented a hypothetical situation to a random sample of parents in the general Dutch population. Parents were asked about their willingness to inform the parents of their children's contact persons, in case their child was diagnosed with measles. A majority of these individuals said they found it important to inform other parents personally, and did not feel guilty or embarrassed about doing so (see Figure 2). A majority (79.2%) also had the contact details (e.g., phone number or email address) of all, or part of, the parents of the children with whom their child most frequently has contact. However, during the same measles outbreak parents of children with measles in a population group with religious concerns about vaccination were not willing to participate in an RDD study at all.

The use of online RDD for contact tracing may raise ethical and privacy concerns, particularly in situations that provoke social barriers related to transmission (e.g., disclosure concerns regarding early pregnancy). Exploiting the social network can reveal the identity of a case to his or her social network[99], even if recruitment is done anonymously, and cases may be reluctant about sharing details of their contact persons. However, these ethical and legal challenges are not new and have always been important issues in source and contact tracing. In [100], we advocate for an ethical framework that offers normative guidance on the use of online technologies for case finding (**Chapter 6**). The ethical and legal boundaries differ between the aims of RDD, i.e., whether it is used for care, or whether it used for research. For research, the privacy law has strengthened, and it becomes a problem to collect, analyse and report on network data.

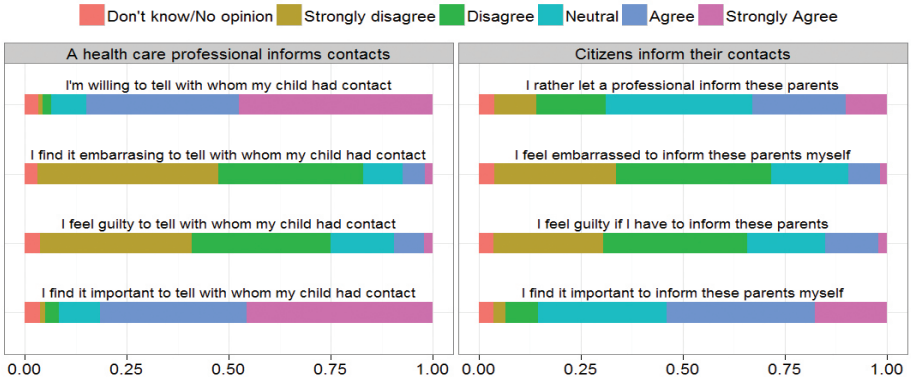


Figure 2. Willingness to cooperate during contact tracing. A representative sample of the general Dutch population (n = 648) were asked during a large measles outbreak about their willingness to help inform contacts of their child, in case their child was diagnosed with measles. In the first situation (left bars), participants were asked to indicate how they would feel if a health care worker asked to share details about their children's contact. In the second situation (right bars), we asked participants what they would feel if they were asked to inform the parents of their children's contacts themselves, e.g., via e-mail or telephone.

Determinants of online peer recruitment

The determinants that drive the online recruitment process have been investigated to a limited extent for hidden populations. For the general population, no data on drivers are available yet, and especially in the situation, where participants are cases of infection, it is unknown whether cases are more motivated or reluctant to recruit other cases. More empirical data is needed to understand the reasons why people choose to invite others, why they choose to invite specific contact persons, which recruitment options are most convenient for them and why, and why their contact persons decide to participate or not. For example, would a larger proportion of participants invite their contact persons if a questionnaire takes less time to fill in? Or, would more participants invite their contact persons if they are asked to send out invitations before answering the questionnaire? In our studies, we observed that reporting any symptoms and having certain socio-demographic characteristics (e.g., being a female and well educated) increases the probability to recruit contact persons. Also, certain recruitment options (e.g., sending an invitation via email or Facebook messenger) are preferred above others, although this differs per country[71, 96].

To further increase our understanding of online peer recruitment, we modelled the recruitment process to analyse the influence of determinants (such as the number of invitations sent out) on sample size, sample composition and shape of network trees (**Chapter 7**). During our data collection[86, 96], we observed a bimodal distribution for the number of invitations sent by participants, with a peak at zero and a peak at four (the maximum available number of coupons). This distribution is the result of asking participants to invite four contact persons.

However, simulations suggested that by motivating participants to invite any number between one and four is more effective in reaching successful peer recruitment[101]. A simulation model also enables to analyse the relationship between mean and variance of number of invitations sent by participants, which allows to identify the circumstances determining optimal peer recruitment[102]. Combining a simulation model with empirical data can provide useful input for future RDD studies in various target populations, e.g., on the required mean number of successfully sent invitations to reach large recruitment trees, a certain sample composition or a certain number of waves[101, 103].

Technical details such as survey design and user friendliness of a website, are undoubtedly important factors for a successful online recruitment. Barriers may already occur when recipients of invitations do not recognize the sender. The invitation is ignored, or even worse, the email invitation goes straight to the SPAM folder of the recipient. If recipients do read the invitation email, it is crucial to redirect them to a survey website. At this point, potential participants may be lost due to problems with opening the website, either because their email provider blocks them from opening websites directly from an email or because the survey website is not compatible with the device they are using. All these factors reduce response rates in online surveys.

Invitations by standard email seem less convenient with the arrival of social media and messenger applications, and less suitable to reach seeds and contact persons. It can therefore be worthwhile to facilitate and promote participants to use other (direct and private) recruitment options than email, e.g., recruitment via applications such as Facebook messenger and WhatsApp[104]. Furthermore, to prevent the recruitment of contact persons who can solely be reached via the Internet, researchers could try to offer participants an offline recruitment option, e.g., via printable coupons or SMS[105]. We emphasize that we only motivate the use of popular social media and mobile phone applications as a way for sending invitations (containing unique identifiers) to contact persons. For privacy reasons, the data collection with RDD must be done via a secure website of the institution conducting the study.

Section VI. Future applications for public health interventions

During outbreaks of certain pathogens, public health professionals carry out special investigations and interventions, such as screening of potential cases, contact tracing, and the active gathering of information through intensified communication with clinics, hospitals, and laboratories. These actions are done to accomplish early case detection in order to minimize health consequences and to prevent further transmission[3].

Contact tracing in particular is an essential element in the control of outbreaks of virulent or new pathogens, for which no therapy exists and an infection can result in death. New identified contact persons are quarantined if asymptomatic and isolated if symptomatic. For example, for diseases like SARS[106], Ebola[5], and the Middle East respiratory syndrome (MERS) [107], tracing of contact persons, with subsequent quarantine and isolation, has proven to be an effective measure, especially when the number of cases is low[6]. However, it causes a heavy burden for public health professionals, as most work is done through outreach work (such as house visits). For certain pathogens it is also not clear which types of contact persons may represent a risk (e.g., hepatitis A virus and newly emerging pathogens)[108, 109]. Depending on the disease severity, the focus often moves to reporting severe cases or infections among specific (high risk) groups during the course of an epidemic to reduce work load. Such reduced reporting complicates the monitoring and control of disease outbreaks.

In the area of STIs and human immunodeficiency virus (HIV), additional barriers such as stigmatization and privacy concerns may be involved that hinder traditional contact tracing by health care workers[110]. In the area of HIV, contact tracing is difficult because transmission may have occurred years ago and it may have involved anonymous contacts. For many STIs, e.g., HIV and chlamydia, most new infections originate from asymptomatic individuals who are unaware of their infection. This stresses the need to identify cases as soon as possible after infection occurred[111].

To improve timely notification and prevent further transmission, the US Centers for Disease Control and Prevention introduced voluntary partner services for STIs[112, 113]. With this approach cases are assisted with notifying their sexual partners of their exposure to STIs or HIV. With patient referral, patients inform their sexual partners. Notified partners can then choose to get tested, and in case of infection, receive medical treatment and prevention services (e.g., risk-reduction counselling). This can also be done confidentially, by means of provider referral. Here a case lets a public health counsellor inform their partner, without mentioning their names or the moment of exposure. Notifying partners is mostly done face-to-face or via telephone, but recently also more and more via the Internet. Few studies experimented with web-based notification via personal or anonymous emails, electronic postcards and text messages[114-119]. This provided limited information on the uptake of these web-based methods, e.g., it is still unclear what proportion of partners notified online actually seek testing and treatment[105].

Partner notification approaches are sometimes combined with home-based testing. This can improve case ascertainment and engage more people in disease testing[120-122]. In a randomised controlled trial conducted by Ostergaard and colleagues, notified partners preferred home-based test kits above regular testing by health care providers[121]. Unfortunately, three studies observed a poor uptake of home based test kits among sexual partners[105, 118, 123].

An online social network approach, like RDD, may further increase the detection of cases who may not be reached when solely contact persons of cases are tracked. Peers influence each other's attitude and actions, not only in general, but also more specific with respect to sexual behaviour and STI testing. In practice, it may therefore be useful to target also non-sexual relationships to identify individuals who are at increased risk for having a STI[124]. Based on this principle, a number of studies empirically explored peer-driven strategies that included the use of an individual's social network to find new cases and notify social contact persons at risk for infection. These studies had, however, mixed success with peer recruitment and consequently with finding new cases[105, 123, 125-128].

For outbreaks of pathogens that are transmitted through person-to-person contact, online RDD in combination with home-based sampling may be a good alternative for disease outbreaks whereby traditional contact tracing fails to track specific cases or contact persons. This may be the case, because the social network is difficult to reach for health care professionals, such as in student networks, or because the pathogen may spread by indirect contact (e.g., via contaminated objects[2]). Online RDD may be a useful addition to traditional contact tracing for diseases where the likelihood of incidentally missing infected cases is acceptable. Diseases like SARS, MERS and Ebola are potentially deadly, and therefore public health authorities are reluctant to rely on public responsibility to find all cases. Nevertheless, during outbreaks of these diseases, RDD may still be used as addition to other approaches. In certain situations, such as disease transmission within an airplane or other public transport[129], RDD is not feasible, because most individuals are not able to inform their incidental contact persons themselves.

Future steps

We are currently investigating whether RDD can be a useful method to support public health professionals with contact tracing during outbreaks of diseases that spread via direct contact, such as mumps, pertussis, and STIs. We are developing an online RDD tool for public health professionals, which facilitates online contact tracing via the contact network of cases, by the cases themselves. Such a tool will not only assist public health professionals in overcoming practical challenges, but may also enhance case detection and the tracking of the geographical spread of pathogens.

As part of another research project in the area of risk communication, we are investigating whether a respondent-driven method can be used to investigate clustering of risk perceptions and behaviour, and to spread educational messages through social networks. Such peer-driven network interventions use social networks to accelerate behaviour change or improve adherence to interventions among peers[130]. Theoretically it has been demonstrated that individual adoption is much more likely when participants receive social reinforcement from

multiple neighbours in their social network[131]. A respondent-driven method may be used to increase treatment uptake of individuals in a network[132], or for spreading information about a new screening or vaccination program to allow individuals to make a well-informed decision about their participation. In practice, intervention mapping approaches have been used to target networks of high-risk young people for Chlamydia trachomatis testing[117]. In a similar way, RDD may be used to spread educational messages during outbreaks of pathogens that are transmitted in other ways than STIs. Via such peer-driven messages, contact persons of cases can be warned quickly about their exposure to infection. This allows them to get early testing and treatment, or to change their behaviour to prevent further transmission.

Conclusions

A society where individuals are increasingly connected via the Internet provides opportunities for innovative methods to further improve public health infectious disease control. In this paper, we introduced online respondent-driven detection for case finding, by discussing results and lessons learned during our research studies. This method takes advantage of contact and social networks by utilizing the relations between individuals to increase case detection, the same contacts that pathogens use to spread through the population. Online RDD has unlimited geographical coverage and can provide timely, detailed information about individuals and their contact persons. Timely finding of cases is in particular important for treating cases and effective interventions to prevent further pathogen transmission. Using RDD also further increases our epidemiological knowledge on contact networks and the spread of infectious diseases within these networks. Such information is useful for informing mathematical models that rely on network data and help to predict the impact of control measures. An important challenge remains in how to motivate participants to invite contact persons, and how to convince invited contact persons to participate.

REFERENCES

1. Kramer A, Kretzschmar M, Krickeberg K (Eds.): Modern Infectious Disease Epidemiology: Concepts, Methods, Mathematical Models, and Public Health. New York: Springer; 2010.

2. Porta M: A dictionary of epidemiology. 6th edn. Oxford: Oxford University Press; 2014.

3. Reintjes R, Krickeberg K: Epidemiologic Surveillance. In Modern Infectious Disease Epidemiology Concepts, Methods, Mathematical Models, and Public Health. Edited by Krämer A, Kretzschmar M, Krickeberg K. LLC: Springer Science+Business Media; 2010: 143-158

4. Paquet C, Coulombier D, Kaiser R, Ciotti M: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill 2006, 11:212-214.

5. Greiner AL, Angelo KM, McCollum AM, Mirkovic K, Arthur R, Angulo FJ: Addressing contact tracing challenges-critical to halting Ebola virus disease transmission. Int J Infect Dis 2015, 41:53-55.

6. Eames KT, Keeling MJ: Contact tracing and disease control. Proc Biol Sci 2003, 270:2565-2571.

7. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, Eames KT, Edmunds WJ, Frost SD, Funk S, et al: Modeling infectious disease dynamics in the complex landscape of global health. Science 2015, 347:aaa4339.

8. Reichler MR, Reves R, Bur S, Thompson V, Mangura BT, Ford J, Valway SE, Onorato IM, Contact Investigation Study G: Evaluation of investigations conducted to detect and prevent transmission of tuberculosis. JAMA 2002, 287:991-995.

9. Dixon MG, Taylor MM, Dee J, Hakim A, Cantey P, Lim T, Bah H, Camara SM, Ndongmo CB, Togba M, et al: Contact Tracing Activities during the Ebola Virus Disease Epidemic in Kindia and Faranah, Guinea, 2014. Emerg Infect Dis 2015, 21:2022-2028.

10. Garnett GP, Anderson RM: Contact tracing and the estimation of sexual mixing patterns: the epidemiology of gonococcal infections. Sex Transm Dis 1993, 20:181-191.

11. Jajosky RA, Groseclose SL: Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health 2004, 4:29.

12. Yoo HS, Park O, Park HK, Lee EG, Jeong EK, Lee JK, Cho SI: Timeliness of national notifiable diseases surveillance system in Korea: a cross-sectional study. BMC Public Health 2009, 9:93.

13. Reijn E, Swaan CM, Kretzschmar ME, van Steenbergen JE: Analysis of timeliness of infectious disease reporting in the Netherlands. BMC Public Health 2011, 11:409.

14. Bonacic Marinovic A, Swaan C, van Steenbergen J, Kretzschmar M: Quantifying reporting timeliness to improve outbreak control. Emerg Infect Dis 2015, 21:209-216.

15. Heymann DL: Control of Communicable Diseases Manual. 19 edn: American Public Health Association; 2008.

16. Rea E, Lafleche J, Stalker S, Guarda BK, Shapiro H, Johnson I, Bondy SJ, Upshur R, Russell ML, Eliasziw M: Duration and distance of exposure are important predictors of transmission among community contacts of Ontario SARS cases. Epidemiol Infect 2007, 135:914-921.

17. Musher DM: How contagious are common respiratory tract infections? N Engl J Med 2003, 348:1256-1266.

18. Bridges CB, Kuehnert MJ, Hall CB: Transmission of influenza: implications for control in health care settings. Clin Infect Dis 2003, 37:1094-1101.

19. Goldmann DA: Transmission of viral respiratory infections in the home. Pediatr Infect Dis J 2000, 19:S97-102.

20. Potterat JJ, Muth SQ, Rothenberg RB, Zimmerman HZ, Green DL, Taylor JE, Bonney MS, White HA: Sexual network structure as an indicator of epidemic phase. Sex Transm Infect 2002, 78:152-158.

21. Edmunds WJ, O'Callaghan CJ, Nokes DJ: Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proc Biol Sci 1997, 264:949-957.

22. Morris M, Zavisca J, Dean L: Social and sexual networks: their role in the spread of HIV/AIDS among young gay men. AIDS Educ Prev 1995, 7:24-35.

23. Kretzschmar M, Morris M: Measures of concurrency in networks and the spread of infectious disease. Math Biosci 1996, 133:165-195.

24. Wendelboe AM, Hudgens MG, Poole C, Van Rie A: Estimating the role of casual contact from the community in transmission of Bordetella pertussis to young infants. Emerg Themes Epidemiol 2007, 4:15.

25. Smith KP, Christakis NA: Social Networks and Health. Annual Review of Sociology 2008, 34:405-429.

26. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA: Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect 2012, 140:2117-2130.

27. Wasserman S, Faust K: Social network analysis: Methods and applications. Cambridge: Cambridge University Press; 1994.

28. Eames KT: Networks of influence and infection: parental choices and childhood disease. J R Soc Interface 2009, 6:811-814.

29. Salathe M, Khandelwal S: Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011, 7:e1002199.

30. Travers J, Milgram S: An Experimental Study of the Small World Problem. Sociometry 1969, 32:425-443.

31. Newman MEJ: Networks: an introduction. New York: Oxford University Press; 2010.

32. Anderson RM, Fraser C, Ghani AC, Donnelly CA, Riley S, Ferguson NM, Leung GM, Lam TH, Hedley AJ: Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. Philos Trans R Soc Lond B Biol Sci 2004, 359:1091-1105.

33. Bansal S, Read J, Pourbohloul B, Meyers LA: The dynamic nature of contact networks in infectious disease epidemiology. J Biol Dyn 2010, 4:478-489.

34. Salathe M, Jones JH: Dynamics and control of diseases in networks with community structure. PLoS Comput Biol 2010, 6:e1000736.

35. Heckathorn DD: Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations Social Problems 2002, 49:11-34.

36. Salganik MJ, Heckathorn DD: Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology 2004.

37. Heckathorn D: Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems 1997, 44:174-199.

38. Wejnert C: An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and out-of-Equilibrium Data. Sociol Methodol 2009, 39:73-116.

39. Handcock MS, Gile KJ, Mar CM: Estimating hidden population size using Respondent-Driven Sampling data. Electron J Stat 2014, 8:1491-1521.

40. Tyldum G, Johnston L: Applying Respondent Driven Sampling to Migrant Populations: Lessons from the Field. Palgrave Macmillan UK; 2014.

41. International Telecommunication Union Measuring the information society 2015. 2015. http://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2015/MISR2015-w5.pdf. Accessed 2 september 2016.

42. Van Deursen AJAM, Van Dijk JAGM: Improving digital skills for the use of online public information and services. Government Information Quarterly 2009, 26:333–340.

43. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A: Digital epidemiology. PLoS Comput Biol 2012, 8:e1002616.

44. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L: Detecting influenza epidemics using search engine query data. Nature 2009, 457:1012-1014.

45. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L: Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013, 9:e1003256.

46. Davidson MW, Haim DA, Radin JM: Using networks to combine "big data" and traditional surveillance to improve influenza predictions. Scientific Reports 2014, 5:8154.

47. Ladbury G, Ostendorf S, Waegemaekers T, Hahné S: Liking" Social Networking Sites – Use of Facebook as a Recruitment Tool in an Outbreak Investigation, The Netherlands, 2012. Epidemiology: Open Access 2013, 3.

48. Lee C, Singal M: New Technologies for Partner Notifications for Sexually Transmitted Infections. In Evidence Review: National Collaborating Centre for Infectious Diseases; 2013.

49. Wojcik OP, Brownstein JS, Chunara R, Johansson MA: Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol 2014, 11:7.

50. Wejnert C, Heckathorn DD: Web-based network sampling: efficiency and efficacy of respondent-driven sampling for online research. Sociological Methods and Research 2008, 37:105-134.

51. Boyd DM, Ellison NB: Social network sites: definition, history, and scholarship. J Comput Commun 2007, 13:210-230.

52. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, Mendes GA, Lau AY: The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015, 22:243-256.

53. Bauermeister JA, Zimmerman MA, Johns MM, Glowacki P, Stoddard S, Volz E: Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (webRDS) strategy. J Stud Alcohol Drugs 2012, 73:834-838.

54. Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, Nguyen TA, Liljeros F, Thorson A: Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. PLoS One 2012, 7:e49417.

55. McCreesh N, Johnston LG, Copas A, Sonnenberg P, Seeley J, Hayes RJ, Frost SD, White RG: Evaluation of the role of location and distance in recruitment in respondent-driven sampling. Int J Health Geogr 2011, 10:56.

56. Johnston LG, Trummal A, Lohmus L, Ravalepik A: Efficacy of convenience sampling through the internet versus respondent driven sampling among males who have sex with males in Tallinn and Harju County, Estonia: challenges reaching a hidden population. AIDS Care 2009, 21:1195-1202.

57. Fenton KA, Korovessis C, Johnson AM, McCadden A, McManus S, Wellings K, Mercer CH, Carder C, Copas AJ, Nanchahal K, et al: Sexual behaviour in Britain: reported sexually transmitted infections and prevalent genital Chlamydia trachomatis infection. Lancet 2001, 358:1851-1854.

58. Garnett GP, Hughes JP, Anderson RM, Stoner BP, Aral SO, Whittington WL, Handsfield HH, Holmes KK: Sexual mixing patterns of patients attending sexually transmitted diseases clinics. Sex Transm Dis 1996, 23:248-257.

59. Gregson S, Nyamukapa CA, Garnett GP, Mason PR, Zhuwau T, Carael M, Chandiwana SK, Anderson RM: Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. Lancet 2002, 359:1896-1903.

60. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH: A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci U S A 2010, 107:22020-22025.

61. Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J: Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med 2011, 8:e1001083.

62. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J, et al: Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008, 5:e74.

63. Smieszek T, Barclay VC, Seeni I, Rainey JJ, Gao H, Uzicanin A, Salathe M: How should social mixing be measured: comparing web-based survey and sensor-based methods. BMC Infect Dis 2014, 14:136.

64. Read JM, Eames KT, Edmunds WJ: Dynamic social networks and the implications for the spread of infectious disease. J R Soc Interface 2008, 5:1001-1007.

65. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M: Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol Infect 2008, 136:813-822.

66. Bernard H, Fischer R, Mikolajczyk RT, Kretzschmar M, Wildner M: Nurses' contacts and potential for infectious disease transmission. Emerg Infect Dis 2009, 15:1438-1444.

67. Horby P, Pham QT, Hens N, Nguyen TT, Le QM, Dang DT, Nguyen ML, Nguyen TH, Alexander N, Edmunds WJ, et al: Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS One 2011, 6:e16965.

68. Kretzschmar M, Mikolajczyk RT: Contact profiles in eight European countries and implications for modelling the spread of airborne infectious diseases. PLoS One 2009, 4:e5931.

69. Christley RM, Pinchbeck GL, Bowers RG, Clancy D, French NP, Bennett R, Turner J: Infection in social networks: using network analysis to identify high-risk individuals. Am J Epidemiol 2005, 162:1024-1031.

70. Salathe M, Bonhoeffer S: The effect of opinion clustering on disease outbreaks. J R Soc Interface 2008, 5:1505-1508.

71. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Chanyasanha C, Tipayamongkholgul M, Thorson AE, Bengtsson L, Lu X, Kretzschmar ME: Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and The Netherlands using respondent-driven sampling. PLoS One 2014, 9:e113711.

72. Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PG, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar ME: Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PLoS One 2014, 9:e85256.

73. Adamic LA, Lukose RM, Puniyani AR, Huberman BA: Search in power-law networks. Phys Rev E Stat Nonlin Soft Matter Phys 2001, 64:046135.

74. Newman ME: Assortative mixing in networks. Phys Rev Lett 2002, 89:208701.

75. Anderson RM, May RM: Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press; 1991.

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

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

78. Rohani P, Zhong X, King AA: Contact network structure explains the changing epidemiology of pertussis. Science 2010, 330:982-985.

79. Eames KT, Tilston NL, Brooks-Pollock E, Edmunds WJ: Measured dynamic social contact patterns explain the spread of H1N1v influenza. PLoS Comput Biol 2012, 8:e1002425.

80. Zhou K, Lokate M, Deurenberg RH, Tepper M, Arends JP, Raangs EG, Lo-Ten-Foe J, Grundmann H, Rossen JW, Friedrich AW: Use of whole-genome sequencing to trace, control and characterize the regional expansion of extended-spectrum beta-lactamase producing ST15 Klebsiella pneumoniae. Sci Rep 2016, 6:20840.

81. Ypma RJ, Bataille AM, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM: Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc Biol Sci 2012, 279:444-450.

82. van der Kuyl AC, Jurriaans S, Back NK, Sprenger HG, van der Werf TS, Zorgdrager F, Berkhout B, Cornelissen M: Unusual cluster of HIV type 1 dual infections in Groningen, The Netherlands. AIDS Res Hum Retroviruses 2011, 27:429-433.

83. Villasenor-Sierra A, Quinonez-Alvarado MG, Caballero-Hoyos JR: Interpersonal relationships and group A streptococcus spread in a Mexican day-care center. Salud Publica Mex 2007, 49:323-329.

84. Stein RA: Super-spreaders in infectious diseases. Int J Infect Dis 2011, 15:e510-513.

85. Bates SJ, Trostle J, Cevallos WT, Hubbard A, Eisenberg JN: Relating diarrheal disease to social networks and the geographic configuration of communities in rural Ecuador. Am J Epidemiol 2007, 166:1088-1095.

86. Stein ML, van der Heijden PG, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, Thorson A, Kretzschmar ME: Tracking social contact networks with online respondent-driven detection: who recruits whom? BMC Infect Dis 2015, 15:522.

87. Barclay VC, Smieszek T, He J, Cao G, Rainey JJ, Gao H, Uzicanin A, Salathe M: Positive network assortativity of influenza vaccination at a high school: implications for outbreak risk and herd immunity. PLoS One 2014, 9:e87042.

88. Teutsch SM, Thacker SB: Planning a public health surveillance system. Epidemiol Bull 1995, 16:1-6.

89. Declich S, Carter AO: Public health surveillance: historical origins, methods and evaluation. Bull World Health Organ 1994, 72:285-304.

90. Nsubuga P, White ME, Thacker SB, Anderson MA, Blount SB, Broome CV, Chiller TM, Espitia V, Imtiaz R, Sosin D, et al: Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In Disease Control Priorities in Developing Countries. 2nd edition. Edited by Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A, Musgrove P. Washington (DC); 2006

91. Henning KJ: What is syndromic surveillance? MMWR Morb Mortal Wkly Rep 2004, 53 Suppl:5-11.

92. Fourquet F, Drucker J: Communicable disease surveillance: the sentinel network. Lancet 1997, 349:794-795.

93. McDonald SA, Presanis AM, De Angelis D, van der Hoek W, Hooiveld M, Donker G, Kretzschmar ME: An evidence synthesis approach to estimating the incidence of seasonal influenza in the Netherlands. Influenza Other Respir Viruses 2014, 8:33-41.

94. Friesema IH, Koppeschaar CE, Donker GA, Dijkstra F, van Noort SP, Smallenburg R, van der Hoek W, van der Sande MA: Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. Vaccine 2009, 27:6353-6357.

95. Eames KT, Brooks-Pollock E, Paolotti D, Perosa M, Gioannini C, Edmunds WJ: Rapid assessment of influenza vaccine effectiveness: analysis of an internet-based cohort. Epidemiol Infect 2012, 140:1309-1315.

96. Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Koppeschaar CE, Bengtsson L, Thorson A, Kretzschmar ME: Enhancing Syndromic Surveillance With Online Respondent-Driven Detection. Am J Public Health 2015, 105:e90-97.

97. Lehmann BA, Ruiter RA, Kok G: A qualitative study of the coverage of influenza vaccination on Dutch news sites and social media websites. BMC Public

Health 2013, 13:547.

98. Van der Schoor AS, Beaujean DJMA, Wong A, Timen A: Public perception, knowledge and behaviour during a measles outbreak in the Netherlands in 2013–2014. In preparation 2016.

99. Mandeville KT, Harris M, Thomas HL, Chow Y, Seng C: Using social networking sites for communicable disease control: innovative contact tracing or breach of confidentiality? Public Health Ethics 2013.

100. Stein ML, Rump BO, Kretzschmar MEE, Steenbergen JE: Social Networking Sites as a Tool for Contact Tracing: Urge for Ethical Framework for Normative Guidance. Public Health Ethics 2014, 7:57-60.

101. Stein ML, Buskens V, Van der Heijden PGM, Van Steenbergen J, Kretzschmar MEE: Drivers of respondent-driven detection. In preparation 2016.

102. Malmros J, Liljeros F, Britton T: Respondent-driven sampling and an unusual epidemic. 2014.

103. Crawford FW: The graphical structure of respondent-driven sampling. Sociological Methodology 2016:1-25.

104. Thornton L, Batterman PJ, Fassnacht DB, Kay-Lambkin F, Calear AL, Hunt S: Recruiting for health, medical or psychosocial research using Facebook: systematic review. Internet Interventions 2016, 4:72-81.

105. Theunissen K, Hoebe C, Kok G, Crutzen R, Kara-Zaitri C, de Vries N, van Bergen J, Hamilton R, van der Sande M, Dukers-Muijrers N: A Web-Based Respondent Driven Sampling Pilot Targeting Young People at Risk for Chlamydia Trachomatis in Social and Sexual Networks with Testing: A Use Evaluation. Int J Environ Res Public Health 2015, 12:9889-9906.

106. Severe acute respiratory syndrome--Singapore, 2003. MMWR Morb Mortal Wkly Rep 2003, 52:405-411.

107. Mollers M, Jonges M, Pas SD, van der Eijk AA, Dirksen K, al. e: Follow-up of Contacts of Middle East Respiratory Syndrome Coronavirus–Infected Returning Travelers, the Netherlands, 2014. Emerging Infectious Diseases 2015, 21.

108. Eames KT, Webb C, Thomas K, Smith J, Salmon R, Temple JM: Assessing the role of contact tracing in a suspected H7N2 influenza A outbreak in humans in Wales. BMC Infect Dis 2010, 10:141.

109. Parry-Ford F, Boddington N, Pebody R, Phin N, Incident Management T: Public health response to two incidents of confirmed MERS-CoV cases travelling on flights through London Heathrow Airport in 2014 - lessons learnt. Euro Surveill 2015, 20.

110. Theunissen KA, Bos AE, Hoebe CJ, Kok G, Vluggen S, Crutzen R, Dukers-Muijrers NH: Chlamydia trachomatis testing among young people: what is the role of stigma? BMC Public Health 2015, 15:651.

111. Marks G, Crepaz N, Janssen RS: Estimating sexual transmission of HIV from persons aware and unaware that they are infected with the virus in the USA. AIDS 2006, 20:1447-1450.

112. Center for Disease Control and Prevention (CDC) HIV, Hepatitis, STD and TB Partners. 2014. http://www.cdc.gov/nchhstp/partners/Partner-Services.html. Accessed 5 May 2014.

113. Center for Disease Control and Prevention: Use of Social Networks to Identify Persons with Undiagnosed HIV Infection. Seven US cities, October 2003-September 2004. MMWR 2005, 54:601-605.

114. Bilardi JE, Fairley CK, Hopkins CA, Hocking JS, Temple-Smith MJ, Bowden FJ, Russell DB, Pitts M, Tomnay JE, Parker RM, et al: Experiences and outcomes of partner notification among men and women recently diagnosed with Chlamydia and their views on innovative resources aimed at improving notification rates. Sex Transm Dis 2010, 37:253-258.

115. Levine D, Woodruff AJ, Mocello AR, Lebrija J, Klausner JD: inSPOT: the first online STD partner notification system using electronic postcards. PLoS Med 2008, 5:e213.

116. Gotz HM, van Rooijen MS, Vriens P, Op de Coul E, Hamers M, Heijman T, van den Heuvel F, Koekenbier R, van Leeuwen AP, Voeten HA: Initial evaluation of use of an online partner notification tool for STI, called 'suggest a test': a cross sectional pilot study. Sex Transm Infect 2014, 90:195-200.

117. Theunissen KA, Hoebe CJ, Crutzen R, Kara-Zaitri C, de Vries NK, van Bergen JE, van der Sande MA, Dukers-Muijrers NH: Using intervention mapping for the development of a targeted secure web-based outreach strategy named SafeFriend, for Chlamydia trachomatis testing in young people at risk. BMC Public Health 2013, 13:996.

118. Dukers-Muijrers NH, Theunissen KA, Wolffs PT, Kok G, Hoebe CJ: Acceptance of Home-Based Chlamydia Genital and Anorectal Testing Using Short Message Service (SMS) in Previously Tested Young People and Their Social and Sexual Networks. PLoS One 2015, 10:e0133575.

119. Mimiaga MJ, Tetu AM, Gortmaker S, Koenen KC, Fair AD, Novak DS, Vanderwarker R, Bertrand T, Adelson S, Mayer KH: HIV and STD status among MSM and attitudes about Internet partner notification for STD exposure. Sex Transm Dis 2008, 35:111-116.

120. Andersen B, Ostergaard L, Moller JK, Olesen F: Home sampling versus conventional contact tracing for detecting Chlamydia trachomatis infection in male partners of infected women: randomised study. BMJ 1998, 316:350-351.

121. Ostergaard L, Andersen B, Moller JK, Olesen F, Worm AM: Managing partners of people diagnosed with Chlamydia trachomatis: a comparison of two partner testing methods. Sex Transm Infect 2003, 79:358-361.

122. Ostergaard L, Moller JK, Andersen B, Olesen F: Diagnosis of urogenital Chlamydia trachomatis infection in women based on mailed samples obtained at home: multipractice comparative study. BMJ 1996, 313:1186-1189.

123. Rose SB, Lawton BA, Bromhead C, MacDonald EJ, Elley CR: Poor uptake of self-sample collection kits for Chlamydia testing outside primary care. Aust N Z J Public Health 2010, 34:517-520.

124. Youm Y, Laumann EO: Social network effects on the transmission of sexually transmitted diseases. Sex Transm Dis 2002, 29:689-697.

125. Rothenberg R, Kimbrough L, Lewis-Hardy R, Heath B, Williams OC, Tambe P, Johnson D, Schrader M: Social network methods for endemic foci of syphilis: a pilot project. Sex Transm Dis 2000, 27:12-18.

126. Rothenberg RB, Sterk C, Toomey KE, Potterat JJ, Johnson D, Schrader M, Hatch S: Using social network and ethnographic tools to evaluate syphilis transmission. Sex Transm Dis 1998, 25:154-160.

127. Loaring J, Hickman M, Oliver I, Campbell R, Trotter C, Macleod J, Pye K, Crichton J, Horner P: Could a peer-led intervention increase uptake of chlamydia screening? A proof of principle pilot study. J Fam Plann Reprod Health Care 2013, 39:21-28.

128. Rosenberg NE, Kamanga G, Pettifor AE, Bonongwe N, Mapanje C, Rutstein SE, Ward M, Hoffman IF, Martinson F, Miller WC: STI patients are effective recruiters of undiagnosed cases of HIV: results of a social contact recruitment study in Malawi. J Acquir Immune Defic Syndr 2014, 65:e162-169.

129. Swaan CM, Appels R, Kretzschmar ME, van Steenbergen JE: Timeliness of contact tracing among flight passengers for influenza A/H1N1 2009. BMC Infect Dis 2011, 11:355.

130. Valente TW: Network interventions. Science 2012, 337:49-53.

131. Centola D: The spread of behavior in an online social network experiment. Science 2010, 329:1194-1197.

132. Deering KN, Shannon K, Sinclair H, Parsad D, Gilbert E, Tyndall MW: Piloting a peer-driven intervention model to increase access and adherence to antiretroviral therapy and HIV care among street-entrenched HIV-positive women in Vancouver. AIDS Patient Care STDS 2009, 23:603-609.

Summary

A broad range of infectious diseases such as influenza, measles, Ebola, and severe acute respiratory syndrome (SARS) are transmitted through direct or close human contact, e.g., shaking hands or a face-to-face conversation. Therefore, these pathogens do not spread randomly through a population, but follow the structure of human contact networks. This entails that infected persons (cases) tend to cluster by time and space, and that their contact persons are at a higher risk for infection. Making use of these networks may help to understand and control the spread of infectious diseases. This thesis presents multidisciplinary research studies in which the use of online respondent-driven detection (RDD), a network-based method via the Internet, was piloted for sampling contact networks of individuals in the general population. The aims were to investigate the feasibility of using the contact network of cases to analyse transmission dynamics within these networks (**part I**) and to enhance case finding during outbreaks of emerging or re-emerging pathogens (**part II**). We also investigated factors driving the recruitment process (**part III**), and highlighted prerequisites for implementation and outlined future applications of RDD for public health interventions (**part IV**).

Part I Studying contact networks using an online respondent-driven method

Understanding pathogen transmission through a population requires empirical data on contact patterns. Such information can inform mathematical models that provide evidence-based support for outbreak preparedness and interventions. The majority of studies that investigated contact patterns relevant for the spread of close contact pathogens used an egocentric design, i.e., participants were sampled independently of one another. These studies provided no information on the network structure beyond the contact persons reported by participants. Therefore, we aimed at developing a method to recruit individuals and their contact persons into a survey in order to analyse links between contact persons of contact persons within the same network.

In **chapters 2 and 3**, we piloted the feasibility of online RDD to study contact networks. We developed a software system for facilitating and tracking recruitment by Facebook and email. One-day diary surveys were conducted by applying online RDD among a convenience sample of university students in Thailand and the Netherlands. Participants were asked to record numbers of contact persons at different settings and self-reported influenza-like-illness (ILI) symptoms, and to invite four contact persons whom they had met in the preceding week. We analysed correlations between recruiters and recruitees to investigate mixing patterns. A total of 358 individuals participated in the Netherlands and 257 in Thailand. Seeds (i.e., the first recruiters in a recruitment tree) in the Netherlands were more successful in inviting recruitees who also completed the questionnaire. Nevertheless, in Thailand, we reached up to six waves of recruitees, compared to five waves in the Netherlands. There were 233 pairs of

recruiter-recruitee in the Netherlands and 140 in Thailand. In both countries, more than half of all participants invited contact persons. Thai participants preferred Facebook for inviting recruitees, while Dutch participants mainly sent invitations by email. Despite strong socio-cultural differences between both countries, we observed assortative mixing patterns by demographic variables and random mixing by numbers of contact persons in both samples. This means that participants invited contact persons of similar age, sex and educational level, but they did not preferably invite persons who also reported similar numbers of contact persons. Both pilot studies provided new insights on number of contact persons and mixing patterns relevant for pathogens that are transmitted via close contact. However, our studies were limited by small sample sizes and surveys remained mainly within age groups.

Based on experience gained with surveys conducted in the Netherlands and Thailand, we combined online RDD with two large participatory syndromic surveillance panels in the Netherlands and Dutch speaking Flanders (Belgium). These Internet-based systems capture voluntarily submitted data on ILI symptoms during the winter season from the general public. In **chapter 4**, volunteers of these panels were invited to a RDD survey and asked to recruit four contact persons whom they had met face to face in the preceding two weeks. In total 1,560 individuals completed the survey, who reported in total 30,591 contact persons. A large number of recruiter-recruitee pairs (n=488) were sampled, consisting of individuals of different ages and backgrounds. Recruitment was assortative (i.e., more recruitment of individuals with similar characteristics) by age, education, household size, influenza vaccination status and sentiments, indicating that participants tended to recruit contact persons similar to themselves with respect to those demographics and sentiments. Furthermore, we observed assortative recruitment by self-reported symptoms, reaffirming the RDD objective of sampling contact persons whom a participant may infect or by whom a participant may get infected in general or in case of an outbreak. Recruitment was random by sex and, again, by numbers of contact persons (i.e., there was no tendency of participants to recruit contact persons of the same sex or with similar numbers of contact persons). Correlations between pairs were influenced by the geographical distance between two persons. Recruitment of persons with the same postal code was more assortative by demographic characteristics, compared to recruitment of persons who lived further away. This means that the geographical distance between a recruiter and a recruitee determines the type of contact networks being sampled with online RDD. A comparison of our 'who-recruits-whom' matrix stratified by age showed similar structures in patterns as the 'who-has-contact-with-whom' matrix stratified by age reported earlier during a large egocentric study. This indicates that, although complex mechanisms influence online peer recruitment, we may conclude that the observed statistical correlations reflect the observed contact network patterns in the general population.

Part II Identifying cases with online respondent driven detection

Monitoring the spread, preparing for and responding to infectious disease outbreaks are key public health priorities. The finding of infectious cases is essential in these three crucial elements of infectious disease control, which aims to prevent further spread in the population and individual health consequences. However, data collected with case finding is for practically all infectious diseases incomplete and biased. For many infectious diseases, an infection can occur without symptoms, but asymptomatic cases can still be infectious to others. For some infections that lead to symptoms, the majority of cases only experience mild symptoms and do not seek health care and are therefore not identified, but still are infectious and can spread the agent. To find these cases is laborious and often challenging. It is one of the most important workloads of public health professionals as most work is done manually, e.g., by telephone. Therefore, we aimed at developing an online respondent-driven method that uses the contact network of cases to detect other cases.

In **chapter 5**, we used the same dataset that we collected with the participatory surveillance panels in **chapter 4** and analysed self-reported symptoms of individuals linked by recruitment chains. We observed that seeds reporting symptoms recruited more frequently contact persons who also reported symptoms than asymptomatic seeds. In other words, symptomatic cases recruited other symptomatic cases. Furthermore, symptomatic participants mostly reported observing similar symptoms among their close contact persons. Participants who reported symptoms were also more likely to invite others, than participants without symptoms. Our results suggested that combining online communities with RDD might enhance identification of hitherto 'hidden' cases that go unnoticed by traditional surveillance systems, especially for infectious disease outbreaks in which the majority of symptomatic patients do not seek health care.

However, as during our previous studies, sample size of recruitment remained small. On average, less than half of all seeds invited a contact person, a proportion not sufficient to generate long recruitment chains. Participants expressed concerns about privacy and not wanting to bother acquaintances with a questionnaire, which withheld some participants from sending invitations to contact persons. These ethical and privacy challenges are not new and have always been important issues in traditional case finding. In **chapter 6**, we discuss the use of social networking sites for case finding, and advocate an ethical framework offering normative guidance on the use of innovative technologies in infectious disease control.

Part III Factors driving peer recruitment

Chapter 7 focuses on the factors important for the success of the chain recruitment process. We implemented a stochastic simulation model, where parameters were suggested by our empirically collected data, to determine the thresholds for obtaining large recruitment trees and the number of waves needed to reach a steady state in the sample composition for individual characteristics. When a steady state is reached, the sample composition remains the same in future waves and can be used to infer prevalence of characteristics in the population from which the sample was obtained.

Our main finding was that by motivating participants to send out any number of invitations between 1 and 4 is more effective in reaching large recruitment trees, than asking to send exactly four. In the latter situation the majority of participants does not send anything and only a low proportion sends out all 4 invitations. In the observed data, the probability of invitation acceptance by invitees was low and large recruitment trees could not be obtained. We also investigated the influence of difference in mixing behaviour on recruitment. Our empirical data suggest that there are only a few active seeds. With assortative mixing behaviour (also seen in our empirical data), and by starting with active seeds, large recruitment trees are reached faster, within a lower number of waves, compared to a hypothetical situation with random mixing (i.e., no preference of recruiters to invite certain contact persons). However, assortative mixing (e.g., females only invite other females) led to samples with an overrepresentation of participants with specific characteristics, and a steady state is reached slower, compared to random mixing. Our simulation model is a helpful tool that can assist researchers or public health professionals with preparing research or contact tracing using RDD. For example, the model can provide input on the required number of active seeds and mean number of successfully sent invitations to reach large recruitment trees, a certain sample composition or a certain number of waves.

Part IV In perspective

In this thesis, we introduced online RDD for studying infectious disease transmission and case finding. In **chapter 8,** we discussed results and lessons learned during our empirical studies, implications and remaining challenges. RDD takes advantage of networks by utilizing the relations between individuals to identify other cases, the same contacts that pathogens use to spread through the population. Online RDD has large geographical coverage and can provide timely, detailed information about individuals and their contact persons. Timely finding of cases is in particular important for treating cases and effective intervention to prevent further pathogen transmission. Using RDD also further increases our epidemiological knowledge on contact networks and the transmission of pathogens within these networks. Such knowledge

is useful for informing mathematical models that rely on network data and help to predict the impact of interventions.

However, RDD has important limitations concerning peer recruitment. It remains challenging to motivate individuals to invite contact persons and to convince invited contact persons to participate. These challenges are relevant both for studying contact networks and for case finding. Although many Internet-users share information with each other via social media, sending an invitation for a questionnaire specifically to a few contact persons is a step that many of our participants were not willing to take. Another limitation is that it is difficult with RDD to determine which contact persons participants invite from their 'pool of contact persons', which invited contact persons complete the questionnaire, and importantly, which contact persons decide not to participate. In our studies, we observed a strong tendency of participants to recruit contact persons with similar traits; this is a requirement for case finding, but biases peer recruitment when it comes to using RDD for studying contact networks. More empirical research is needed to understand the reasons why people choose to invite others, why they choose to invite specific contact persons, which recruitment options are most convenient for them and why or why not their contact persons decide to participate.

We are currently investigating whether online RDD can be a useful method to support public health professionals with contact tracing during outbreaks of diseases that spread via contact networks, such as pertussis and sexually transmitted infections. Furthermore, we are exploring the use of online RDD to measure clustering of the same risk perceptions and behaviour in social networks, and to spread educational messages about a (new) screening or vaccination program to allow individuals to make a well-informed decision about their participation.

Nederlandse samenvatting (Dutch summary)

Veel infectieziekten zoals influenza, mazelen, ebola en SARS worden overgedragen via direct of nauw contact tussen individuen, bijvoorbeeld door het schudden van handen of in een fysiek gesprek. Deze infectieziekten verspreiden zich hierdoor niet willekeurig in een populatie, maar volgen de structuur van contactnetwerken. Dit leidt vaak tot clustering van geïnfecteerden in tijd en plaats. Contacten van een geïnfecteerde hebben daardoor ook een groter infectierisico. Inzicht in, en gebruik maken van deze netwerken kan helpen bij het beter begrijpen én bestrijden van de verspreiding van infectieziekten. Dit proefschrift beschrijft multidisciplinair onderzoek waarin de ontwikkeling en bruikbaarheid is getest van online respondentgestuurde detectie (RDD), een netwerkgebaseerde methode via het internet om steekproeven te nemen van contactnetwerken van individuen in de algemene populatie. Onze doelstellingen waren om te onderzoeken of contactnetwerken van geïnfecteerde individuen gebruikt kunnen worden voor het bestuderen van de verspreiding van infectieziekten binnen deze netwerken (**deel I**) en voor het verbeteren van de opsporing van geïnfecteerden tijdens uitbraken van bestaande en nieuwe infectieziekten (**deel II**). We onderzochten ook factoren die belangrijk zijn voor het rekruteringsproces (**deel III**), en belichtten de randvoorwaarden voor uitvoering en beschreven toekomstige toepassingen van RDD voor volksgezondheidsinterventies (**deel IV**).

Deel I Contactnetwerken bestuderen met een online respondentgestuurde methode

Om een goed beeld te krijgen van de verspreiding van infectieziekten in een populatie is het nodig om te weten hoe contactpatronen er in werkelijkheid uit zien. Deze gegevens leveren belangrijke bouwstenen voor wiskundige modellen die gebruikt worden om het beleid ten aanzien van preventie- of interventiemaatregelen te onderbouwen. De meeste studies die contactpatronen onderzochten, die relevant zijn voor infectieziekten die via nauw contact verspreiden, gebruikten hierbij een 'egocentrische' onderzoeksopzet. Dit betekent dat deelnemers onafhankelijk van elkaar werden benaderd. Deze studies geven hierdoor geen informatie over de netwerkstructuur, afgezien van de contacten die door deelnemers worden gerapporteerd. Wij hadden daarom het doel om een methode te ontwikkelen waarmee zowel individuen als hun contacten meedoen met een vragenlijst, om vervolgens de onderlinge verbindingen tussen contacten in hetzelfde netwerk te kunnen analyseren.

In de onderzoeken van de **hoofdstukken 2 en 3** hebben we de haalbaarheid getest van online RDD voor het bestuderen van contactnetwerken. Allereerst hebben we een softwaresysteem ontwikkeld dat het mogelijk maakt dat mensen elkaar uitnodigen via Facebook en e-mail. Een beperkt aantal universiteitsstudenten in Thailand en Nederland is uitgenodigd voor een RDD survey. We vroegen de deelnemers naar het aantal contacten op één dag, op verschillende locaties en naar ervaren griepachtige symptomen. Aan het einde van de vragenlijst stond het verzoek om vier contacten uit te nodigen die men in de voorgaande week fysiek had ontmoet.

Vervolgens analyseerden we correlaties tussen rekruteerders en hun gerekruteerden om contactpatronen te onderzoeken. In totaal deden 358 individuen mee in Nederland en 257 in Thailand. 'Seeds' (dit zijn de eerste rekruteerders in een rekruteringsboom) in Nederland waren succesvoller in het rekruteren van contacten die de vragenlijst ook voltooiden, met andere woorden, de Nederlandse gerekruteerden deden vaker volledig mee. In Thailand bereikten we echter zes stappen in een keten van contacten, vergeleken met vijf stappen in Nederland. Er waren 233 paren van rekruteerder-contactpersoon in Nederland en 140 in Thailand. In beide landen nodigden meer dan de helft van alle deelnemers contacten uit. Thaise deelnemers gaven hierbij de voorkeur aan Facebook, terwijl Nederlandse deelnemers voornamelijk e-mail gebruikten. Ondanks dat beide landen aanzienlijk van elkaar verschillen op sociaal-cultureel niveau, was de opvallende overeenkomst dat deelnemers in beide landen individuen uitnodigden die vergelijkbaar met hen waren ten aanzien van bepaalde demografische kenmerken, zoals leeftijd, geslacht en opleidingsniveau. Een dergelijke overeenkomst was er niet voor het aantal contacten per persoon. Hier bleek geen patroon in te zitten: mensen met veel contacten nodigden niet vaker andere mensen uit die ook veel contacten hadden. Onze teststudies gaven nieuwe inzichten in het aantal contacten en contactpatronen die relevant zijn voor pathogenen die zich via direct of nauw contact verspreiden. Echter, onze studies berustten op kleine steekproeven en de vragenlijsten gingen voornamelijk rond in dezelfde leeftijdsgroepen.

De ervaringen van de studies in Nederland en Thailand pasten we vervolgens toe bij het combineren van online RDD met twee grote burgerpanels voor syndroomsurveillance van griepachtige symptomen. Deze panels zijn tijdens het griepseizoen actief in Nederland en in Nederlands sprekend Vlaanderen (België): de Grote Griepmeting en de Grote Longontstekingmeting. In **hoofdstuk 4** beschrijven we hoe we de vrijwilligers van deze panels uitnodigden voor een RDD-vragenlijst en hen vroegen om vier contacten uit te nodigen die ze fysiek hadden ontmoet in de voorgaande twee weken. In totaal hebben 1.560 deelnemers de vragenlijst volledig ingevuld, waarbij ze allen bij elkaar aangaven in totaal 30.591 contacten te hebben. Een groot aantal (rekruteerder-gerekruteerde-) paren (N=488) heeft uiteindelijk meegedaan, bestaande uit individuen met verschillende leeftijden en achtergronden. Rekruteerders nodigden mensen uit van vergelijkbare leeftijd, opleidingsniveau, gezinsgrootte, influenzavaccinatiestatus en perceptie van het vaccin. Deze overeenkomst rekruteerder-gerekruteerde vonden we ook voor ervaren griepachtige klachten. Dit bevestigde het doel van RDD, namelijk: met RDD bereiken we de contacten die geïnfecteerd zouden kunnen worden door een deelnemer of die de deelnemer zouden kunnen infecteren in het algemeen en in geval van een uitbraak. Vrouwen nodigden niet speciaal vrouwen uit, en mannen ook niet mannen. Ook de omvang van de contactenkring leidde niet tot het uitnodigen van mensen met eenzelfde grootte van de contactenkring. De correlaties tussen twee personen werden beïnvloed door de geografische afstand tussen deze personen. Uitgenodigde contacten die

op dezelfde postcode woonden hadden vaker dezelfde demografische kenmerken als de rekruteerder, vergeleken met contacten die verder weg woonden. Dit betekent dat de geografische afstand tussen een rekruteerder en een contact bepaalt welke soort contactnetwerk er wordt bereikt met online RDD. We vergeleken ook de 'wie-heeft-contact-met-wie'-matrix naar leeftijd, die eerder werd verzameld tijdens een grote egocentrische studie (gegevens verzameld van losse individuen), met onze 'wie-nodigt-wie-uit'-matrix naar leeftijd. Deze matrixen bleken sterk overeen te komen. Dit betekent dat we kunnen concluderen dat de geobserveerde statistische correlaties met deze nieuwe methode de geobserveerde contactpatronen in de algemene populatie goed weerspiegelen, ondanks de complexe mechanismen die een rol spelen bij het online rekruteringsproces.

Deel II Opsporen van geïnfecteerden met online respondentgestuurde detectie

Het monitoren van de verspreiding van infecties, en het voorbereiden en reageren op infectieziekte-uitbraken zijn basistaken voor de volksgezondheid. Het opsporen van geïnfecteerden die besmettelijk zijn voor anderen is daarbij een essentieel element, om zo verdere verspreiding in de populatie en individuele gezondheidsschade te beperken. De standaard werkwijze in de openbare gezondheidszorg is daarvoor bron- en contactonderzoek. Echter, de gegevens die met bron- en contactopsporing worden verzameld zijn vrijwel voor alle infectieziekten onvolledig en vertekenen de werkelijke situatie. Van veel verwekkers verloopt een infectie in een groot deel van de bevolking zonder symptomen, maar asymptomatische geïnfecteerden kunnen wèl (nog steeds) besmettelijk zijn voor anderen. Bovendien ervaren de meeste symptomatische geïnfecteerden (mensen mèt klachten) vaak alleen milde klachten en bezoeken daarvoor niet de huisarts, waardoor ze onzichtbaar blijven voor alle registratiesystemen. Bovendien is de gebruikelijke manier van bron- en contactopsporing moeilijk en arbeidsintensief voor het personeel in de openbare gezondheidszorg, omdat het meeste werk alleen handmatig kan, zoals via telefoongesprekken. We hadden daarom als doel om een online methode te ontwikkelen die het opsporen van geïnfecteerden via de contactnetwerken van burgers mogelijk maakt via de burgers zelf.

In **hoofdstuk 5** gebruikten we opnieuw de gegevens die we eerder verzamelden met de burgerpanels uit **hoofdstuk 4** en analyseerden de ervaren griepachtige klachten van individuen die elkaar uitnodigden. Uit deze analyse bleek dat 'seeds' met griepachtige klachten meer contacten rekruteerden met vergelijkbare klachten dan 'seeds' zonder klachten. Dit ging op voor algemene symptomen, verkoudheid en koorts. Met andere woorden, symptomatische geïnfecteerden nodigden andere symptomatische geïnfecteerden uit. Ook rapporteerden deelnemers met ziekteverschijnselen bij nauwe contacten (al dan niet uitgenodigd om deel te nemen) vergelijkbare verschijnselen. Deelnemers met symptomen waren ook actiever met het versturen van uitnodigingen naar contacten dan deelnemers zonder symptomen. Deze

resultaten suggereren dat het combineren van RDD met internetpanels de opsporing van geïnfecteerden kan verbeteren; met name mensen die tot dan toe verborgen bleven voor traditionele surveillancesystemen. In het bijzonder tijdens infectieziekte-uitbraken, waarbij vaak de meerderheid van de symptomatische patiënten geen gebruik maakt van de gezondheidszorg, is dit mogelijk een belangrijke aanvulling voor het krijgen van een goed beeld van de verspreiding en ziektelast van de uitbraak. Echter, net als bij onze voorgaande studies, nodigde maar een zeer klein gedeelte van de deelnemers ook anderen uit. Gemiddeld nodigde minder dan de helft van alle 'seeds' een contactpersoon uit. Dit percentage was te laag om lange kettingen van onderlinge rekrutering te verkrijgen. Deelnemers uitten hun bezorgdheid over privacy en vonden het onprettig kennissen met een vragenlijst lastig te vallen. Dit weerhield sommige deelnemers van het versturen van uitnodigingen. Deze kwesties omtrent ethiek en privacy zijn niet nieuw en zijn ook altijd (al) belangrijk bij traditionele bron- en contactopsporing. In **hoofdstuk 6** bediscussiëren we het gebruik van sociale media voor bron- en contactopsporing en pleiten we voor invoering van een ethisch kader, dat normatieve richtlijnen aangeeft voor het gebruik van dit soort innovatieve technieken in de infectieziektebestrijding.

Deel III Factoren relevant voor het rekruteringsproces

Hoofdstuk 7 focust zich op de factoren die belangrijk zijn voor het succes van het rekruteringsproces. We maakten een stochastisch simulatiemodel om de drempels te bepalen voor het verkrijgen van grote rekruteringsbomen en het aantal benodigde contactstappen om een stabiele toestand te bereiken in individuele kenmerken van gerekruteerden. Vanaf het moment dat er een stabiele toestand wordt bereikt voor een bepaald kenmerk, blijft de samenstelling van de steekproef bij toenemende contactstappen gelijk en kunnen er prevalentieschattingen worden gemaakt voor kenmerken van de populatie waaruit de steekproef werd getrokken. Voor de parameters in het model (zoals de kansen op het versturen en accepteren van uitnodigingen; wie-nodigt-wie-uit) gebruikten we de in werkelijkheid geobserveerde data.

Onze belangrijkste bevinding was dat het motiveren van deelnemers om tussen de één en de vier uitnodigingen te versturen effectiever is voor het bereiken van grote rekruteringsbomen, dan mensen te vragen exact vier mensen uit te nodigen, omdat we weten dat de meerderheid van de deelnemers dan geen enkele uitnodiging verstuurt en een laag percentage deelnemers alle vier de uitnodigingen. In de in werkelijkheid geobserveerde data was de kans op het accepteren van een uitnodiging door een contactpersoon te laag en was het niet mogelijk om grote rekruteringsbomen te verkrijgen. We onderzochten ook de invloed van verschillend rekruteringsgedrag (oftewel, wie-nodigt-wie-uit? En wat gebeurt er als deelnemers andere individuen uitnodigen). In situaties waar individuen gelijksoortige mensen uitnodigden en

waarbij de simulaties startten met 'seeds' die zeer actief waren met het versturen van uitnodigingen, ontstonden grote rekruteringsbomen sneller, binnen een lager aantal contactstappen, dan in een hypothetische situatie waar individuen willekeurig anderen uitnodigen. Omdat we in werkelijkheid ook zien dat mensen gelijksoortige mensen uitnodigen is het advies om vooral te proberen actieve rekruteerders (dit waren in ons geval hoogopgeleide vrouwen in de leeftijdsgroep 60 jaar en ouder) als eerste 'seeds' in te schakelen. Het nadeel is wel dat in situaties waarin deelnemers gelijksoortigen uitnodigden (bijvoorbeeld vrouwen nodigden andere vrouwen uit) leidt tot onderzoekspopulaties met een oververtegenwoordiging van deelnemers met deze specifieke kenmerken en bovendien duurt het langer voordat er een stabiele toestand wordt bereikt. Het simulatiemodel is een handig hulpmiddel voor ons en andere onderzoekers of personeel in de openbare gezondheidszorg bij het opzetten van online RDD-vragenlijsten voor onderzoek of bron- en contactopsporing. Het model kan bijvoorbeeld inzicht geven in het benodigde aantal actieve 'seeds' en het gemiddelde aantal succesvol verstuurde uitnodigingen, om grote rekruteringsbomen, een bepaalde steekproefgrootte of om een bepaalde samenstelling van de steekproef te bereiken.

Deel IV In perspectief

In dit proefschrift introduceren we online RDD voor het bestuderen van infectieziektetransmissie en de opsporing van geïnfecteerden. In **hoofdstuk 8** bespreken we de resultaten en leerervaringen van onze empirische studies, implicaties en de resterende uitdagingen. RDD maakt gebruikt van netwerken door de relaties tussen individuen te gebruiken bij de opsporing van geïnfecteerden, dezelfde verbindingen die infectieziekten gebruiken om zich te verspreiden in een populatie. Online RDD heeft een groot geografisch bereik en kan tijdige, gedetailleerde informatie over individuen en hun contacten opleveren. Een tijdige detectie van geïnfecteerden is vooral belangrijk om patiënten op tijd te kunnen behandelen en maakt effectievere interventies mogelijk om verdere verspreiding van infectieziekten te voorkomen. Het inzetten van RDD vergroot ook onze epidemiologische kennis over contactnetwerken en de verspreiding van infectieziekten binnen deze netwerken. Deze kennis is nodig voor het informeren van wiskundige modellen die afhankelijk zijn van netwerkgegevens en waarmee het effect van interventies kan worden voorspeld.

Echter, het motiveren van deelnemers om hun contacten uit te nodigen, en het overtuigen van deze contacten om mee te doen, is in al ons onderzoek een groot struikelblok gebleken en is daardoor de belangrijkste uitdaging voor toekomstige verdere ontwikkeling van RDD. Dit is niet alleen belangrijk voor de studie naar contactnetwerken, maar zeker ook voor het opsporen van geïnfecteerden. Alhoewel veel internetgebruikers via sociale media informatie met elkaar delen, waren veel deelnemers niet vertrouwd met het gericht doorsturen van een aantal uitnodigingen voor een vragenlijstonderzoek naar contacten. Een andere beperking

is dat het moeilijk is om met RDD te bepalen welke contacten worden uitgenodigd van het totaal aantal contacten van een deelnemer. Maar ook welke uitgenodigde contacten vervolgens meedoen en, nog belangrijker, wie van de uitgenodigde contacten niet meedoen. In onze studies zagen we een sterke neiging onder onze deelnemers om soortgelijke individuen uit te nodigen; dit is weliswaar een voorwaarde voor bron- en contactopsporing, maar vertekent het rekruteringsproces wanneer RDD wordt gebruikt voor het bestuderen van contactnetwerken. Er is meer onderzoek nodig om beter te begrijpen waarom individuen ervoor kiezen anderen wèl uit te nodigen, waarom ze specifieke contacten uitnodigen, welke rekruteringsopties het handigst voor hen zijn, en waarom hun contacten deelnemen, of waarom juist niet.

We zijn momenteel aan het onderzoeken of online RDD een nuttig instrument kan zijn voor de ondersteuning van GGD-personeel bij bron- en contactopsporing van infectieziekten die zich via contactnetwerken verspreiden, zoals kinkhoest en seksueel overdraagbare aandoeningen. Daarnaast verkennen we het gebruik van online RDD om clustering van dezelfde risicopercepties en gedrag in sociale netwerken te meten en daarna voor het verspreiden van educatieve berichten over (nieuwe) screening- of vaccinatieprogramma's, zodat individuen een goedgeïnformeerde keuze kunnen maken over hun deelname.

Dankwoord
(Acknowledgements)

'*It kin net*' dacht ik vaak in de eerste jaren van mijn promotieonderzoek. Mijn project kende een moeizame start en er was weinig houvast bij andere projecten of medeonderzoekers. Dat dit proefschrift er nu ligt voelt als een hele overwinning, maar het was niet gelukt zonder de hulp en ondersteuning van mijn begeleiders, collega's, vrienden en familie. Ik wil graag een aantal personen in het bijzonder bedanken die belangrijk waren voor de totstandkoming van dit proefschrift.

Allereerst wil ik graag mijn drie promotoren en co-promotor bedanken voor hun begeleiding en ondersteuning. Hooggeachte promotor Kretzschmar, beste Mirjam, hartelijk dank voor al je kritische beoordelingen en scherpe opmerkingen op mijn projectideeën, projectuitvoeringen en manuscripten. Dit zorgde voor een exponentiële leercurve (*of toch beta-binomiaal..?*) en maakte onze samenwerking alles behalve saai. Ook dank ik je voor je geduld, de nodige sturing en het aanbrengen van focus, maar ook voor het vertrouwen en de vrijheid die je mij gaf om een deel van mijn onderzoek in Thailand uit te voeren. Ik herinner mij nog vaak je uitspraken in de periode 2010-2012: "*moet je nu alweer naar Thailand..?*". Ik ben ook mijn twee andere promotoren van de faculteit Sociale Wetenschappen van de Universiteit Utrecht, prof. dr. ir. V. Buskens en prof. dr. P.G.M. van der Heijden, zeer dankbaar voor hun betrokkenheid, sturing en ondersteuning. Vincent en Peter, bedankt dat ik altijd bij jullie mocht langskomen, met welke problemen dan ook. Ik heb dit altijd zeer gewaardeerd.

Zeer geachte copromotor dr. J.E. van Steenbergen, beste Jim, ook al gaan we nog even fijn door met ons respondentgestuurde 'speeltje', ik wil je ontzettend bedanken voor de plezierige begeleiding en ondersteuning op alle momenten dat het echt nodig was. Ik mocht je altijd lastigvallen, wanneer dan ook (*deze service verdwijnt nu toch niet hoop ik...?*). Je geduldige en uitvoerige manier van uitleggen is inspirerend; daar zouden velen een voorbeeld aan kunnen nemen. Als dank voor al je hulp wil ik je graag één jaar gratis Mart-ICT-noodhulp aanbieden voor de vele dagelijkse RIVM ICT-uitdagingen. Dit komt inclusief een cursus 'Track Changes', het ontcijferen van je manuscriptcorrecties met pen duurde namelijk meestal langer dan het verwerken ervan.

I would like to thank the members of the reading committee, prof. dr. ir. H.A. Smit, prof. dr. T.J.M. Verheij, prof. dr. J.B.F. de Wit, prof. dr. C.J.P.A. Hoebe and dr. S.D.W. Frost for studying and evaluating my thesis.

I would also like to acknowledge and thank prof. dr. A. Thorson and dr. L. Bengtsson. Dear Anna and Linus, many thanks for the warm welcomes in Stockholm, our collaboration, your input on my project plans and manuscripts, and for giving me the opportunity to develop and use the software systems. I sincerely hope we can continue our collaboration in the years to come (wherever you are working or travelling to) and further develop online respondent-

driven research. I would also like to thank dr. Xin Lu for his useful comments on my manuscripts and for helpful discussions on RDS estimators.

Furthermore, I thank dr. Charnchudhi Chanyasanha and dr. Mathuros Tipayamongkholgul of the Faculty of Public Health of the Mahidol University in Thailand for the pleasant collaboration and their help with setting up part of my research project in Thailand. Dear Charnchudhi, many thanks for your kind hospitality, facilitating my visits to the university and for giving me the opportunity to meet your nice colleagues of the Microbiology department. I also thank all staff and students of the Microbiology department for the great times. In particular P'Tah and N'Pao for their support with my project and the very nice (but very spicy) Thai dinners.

I am very grateful to prof. dr. Richard Coker and his London School of Hygiene and Tropical Medicine team in Bangkok for providing me during each visit to Thailand with working facilities at the CDPRG. Also many thanks to Jame(s), Nat, P'Gig, Tom & Jess, Piya & Pam, Aronrag & Ben for the great times and dinners, let's meet again soon! (and P'Gig, I am still saving to be able to pay your bills... ขอโทษครับ).

Carl Koppeschaar en Ronald Smallenburg van de GroteGriepmeting, hartelijk dank voor de samenwerking en de mogelijkheid om gebruik te maken van jullie online burgerpanel. Ik kijk met veel plezier terug op onze (onverwacht gezamenlijke) trip naar de conferentie in Turijn.

Geachte prof. dr. Marc Bonten, beste Marc, bedankt dat ik zo nu en dan bij je terecht kon en voor het mogen bijwonen van de Wednesday Morning Meeting. Ik heb hier veel van geleerd. Ik wil ook graag alle andere Julius-collega's bedanken voor hun hulp en gezelligheid tijdens mijn promotietraject. In het bijzonder het modelleringsgroepje: Anneke, KaYin, Manon, John, Axel, Welling en Gijs, dank voor de fijne motiverende discussies (en gezellige gesprekken). Ik ben ook Ewoud (Schuit), mijn mentor, zeer dankbaar voor alle ondersteuning op persoonlijk vlak, én met R. Beste Martin (Bootsma), bedankt dat je deur altijd open stond en dat je altijd bereid was om mee te denken. Ook dank ik alle kamer- en afdelingsgenoten op alle verschillende kamers over de tijd (Christiana, Carla, Jorien, Marloes, Floor, Marleen, Marijn, Paula, Janneke, Mirjam, Judith, Henok, Linda, Irene, Kim) voor de gezellige tijd, lunches en etentjes.

Zeergeleerde opponent dr. A. Timen, beste Aura, bedankt voor alle hulp en steun. Dankzij jou kwam ik bij het RIVM én in Azië terecht. Zonder jou géén vrouw en géén kind. Ik denk nog vaak met veel plezier terug aan onze reizen naar het verre oosten en onze gesprekken op vliegvelden, bijvoorbeeld over de 'prachtige' vloerbedekking in Singapore. Luang Prabang was toch wel de meest bijzondere bijeenkomst, met collega's die niet kwamen opdagen voor het etentje met de projectgroep. Ik herinner me ook mijn genereuze aanbod op Suvarnabhumi, om mijn vlucht te nemen omdat jouw vliegtuig vertraagd was. Helaas lukte het je toch om een

andere oplossing te vinden. Ik bedank ook (soon-to-be) dr. Desirée voor alle coaching in de afgelopen jaren. Desirée, je gaf mij de nodige sturing en leerde mij om anders te kijken naar dagelijkse situaties: niet het probleem blijven benoemen ('ja, dat weten we nu wel'), maar oplossen of er zelf mee leren omgaan. Erg nuttig, zeker nu ik voor jou werk. Ik wil ook alle anderen van mijn LCI-familie bedanken voor hun gastvrijheid, ondersteuning én gezelligheid tijdens mijn promotietraject. Beste Albert Wong, bedankt voor alle statistische hulp en ondersteuning met R.

Ilona en Jaike, bedankt dat jullie mijn paranimfen willen zijn, ik vind het een eer dat jullie mij willen bijstaan op de 13e. Ilona, ook na Uilenstede bleven we elkaar gelukkig zien. Ik geniet altijd erg van onze gezellige etentjes samen met de jongens. We moeten dit volhouden de komende jaren, ongeacht de drukke agenda's. Jaike, vreemd genoeg leerden we elkaar pas in het laatste jaar kennen, ondanks onze gezamenlijke interesse voor de wetenschap. Dit had best wat eerder mogen gebeuren, je bent namelijk één van de weinige collega's die mijn 'geflauwekul' doorziet (en er ook nog om kan lachen).

Beste vrienden en familieleden, bedankt voor jullie gezelligheid en steun in de afgelopen jaren. David & Inge, Luuk, Milou, Tong, Tae, Pae, Pong, Mai, Marwin, mijn vijf broers én schoonzussen, Tae and parents in law: thank you very much for bringing me joy during hard times and for your interest in my research progress. Beste Jasper, bedankt voor al je hulp met het programmeren van mijn vragenlijstsystemen, soms tot in de late uurtjes. Maar ook voor alle momenten dat ik bij je terecht kon voor advies over software en servers. Lieve ouders, bedankt voor jullie vertrouwen in mijn kunnen en de hulp die jullie mij gaven wanneer dat nodig was.

ถึง Tam ที่รักของผม งานวิจัยนี้จะไม่เกิดขึ้นเลยถ้าขาดคุณ, จริง จริง. เราทั้งคู่ผ่านเวลาที่มากล้ำมาก แต่คุณก็มั่งมีนอยู่ข้างผม ช่วยเหลือผมและฟังผมบ่นเรื่องงานที่ไม่เคยมีที่ท่าว่าจะจบลงง่ายๆ และที่สำคัญคุณเลือกตัดสินใจที่จะจากครอบครัว เพื่อนมันพันๆ คนและงานที่เมืองไทยแต่เฟี้ยงเพื่อที่จะมาอยู่กับผมที่อูเทรค เพื่อที่จะให้ผมทำสิ่งที่ริ่มต้นนี้ให้สำเร็จ เพราะสิ่งเหล่านี้เรากังมีอันนี้ ผมภูมิใจในตัวคุณมาก คุณหริมหที่จะอ่าน พูด ฟัง และเฟี้ยนภาษาดัตช์ได้อย่างเร็ว ทำความรู้จักเพื่อนใหม่ๆ ในอูเทรคและมังหางานทำได้อีกด้วย แต่ที่สำคัญที่สุดคุณเป็นซุปเปอร์มัมสำหรับทิมเจ้าตัวน้อยของเรา ผมรักคุณและTim มากๆ และต่อจากนี้ก็เป็นเวลาที่เราจะสนุกกับครอบครัวม้าๆ เล็กๆ ของเรานะ (and yes, I will learn to speak more Thai now).

Mart Stein
Oktober 2016

About the author

About the author

Mart was born on 26 March 1987 in Houten, the Netherlands. He followed his secondary education in Sneek (RSG Magister Alvinus), from which he graduated in 2005. In that year, he started the study Health and Life Sciences at the VU University in Amsterdam. After completion of his bachelor in 2008, he continued with a Masters program in Health Sciences, with specialisation Infectious Diseases and Public Health. During his master studies he was an intern at the National Coordination Centre for Communicable Disease Control (LCI), part of the Centre for Infectious Disease Control of the National Institute for Public Health and the Environment (RIVM) in Bilthoven. His internship contributed to the international EU-funded AsiaFluCap project, which aimed to assess and strengthen the health care capacity to respond to pandemics in several countries in Southeast Asia.

In 2009, he started to work as a junior researcher at the LCI. He continued working on the AsiaFluCap project and was involved in the Dutch national evaluation of the infectious disease control policy in the 2009 (H1N1) pandemic. His work at the LCI made clear to him that he wanted to pursue a scientific career in infectious disease epidemiology.

In 2011, Mart started his PhD research at the Julius Center for Health Sciences and Primary Care of the University Medical Center Utrecht, resulting in this thesis. He collaborated with researchers at Utrecht University and the Centre for Infectious Disease Control, and additionally set up international collaborations with Mahidol University Bangkok, Thailand, and Karolinska Institutet, Sweden. Part of this PhD took place in Thailand, and he participated in several conferences in Europe and the US. During his PhD, he completed a Master in Epidemiology at Utrecht University, with specialisation Epidemiology of Infectious Diseases.

In 2015, Mart received the Young Researcher Award and Innovation prize at the RIVM. Since October 2016, he is employed as a postdoctoral researcher at the RIVM. At the LCI he is involved in several projects on the use of network-based methods, among which an international project on using webRDS among MSM in Vietnam, and a project on the dynamics of decision-making and preventive health behaviors within social networks. He also coordinates the development of renewed software for online respondent-driven surveys.

Courses and Training

2nd Summer school on network models in STI epidemiology. Five-day course on network analyses by I-Biostat, University Hasselt and Centre for Health Economics Research & Modelling Infectious Diseases (Chermid), University of Antwerp. Bruges, Belgium, September 2015.

Respondent-Driven Sampling. Two-day course by Tulane University. KNCV, The Hague, the Netherlands, April, 2014.

Master Epidemiology Postgraduate, University of Utrecht, the Netherlands. 2012-2014.

Core courses

Introduction to Epidemiology	3.0 EC
Introduction to Statistics and SPSS	1.5 EC
Study Design in Etiologic Research	4.5 EC
Classical methods in Data Analysis	6.0 EC
Modern methods in Data Analysis	4.5 EC
Presentation and Writing Research Proposal	1.0 EC
Research Ethics and Society	1.0 EC

Specialisation course: Epidemiology of Infectious Diseases

Clinical Epidemiology	1.0 EC
Advanced Topics in Etiologic Research: Confounding and Effect Modification	1.5 EC
Meta-Analysis	1.5 EC
Epidemiology of Infectious Diseases	1.5 EC
Basics of mathematical modelling of infectious diseases	1.5 EC
Advanced mathematical modelling of infectious diseases	1.5 EC

Bayesian Statistics	1.5 EC
Boekenclub Julius Center	1.0 EC
Journal club Julius Center	1.0 EC

Mathematical modelling course: within and between-host dynamics of drug-resistant pathogens. Two-day course by Harvard School of Public Health, University of Hong Kong and Mahidol-Oxford Tropical Medicine Research Unit. Bangkok, Thailand. June 2012.

A practical short course on Infectious Disease Modelling. A three-day course by Harvard School of Public Health, University of Hong Kong, John Hopkins University and Mahidol-Oxford Tropical Medicine Research Unit. Bangkok, Thailand. June 2012.

Oral presentations

Stein ML, Van der Heijden PGM, Buskens V, Van Steenbergen JE, Bengtsson L, Koppeschaar CE, Thorson A, Kretzschmar MEE. Tracking social contact networks with online respondent-driven detection. Epidemics 5, Florida, the United States. December 2015.

Stein ML. Social Networks and Control Strategies. VvAwT Nascholing. Amersfoort, the Netherlands, January 2015.

Stein ML, Van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, Van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar MEE. Web-based respondent-driven sampling and contact patterns relevant for the spread of respiratory pathogens: a pilot study in Thailand. Werkgroep Epidemiologisch Onderzoek Nederland (WEON), Utrecht, the Netherlands, June 2013.

Stein ML. Online Respondent-Driven Sampling for studying contact patterns relevant for the spread of close-contact pathogen. UCID symposium, Utrecht, The Netherlands. June 2013

Stein ML, Van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, Van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar MEE. Online Respondent-Driven Sampling for studying contact patterns relevant for spread of close-contact pathogens: a pilot study in Thailand. Digital Epidemiology Workshop, Turin, Italy. May 2013.

Stein ML. Respondent-driven sampling for studying contact network patterns relevant for the spread of close-contact pathogens. Seminar Microbiology department Mahidol University, Bangkok, Thailand. January 2013.

Stein ML, Kretzschmar MEE. Respondent Driven Sampling and control strategies for emerging infectious diseases. Workshop on Respondent-Driven Sampling. Department of Mathematics, Stockholm University, Stockholm, Sweden. December 2011.

Poster presentations

Stein ML, Van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, Van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar MEE. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. Julius PhD Seminar, Zeist, the Netherlands. February 2014.

Stein ML, Van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, Van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar MEE. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. Epidemics 4, Amsterdam, the Netherlands. December 2013.

Stein ML, Chanyasanha C, Tipayamongkholgul M, Sabaiwan W, Van Steenbergen JE, Buskens V, Van der Heijden PGM, Bengtsson L, Thorson AE, Kretzschmar MEE. Respondent-driven sampling for studying contact network patterns relevant to the transmission of respiratory-spread pathogens in the Netherlands and Thailand. Julius PhD Seminar, Zeist, the Netherlands. February 2013.

Other activities

Supervision of medical students "Architectuur Klinisch Wetenschappelijk Onderzoek" (AKWO). Course years 2012, 2013 and 2014.

Reviewer of international scientific journals:

BMC Infectious Diseases
PLoS ONE
Lancet Respiratory Medicine
International Journal of Geographical Information Science
PharmacoEconomics

Coordinator of the development of two generic software systems for conducting web based respondent-driven surveys.

Grants and awards

The contagiousness of social networks: analysing the dynamics of risk perception regarding and participation in screening programs within real-world social networks: Contagion-RWS. (Applicant) Strategisch Programma RIVM (SPR) 2015. [€285,000]

AroundYou: Development of a digital tool for contact tracing with citizens. (Project leader) RIVM Innovation Prize 2015. [€60,000]

Young Researcher Award. RIVM, Bilthoven. 2015.

Publications

Stein ML, van Vliet JA, Timen A. Chronological overview of the 2009/2010 H1N1 influenza pandemic and the response of the Centre for Infectious Disease Control of the National Institute for Public Health and Environment (RIVM). RIVM, July 2011; report 215011006. [Available in Dutch and English]

Krumkamp R, Kretzschmar M, Rudge JW, Ahmad A, Hanvoravongchai P, Westenhoefer J, **Stein M**, Putthasri W, Coker R. Health service resource needs for pandemic influenza in developing countries: a linked transmission dynamics, interventions and resource demand model. Epidemiology and Infection, 2011; 139; 59-67.

Stein ML, Rudge JW, Coker R, van der Weijden C, Krumkamp R, Hanvoravongchai P, Chavez I, Putthasri W, Phommasack B, Adisasmito W, Touch S, Sat LM, Hsu YC, Kretzschmar M and Aura Timen. Development of a resource modelling tool to support decision makers in pandemic influenza preparedness: The AsiaFluCap Simulator. BMC Public Health, 2012;12:870.

Rudge JW, Hanvoravongchai P, Krumkamp R, Chavez I, Adisasmito W, Chau PN, Phommasak B, Putthasri W, Shih CS, **Stein M**, Timen A, Touch S, Reintjes R, Coker R on behalf of the AsiaFluCap Project Consortium. Health system resource gaps and associated mortality from pandemic influenza across six Asian territories. PLoS One, 2012;7(2):e31800.

van der Weijden CP, **Stein ML**, Jacobi AJ, Kretzschmar ME, Reintjes RR, van Steenbergen JE, Timen A. Choosing pandemic parameters for pandemic preparedness planning: A comparison of pandemic scenarios prior to and following the influenza A(H1N1) 2009 pandemic. Health Policy, 2013; 109(1):52-62.

Stein ML, Rump BO, Kretzschmar MEE, van Steenbergen JE. Social networking sites as a tool for contact tracing: urge for ethical framework for normative guidance. Public Health Ethics 2014; 7(1):57-60.

Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, Kretzschmar ME. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand. PLoS One. 2014 Jan 8;9(1):e85256.

Stein ML, van Steenbergen JE, Buskens V, van der Heijden PG, Chanyasanha C, Tipayamongkholgul M, Thorson AE, Bengtsson L, Lu X, Kretzschmar ME. Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling. PloS one. 2014;9(11):e113711.

Stein ML, Steenbergen JE, Buskens V, van der Heijden PGM, Koppeschaar CE, Bengtsson L et al. Enhancing syndromic surveillance with online respondent-driven detection. American Journal of Public Health. August 2015, Vol. 105, No. 8, pp. e90-e97.

Stein ML, van der Heijden PGM, Buskens V, Steenbergen JE, Bengtsson L, Koppeschaar CE, Thorson A and Kretzschmar MEE. Tracking social contact networks with online respondent-driven detection: who recruits whom? BMC Infectious Diseases. November 2015, 12: 522.

Stein ML, Steenbergen JE, Dukers-Muijrers NHTM, Buskens V, van der Heijden PGM, Kretzschmar MEE. Online respondent-driven detection for case finding and public health interventions. [Submitted]

Stein ML, Buskens V, van der Heijden PGM, Steenbergen JE, Kretzschmar MEE. A stochastic simulation model to study respondent-driven recruitment. [Submitted]