FOA: Combining Subjective and Objective Methods for Quantifying Contact Rates and Mixing Pattern in School-Aged Children

Centers for Disease Control and Prevention (CDC)
National Center for Emerging and Zoonotic Infectious Diseases (NCEZID)
U01 Research Project Cooperative Agreements
RFA-CK-11-006

This FOA addresses the "Healthy People 2010" focus area of Immunization and Infectious Diseases,. This FOA is in alignment with HHS/CDC/NCIRD's performance goal(s) to Prevent disease, disability, and death from infectious diseases, including vaccine-preventable diseases.. For more information, see www.health.gov/healthypeople and http://intra-apps.cdc.gov/fmo/

The purpose of this funding is to facilitate research that describes individual social contact and mixing patterns in school-aged children across a range of community settings to improve the precision of contact rate estimation and parameterization for infectious disease transmission models to support the development of disease prevention and control strategies. The objectives of these studies are to quantify age-specific individual social contact and mixing patterns of school-aged children as they migrate within and across a range of community settings (schools, home, dormitories, community) using both previously established and newer methods of assessment.

Background
During the response to the emerging 2009 H1N1 pandemic public health officials leveraged infectious disease models to develop a range of plausible pandemic severity scenarios to explore the implementation of both pharmaceutical and non pharmaceutical interventions aimed at mitigating the impact of the pandemic. Essential components of infectious disease models which influence disease transmission dynamics are the assumed contact rates and mixing patterns of infectious and susceptible members of the modeled population. Due to a paucity of empirical data from the United States supporting these estimates, recent efforts have attempted to directly quantify these parameters within and across various age groups and community settings using population based surveys and self reported diaries of contact and mixing patterns. Most of these studies rely on self reporting and may not objectively capture individual proximity or the duration of contact both of which have been shown to vary by location and which may be important for disease transmission. Quantification of individual contact rates and mixing patterns through a combination of subjective and objective methods may provide closer approximations of these variables within infectious disease models as well as offer a means of validating current methods of assessment. The role of contact rates and mixing patterns are of particular interest in school-aged children who have been shown to be the sentinel cohort diagnosed with influenza in syndromic surveillance systems, experience higher rates of infection during the influenza season and shed influenza virus for approximately twice as long as the estimated period of infectiousness for adults. Due to uncertainty regarding the benefit

of social distancing measures such as school closure which is sensitive to contact rates and mixing patterns, more precise estimates of these variables will support further research into the benefit of this and other non pharmaceutical interventions.

Below are the objectives, which the applicant must respond to:

- Describe and quantify age-specific individual social contact and mixing patterns of school-aged children (K through 12th grade) across a range of community settings (schools, dormitory, home, community) using both previously established and newer methods of quantitative assessment.

- Studies must use (1) representative cross-sectional studies (surveys), or prospective observational studies, (2) prospective self diaries/evaluations, and a smaller proportion of data collection through the use of (3) GPS technologies in the form of Personal Tracking Devices to collect individual contact and mixing data consistent with or superior to the proximity determination limitations of current technology

- Studies should attempt to study as many of the following as possible:

    - elementary schools
    - high schools selected from both urban and rural settings
    - one specialized setting (boarding school, camp).

- The studies must propose methods to collect variables that may be related to contact rates and mixing patters including but not limited to the gender and age of study subjects, the type, location, frequency and duration of contacts, classroom sizes, and relevant characteristic of the physical environment such as square footage.

- Study proposals must incorporate an active surveillance component to permit correlation analysis between observed contact/mixing patterns and the incidence of influenza-like illness over the course of the study period.

- Since seasonality may influence contact rates and mixing patterns, the potential study period should include all twelve months of the year, including weekdays and weekends during the chosen study period

- All studies must provide a plan to ensure the security and protection of the identities of the study participants.

- Studies must compare the observations from GPS technologies in the form of Personal Tracking Devices to assess the level of agreement in estimating contact rates and mixing patterns.

- Study proposals must include a plan to document uptake of any routine disease prevention measures among participants (e.g. seasonal influenza vaccination), as well as any non-pharmaceutical interventions that are routinely or reactively implemented during the course of an outbreak of a influenza-like respiratory illness

- Propose a plan for laboratory testing for influenza to be implemented if an acute outbreak of influenza-like illness is observed during the observation period of the study.

- Use reverse transcription polymerase chain reaction (RT-PCR) testing to identify influenza viruses and not rapid diagnostic which lack optimal sensitivity in low prevalence settings.

Society For Epidemiological Research
Annual Conference 2013: Boston

Analyzing Social Contact Networks For Modeling Infectious Disease Transmission

Biosketches of Panel

PRESENTERS

Derek Cummings is an Associate Professor at the Department of Epidemiology at Johns Hopkins Bloomberg School of Public Health. The goal of his research is to understand the temporal and spatial dynamics of the spread of infectious diseases in order to inform interventions to control their spread. His research includes empirical work and theoretical approaches to simulate the spread of pathogens in populations. He is specifically interested in the dynamics of influenza, dengue hemorrhagic fever, measles and chikungunya. Dr. Cummings is Co-PI of SMART (Social Mixing and respiratory transmission). Email: *dcumming@jhsph.edu*

Molly Leecaster is Assistant Professor at the University Of Utah School Of Medicine, Division of Epidemiology and a researcher with the Salt Lake City Veterans Health Administration. She received her BS in mathematics at the University of Wisconsin, MS in statistics from Virginia Tech, and PhD in statistics from Colorado State University. Her statistical expertise is in sample design and modeling and has covered applications from environmental characterization to national security, and infectious diseases, healthcare epidemiology, and health services. She has been involved in modeling projects for MRSA, C. difficile, RSV, H1N1, and seasonal influenza, applying compartmental, individual-level, and agent-based models. The methods include frequentist and Bayesian statistics, data-driven, and model-based approaches to estimate model parameters and predict transmission and epidemic characteristics. Dr. Leecaster is the Co-PI of Contact among Utah's School-age Population (CUSP). Email: *Molly.Leecaster@utah.edu*

Marcel Salathé, PhD is an assistant professor of biology at the Center for Infectious Disease Dynamics. A Branco Weiss Society in Science fellow, he studies how social networks affect the spread and control of infectious diseases. His research group currently uses complex systems models, wireless sensor network technology and large-scale data sets from online social media sites to analyze the spread of disease and health behaviors on social networks. The group's main goal is to measure and improve health outcomes with basic research, mobile technology and social media. His research program is rooted in four observations (in no particular order of priority): Fundamentally, health and disease are biological phenomena, but ignoring the effect of human behaviors on health and disease outcomes would be ignoring the main drivers of health and disease dynamics in the 21st century; The internet - in all its flavors, ranging from static websites, to communication tools such as email, to social media, to the mobile internet (smartphones, sensors, etc.) - has become a source of information about human behaviors at an unparalleled scale. This opens up completely new

research fields; The ability to collect, mine, filter, analyze and visualize enormously large data sets from this data source is one of the great practical and educational challenges of our times; Programming is becoming the lingua franca of science. Email: *salathe@psu.edu*

Shanta Zimmer is Associate Professor of Medicine at the University of Pittsburgh School of medicine. She is an infectious diseases physician with research and clinical experience in immunology and vaccinology, clinical research, infectious disease modeling, and education and translational research examining dynamics of influenza transmission and epidemic prevention. She is a research consultant to research projects at the Center for Vaccine Research and the Public Health Dynamics Lab at the University of Pittsburgh. She is also involved in implementation of hospital and outpatient clinic quality improvement strategies to reduce respiratory viral transmission in the clinical setting. Dr. Zimmer is Co-PI on the Social Mixing and Respiratory Transmission (SMART). Email: *zimmersm2@upmc.edu*

Jonathan Read is a Lecturer within the Epidemiology and Population Health Department of the University of Liverpool, UK. His research interests include the transmission and evolutionary dynamics of infectious diseases, for both human and animal pathogens. A substantial part of this interest lies in understanding and quantifying patterns of mixing and travel, and the formation of contact networks. He currently works on infectious disease projects in UK, USA, China, Hong Kong, Thailand, Vietnam and Malawi. Dr Read is a co-I on the SMART project. Email: *Jonathan.Read@liverpool.ac.uk*

Damon Toth, PhD, is an Assistant Professor in the Division of Epidemiology and an Adjunct Assistant Professor in the Department of Mathematics at the University of Utah. He received a PhD in Applied Mathematics from the University of Washington. He has designed mathematical models and computer simulations of infection and transmission of disease in schools, hospitals, and larger communities, for use in risk assessment and in planning strategies for interventions to reduce transmission. As part of the CUSP (Contact among Utah's School-age Population) study, he has designed an agent-based influenza transmission model that makes direct use of high fidelity contact data from several schools and other venues in Utah. Email: *toth@math.utah.edu*

Amra Uzicanin MD MPH is a medical epidemiologist with the Centers for Disease Control and Prevention (CDC) in Atlanta, GA. Since 2010, she has been leading a new group at CDC dedicated to developing the scientific evidence base for use of non-pharmaceutical interventions for infectious disease control with focus on pandemic influenza. Her research interests include influenza and other respiratory infections, respiratory infectious disease transmission dynamics, and measles, rubella, and other vaccine-preventable diseases. Dr. Uzicanin has been with CDC since 1998, first as an Epidemic Intelligence Service Officer, and then serving in various scientific positions with CDC's global immunization programs. Prior to joining CDC she led multiple international medical programs for the International Federation of Red Cross and Red Crescent Societies and worked as a practicing physician in her native Bosnia and

Herzegovina. Dr. Uzicanin leads a unit at CDC that initiated, supported and funded the research presented here. Email: *aau5@cdc.gov*

MODERATORS

Charles J. Vukotich, Jr. is a Senior Project Manager in the Center for Public health Practice at the Graduate School of Public Health, University of Pittsburgh. He retired after 30 years from the Allegheny County Health Department (ACHD). His research interests includes studying how children catch, spread, and prevent diseases in schools (k-12), focusing primarily on pandemic and seasonal influenza. He has also studied the integration of research into schools, directing the School Based Research and Practice Network, and worked on public health preparedness. Mr. Vukotich is the Project Manager for SMART. Email: *charlesv@pitt.edu*

Jeanette Rainey, MPH, PhD is an epidemiologist with the Division of Global Migration and Quarantine at the Centers for Disease Control and Prevention (CDC). She assists with the coordination and implementation of research on the effectiveness and feasibility of non-pharmaceutical interventions to mitigate pandemic influenza and other acute respiratory infections. Dr. Rainey has extensive international research experience related to vaccine preventable disease surveillance and vaccination coverage assessments. Prior to her time with CDC, she worked with the Los Angeles County and California State Health Departments. Dr. Rainey is the CDC program officer for the 'Quantifying Contact Rates and Mixing Pattern in School-Aged Children' projects. Email: *jkr7@cdc.gov*

.

Put your ID sticker in this box

The SMART study

Social Mixing And Respiratory Transmission in schools

Please write carefully and, where appropriate, mark boxes with an 'X'

About You

1 What is your school grade?

☐☐

2 What is your school ID code?

☐☐☐☐☐☐☐☐

Write one letter or number in each box, like this 1 2 3 4

3 How old are you?

I am ☐☐ years old

4 Are you a boy or a girl?

a boy ☐ a girl ☐

Put an X in one box, like this ☒

About your family and home

5 Not counting you, how many people live in your home?

☐☐ people

6 Does any other person sleep in your bedroom?

No ☐ Yes ☐

7 If yes, how many people sleep in your bedroom, not counting you?

☐☐ people

7 Does anyone in your house go to pre-school or day care?

No ☐ Yes ☐

If yes, how many people go to pre-school or day care?

☐☐ people

8 Does anyone in your house go to elementary school (grades K to 6)?

No ☐ Yes ☐

If yes, how many people?

☐☐ people

9 Does anyone in your house go to middle school (grades 7 or 8)? *Do not count yourself*

No ☐ Yes ☐

If yes, how many people?

☐☐ people

10 Does anyone in your house go to high school (grades 9 to 12)? *Do not count yourself*

No ☐ Yes ☐

If yes, how many people?

☐☐ people

11 What is the zipcode of your home address?

☐☐☐☐☐

About where you go

12 Where is the furthest place from your home you went in the past **7** days?

City or town
County
State
Country

13 Where is the furthest place from your home you went in the past **30** days?

City or town
County
State
Country

About feeling sick and staying away from school

Sometimes when we are sick, we don't go to school. These questions are about the last time you were sick **and** didn't go to school.

14 When you missed school, did you stay at your home?

No ☐ Yes ☐

15 When you missed school, who took care of you?

No-one ☐ Your Mom or Dad ☐

A brother or sister, or someone else you live with ☐ Someone who doesn't live with you ☐

Don't know ☐

16 Did you get a flu vaccination this school year?

No ☐ Yes ☐ Don't know ☐

17 Do you believe flu vaccines protect you against the flu?

No ☐ Yes ☐ Don't know ☐

About YESTERDAY

The rest of this questionnaire asks about your day YESTERDAY. If you can't remember what you did yesterday, ask your teacher for help.

18 Did you attend school YESTERDAY?

No ☐ Yes ☐

19 If you missed school, why was this?

I was sick or ill ☐

School was closed ☐

Some other reason ☐

20 How did you get to school YESTERDAY?

walked or biked ☐ public bus ☐

by car ☐ some other way ☐

school bus ☐ didn't go to school ☐

Your contact diary

This page asks you about the people you met yesterday

- people you talked with
- people you played with
- people you touched with your hands or face

Steps to follow

step 1 In SECTION A write down the name or nick-name for everyone you met **yesterday**.
For example, "mom", "Derek", "my best friend", "mailman".
Don't write more than one person's name in each box.

step 2 Answer questions 21 to 32 for each of your contacts.

step 3 Ask your teacher for your two random numbers.
Write these in the spaces at the top of questions 33 and 34
These numbers refer to two of your contacts.
Answer questions 33 and 34 for all of your contacts.

Do include:

- ✓ children and teachers you spoke to at school
- ✓ children or adults you to outside of school
- ✓ people whose skin you touched *(e.g. while playing games)*
- ✓ anyone you briefly spoke with on the way home from school

Do **not** include:

- ✗ people you did not talk to
- ✗ people you only talked with through a telephone or computer
- ✗ pets or toys
- ✗ someone you walked past, but did not speak to or touch their skin
- ✗ people you did not meet yesterday

Put your ID sticker in this box

START HERE

SECTION A

Your contacts
Write the name or a description of each person you met yesterday

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

21 Is this person ... ? — a boy or man; a girl or woman

22 How old is this person? *Enter their age, or your best guess*

23 Does this person go to school? — yes; no; don't know

What grade are they? *e.g. K, 7, 12*

24 If they go to school ... — Do they go to your school? yes; no

25 yes; no

26 Do you live with this person? — yes; no

27 Where did you meet yesterday? *Mark all the places you met them yesterday* — my home; a friend's home; a relative's home; school, in class; school, not in class; on school bus; sports practice or event; mall or shopping centre; grocery store; another type of store; restaurant; religious service; somewhere else

28 Did you touch their skin? *e.g. holding hands, playing together* — yes; no

Did you both hold the same toy or object at all yesterday? *e.g. you shared a crayon or hair brush, you swapped food or drink* — yes; no

29 How long did you spend together during the whole of yesterday? — less than 10 min; 10 to 29 min; 29 to 59 min; at least 1 hour, but less than 2; at least 2 hours but less than 4; 4 or more hours

30 How often do you normally see this person? — every day; most school days; once a week; just met them

31

32 Deliberately blank

Write your 1st random number here → 33

Write your 2nd random number here → 34

Did this person meet contact number ___ yesterday? — no; probably no; probably yes; yes

Did this person meet contact number ___ yesterday? — no; probably no; probably yes; yes

The SMART study
Social Mixing And Respiratory Transmission in schools

Put your ID sticker in this box

35 Did you meet any more people yesterday that you haven't told us about?

No ☐ Yes ☐

36 If yes, how many more people did you meet yesterday?
Write how many in the boxes

Babies and infants (0-4 years old) ☐☐☐

Children (5-18 years old) ☐☐☐

Grown-ups (19 or older) ☐☐☐

37 Did you meet more or less people yesterday than normal?

Less ☐

About the same ☐

More ☐

38 How easy did you find this questionnaire?

Very easy ☐

Easy ☐

Hard ☐

Very hard ☐

Don't know ☐

39 What was hard about it?
Please write in the box below.

When you have finished and checked your form:

- tear off and keep SECTION A from your contact diary page
- make sure you have put your ID stickers in the boxes on all three pages
- hand all the pages back to your teacher

Questionnaire for Year 2 contact and symptom surveillance

Version: yr2A.4
Date: ##########

Additional requirements: 1. List of 12 named individuals specifc to the subject

About the subject:

ID of subject	
School ID	
Date of interview	*mm: dd: yy:*

Subject absent (or unable to interview)?
☐ YES [1] ☐ NO [0]

NOTE: When delivered on a MONDAY, YESTERDAY refers to Friday the previous week, in all sections of this questionnaire.

SECTION A. ACTIVITIES YESTERDAY

Q1 How many people did you meet in your home yesterday? (including people you live with)
Write an integer number (zero for nobody) or DON'T KNOW [-2], or DECLINE [-1].

Q2 Did you go to any of these places or do these things yesterday?

Someone else's home Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

School Bus Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Public Transport (including Port Authority bus) Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

After-school programme Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Sports Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Shopping Mall or Store Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Restaurant Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Movies or Pictures Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Church Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Doctors Office Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Other or special event
(particularly if involving lots of individuals)
Please specify

Q3 Were you absent from school yesterday?
Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

Q4 Did you get a flu vaccination this school year? *(since August)*
Yes [1] ☐ No [0] ☐ *Don't know [-2]* ☐ *Decline [-1]* ☐

SECTION B.RELATIONSHIP AND CONTACT WITH NAMED INDIVIDUALS

Please refer to your list of names. These correspond to the numbered rows below.
Ask the subject questions about each of the named individuals.

Q5 Do you know who [NAME] is?

IF Q5=YES
Q6 Is this person a friend of yours?

IF Q5=YES
Q7 Did you talk with [NAME] yesterday? *If NO go to next NAME and Q5.*

IF Q7=YES
These questions are all about YESTERDAY (or the previous school day if yesterday was a holiday/weekend)

Q8 Where did you talk to this person? *Tick all that apply. [0 or 1]*

Q9 Did you touch their skin?

Q10 Did you both hold the same toy or object yesterday?

Q11 How long did you spend together overall yesterday?

Only ask question Q12 when completed Q5-Q11 for all 12 individuals.

Q12 Out of all of the people that I have asked about, which one do you think has the most friends? *Mark the individual. [0 or 1]*

Name of Named Individual	NAME ID NUMBER	Q5 No [0]	Q5 Yes [1]	Q5 *Don't Know [-2]*	Q6 No [0]	Q6 Yes [1]	Q6 *Don't Know [-2]*	Q7 No [0]	Q7 Yes [1]	Q7 *Don't Know [-2]*	Q8 School, in class	Q8 School, not in class	Q8 Home	Q8 Someone else's home	Q8 Somewhere else	Q8 Don't Know	Q8 Decline	Q9 No [0]	Q9 Yes [1]	Q9 Don't Know [-2]	Q9 Decline [-1]	Q10 No [0]	Q10 Yes [1]	Q10 Don't Know [-2]	Q10 Decline [-1]	Q11 less than 10 min [1]	Q11 10 min - 1 hour [2]	Q11 more than 1 hour [3]	Q11 Don't Know [-2]	Q11 Decline [-1]	Q12
	1	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	2	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	3	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	4	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	5	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	6	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	7	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	8	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	9	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	##	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	##	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	##	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

SECTION C. SYMPTOMS OF INFLUENZA LIKE ILLNESS

Q13 Have you been sick in the last seven days (since last [DAY])?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

If Q13 = YES

Q14a Did you have a cough?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14b Did you have a sore throat?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14c Did you have a runny nose?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14d Did you have a fever?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14e Did you have a headache?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14f Did you have vomiting / diarrhea / nausea?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q14g Did you have muscle aches?

- ☐ No [0]
- ☐ Yes [1]
- ☐ *Don't Know [-2]*
- ☐ *Decline [-1]*

Q15 How many people you live with have been sick in the past seven days?

Write an integer number (zero for nobody) or DON'T KNOW [-2], or DECLINE [-1].

THIS IS THE END OF THE INTERVIEW. THANK THE SUBJECT FOR THEIR TIME AND ASSISTANCE.
DIRECT THE SUBJECT BACK TO THEIR CLASS OR WAITING AREA.

post-gazette.com
Pittsburgh Post-Gazette

Washington County students carry devices to help Pitt monitor spread of flu

November 5, 2012 12:26 am
By Jack Kelly / Pittsburgh Post-Gazette

With the help of children who'll be off of school for Tuesday's election, researchers at the University of Pittsburgh hope to find out if school closings can slow the spread of flu and other disease.

Today researchers will distribute remote sensors called motes to about 450 students at Borland Manor Elementary and North Strabane Intermediate schools in the Canon-McMillan School District in Washington County. Students will wear the motes, the size of a beeper and weighing 3 ounces, on lanyards around their necks today, Tuesday and during the school day Wednesday. Researchers will collect the motes before school is dismissed.

Powered by batteries, motes send out a signal that will detect another mote when they get close to each other, and the encounter is electronically recorded. Data collected from motes should give researchers a comprehensive picture of how often children interact.

"This is the first time this is being done anywhere, ever," said Charles Vukotich Jr., senior project manager at Pitt's Graduate School of Public Health.

The Social Mixing And Respiratory Transmission in Schools, or SMART, study is funded by a $700,000 grant from the U.S. Centers for Disease Control and Prevention.

The idea for the study was prompted by the 2009 H1N1 influenza pandemic, which infected more than a million Americans and led to the death of more than 18,000 people worldwide, according to the World Health Organization.

Seasonal flu typically affects older people. But H1N1 was first detected in a 10-year-old in California, then next in an 8-year-old.

The data from the motes will help Pitt researchers identify nonpharmaceutical means of containing flu and other epidemics among schoolchildren, Mr. Vukotich said. This is vitally important, he said, because "in the early part of a pandemic, there won't be a vaccine."

This was certainly true of H1N1, which, the CDC said, was "a unique combination of influenza virus genes never previously identified in either animals or people."

The purpose of the SMART study is to find answers to several questions about transmission during a flu outbreak: What are the most effective means of keeping it from spreading? During the school day, should

movement among classes be restricted? Would more vaccinations help? Should children who show signs of the flu be sent home? Be kept in a separate room? Should sick leave policies for teachers and administrators be changed?

Pitt researchers will use the data to construct models of how schoolchildren interact so they can develop the most effective preventive measures.

This is the second year for the SMART study. Last year elementary, middle and high school students in eight schools in the Canon-McMillan School District and Propel Charter Schools in Allegheny County wore the motes during a school day.

Last year's study showed that the typical student interacts with 109 children during the school day. High school students have more interactions than do younger students. Most interactions occur at lunchtime.

The expanded study this year will give Pitt researchers information about how often children interact outside of school.

Jack Kelly: jkelly@post-gazette.com or 412-263-1476.

First Published November 5, 2012 12:24 am

Print This Page

TRIBLIVE

Washington County students participate in CDC-Pitt study of how flu spreads

This is a mote, a three-ounce electronic device about the size of a beeper, that students in some Canon-McMillan schools will be wearing Nov. 5 - Nov.7 to measure who they come in contact with as part of a CDC study on how influenza spreads. The device will be worn on lanyards around their necks

By Craig Smith
PITTSBURGH TRIBUNE-REVIEW

Published: Saturday, November 3, 2012, 12:01 a.m.
Updated: Saturday, November 3, 2012

About 450 Washington County students will help University of Pittsburgh researchers next week learn more about how flu spreads.

The students, who attend Borland Manor Elementary and North Strabane Intermediate schools in the Canon-McMillan School District, will come home on Monday wearing electronic proximity sensors. The devices, called motes, will record when the students come in contact with each other.

These “electronic tags,” which could collectively record as many as 1 million pieces of data in a typical day, will tell researchers how many times kids come together for conversations, sharing items or other activities, and how far apart they are, said Charles Vukotich Jr., senior project manager at Pitt’s Graduate School of Public Health.

From that data, researchers will be able to better measure how influenza spreads in schools. If, that is, the kids hold up their end of the bargain and wear the 3-ounce devices all day on Monday,

About Craig Smith

Pittsburgh Tribune-Review Staff Reporter Craig Smith can be reached via e-mail or at 412-380-5646

Tuesday and Wednesday, when they'll be returned.

"The little kids kind of think they're cool," said Shanta Zimmer, associate professor in Pitt's School of Medicine. "The older kids ... we can tell there's been some tampering."

The motes send out a weak signal every 20 seconds to detect other motes and record when they detect one.

Students in Canon-McMillan schools participated in the study last year, but this will be the first time they will wear the motes on a scheduled day off from school. Preliminary data from last year's study, which also included Propel charter school students in Allegheny County, showed that each child interacted with an average of 109 other children during the school day.

One of the key questions the study hopes to answer is how effective closing schools might be in stopping the spread of flu, the researchers said.

"Last year there were significant numbers recorded overnight," Vukotich said, providing some evidence that simply closing schools for a few days won't stop children from interacting with each other.

The two-year study, funded by Centers for Disease Control and Prevention, is also being conducted at Penn State and Utah.

Dubbed the "Social Mixing and Respiratory Transmission in Schools," or SMART Schools study, it is part of a CDC effort to create a national policy on school response to flu and other pandemics.

"There was very little resistance to the project. It was only done with (parents') approval," said Michael Daniels, Canon-McMillan superintendent. "We hope this study will mean fewer illnesses and fewer absences."

Zimmer said researchers "know that children can drive influenza outbreaks, but we don't know how or why. Knowing their interaction and contact patterns will give us much-needed real-world data."

Craig Smith is a staff writer for Trib Total Media. He can be reached at 412-380-5646 or csmith@tribweb.com.

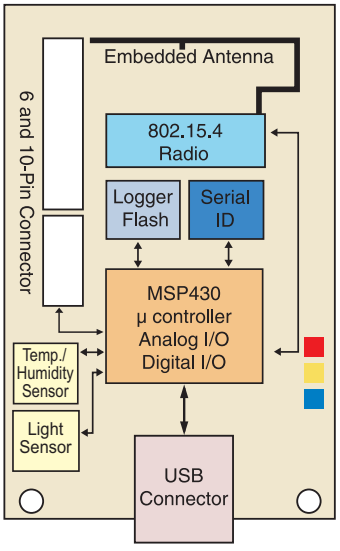
TELOSB

TELOSB MOTE PLATFORM

- IEEE 802.15.4 Compliant
- 250 kbps, High Data Rate Radio
- TI MSP430 Microcontroller with 10kB RAM
- Integrated Onboard Antenna
- Data Collection and Programming via USB Interface
- Open-source Operating System
- Integrated Temperature, Light and Humidity Sensor

Applications

- Platform for Low Power Research Development
- Wireless Sensor Network Experimentation



TPR2420CA Block Diagram

TELOSB

MEMSIC's TelosB Mote TPR2420 is an open-source platform designed to enable cutting-edge experimentation for the research community. The TPR2420 bundles all the essentials for lab studies into a single platform including: USB programming capability, an IEEE 802.15.4 radio with integrated antenna, a low-power MCU with extended memory and an optional sensor suite. TPR2420 offers many features, including:

- IEEE 802.15.4 compliant RF transceiver
- 2.4 to 2.4835 GHz, a globally compatible ISM band
- 250 kbps data rate
- Integrated onboard antenna
- 8 MHz TI MSP430 microcontroller with 10kB RAM
- Low current consumption
- 1MB external flash for data logging
- Programming and data collection via USB
- Sensor suite including integrated light, temperature and humidity sensor
- Runs TinyOS 1.1.11 or higher

The TelosB platform was developed and published to the research community by UC Berkeley. This platform delivers low power consumption allowing for long battery life as well as fast wakeup from sleep state. The TPR2420 is compatible with the open-source TinyOS distribution.

TPR2420 is powered by two AA batteries. If the TPR2420 is plugged into the USB port for programming or communication, power is provided from the host computer. If the TPR2420 is always attached to the USB port no battery pack is needed.

TPR2420 provides users with the capability to interface with additional devices. The two expansion connectors and onboard jumpers may be configured to control analog sensors, digital peripherals and LCD displays.

TinyOS is a small, open-source, energy-efficient software operating system developed by UC Berkeley which supports large scale, self-configuring sensor networks. The source code software development tools are publicly available at: http://www.tinyos.net

Phone: 408.964.9700 | Fax: 408.854.7702 | E-mail: infoca@memsic.com | www.memsic.com

Specifications	TPR2420CA	Remarks
Module		
Processor Performance	16-bit RISC	
Program Flash Memory	48K bytes	
Measurement Serial Flash	1024K bytes	
RAM	10K bytes	
Configuration EEPROM	16K bytes	
Serial Communications	UART	0-3V transmission levels
Analog to Digital Converter	12 bit ADC	8 channels, 0-3V input
Digital to Analog Converter	12 bit DAC	2 ports
Other Interfaces	Digital I/O,I2C,SPI	
Current Draw	1.8 mA	Active mode
	5.1 µA	Sleep mode
RF Transceiver		
Frequency band[1]	2400 MHz to 2483.5 MHz	ISM band
Transmit (TX) data rate	250 kbps	
RF power	-24 dBm to 0 dBm	
Receive Sensitivity	-90 dBm (min), -94 dBm (typ)	
Adjacent channel rejection	47 dB	+ 5 MHz channel spacing
	38 dB	- 5 MHz channel spacing
Outdoor Range	75 m to 100 m	Inverted-F antenna
Indoor Range	20 m to 30 m	Inverted-F antenna
Current Draw	23 mA	Receive mode
	21 µA	Idle mode
	1 µA	Sleep mode
Sensors		
Visible Light Sensor Range	320 nm to 730 nm	Hamamatsu S1087
Visible to IR Sensor Range	320 nm to 1100nm	Hamamatsu S1087-01
Humidity Sensor Range	0-100% RH	Sensirion SHT11
Resolution	0.03% RH	
Accuracy	± 3.5% RH	Absolute RH
Temperature Sensor Range	-40°C to 123.8°C	Sensirion SHT11
Resolution	0.01°C	
Accuracy	± 0.5°C	@25°C
Electromechanical		
Battery	2X AA batteries	Attached pack
User Interface	USB	v1.1 or higher
Size (in)	2.55 x 1.24 x 0.24	Excluding battery pack
(mm)	65 x 31 x 6	Excluding battery pack
Weight (oz)	0.8	Excluding batteries
(grams)	23	Excluding batteries

TPR2420 with Sensor Suite

Notes

[1]Programmable in 1 MHZ steps, 5 MHz steps for compliance with IEEE 802.15.4/D18-2003.

Specifications subject to change without notice

Ordering Information

Model	Description
TPR2420CA	IEEE 802.15.4 TelosB Mote with Sensor Suite

Supporting Information

Salathé et al. 10.1073/pnas.1009094108

SI Methods

Data Collection. ***General.*** On January 14, 2010, we distributed wireless senor network motes (TelosB; Crossbow Technologies Inc.) to all students, teachers, and staff at an American high school (the date was chosen because it represented a typical school day). Participants were asked to sign an assent form on which they indicated at what time the mote was turned on. The assent form also asked participants to indicate their role/status at the school, with the following four options available: "student," "teacher," "staff," and "other." At the end of the day, we collected the motes and assent forms and then obtained data with the corresponding assent from 789 motes/individuals. Some of the motes had not been used (because of people either being absent from the school or not participating in the project), and we did not obtain written assent to use the data for some motes with data. The remaining data cover 94% of the entire school population. We also deployed motes at fixed locations (stationary motes), but these are not part of the dataset described here except for one stationary mote in the main cafeteria; the signal of this mote was used to reconstruct the global timestamp (see below).

Deployment details. Motes were distributed in batches (with an average of ∼11 motes) the night before the deployment and handed out to participants starting around 6:00 AM (with the vast majority of participants receiving and activating their motes on arrival at 8:00 AM). Participants were asked to put their mote in a thin plastic pouch attached to a lanyard (provided by us) and to wear the lanyard around the neck, with the mote being located in front of the chest at all times. The participants handed the motes back to us when leaving the school or at the end of the school day (the vast majority was received between 4:00 and 4:30 PM). The technical details regarding code design, signal strength considerations, and other issues have been described elsewhere (1); however, briefly, each participant's mote was programmed to broadcast beacons at −16.9 dBm at a regular 20-s interval; the packet included the sender's local sequence number. On receiving a beacon, the mote checked the received signal strength indicator (RSSI) value of the packet. Note that the motes are always scanning; thus, no interactions with a duration of at least 20 s will be missed. If the signal strength was lower than −80 dBm, the packet was discarded (this decision was based on experimental data showing that when subjects were facing each other, packets within 3 m had an RSSI value of roughly −80 dBm or above; packets sent when one subject was facing the other person's back had a lower RSSI value (1) (Fig. S4). Otherwise, the receiver created a contact entry, consisting of the sender's ID and beacon sequence number as well as the local mote's sequence number and the RSSI value of the packet.

TelosB motes have a 1-MB flash memory in which interactions can be stored, thus eliminating the need to broadcast interactions to any other external hardware for storage. As a consequence, interactions between subjects can be captured anywhere on the campus of the school, an area of more than 45,000 m^2.

Reconstructing the full contact network required a global timestamp, relative to which all interactions between subjects occurred. Local sequence numbers in each data trace acted as relative clocks, and they could be used as offsets from one stationary mote (the "master stationary mote," located in the main cafeteria), which would be the master clock providing global time. Packets originating from this mote were transmitted at high power (−11 dBm) and were not subject to RSSI filtering at the receiver. More than 90% of mobile motes had received one or more beacons from the master stationary mote. For these mobile motes, we calculated the offset between the master and local sequence numbers. In addition, we created a table of offsets to serve as a lookup table, which included mobile motes as well as other stationary motes. To process data traces from mobile motes that did not hear directly from the master stationary mote, we used the offsets table transitively to compute a timestamp from another mote that already had its global time.

After processing the raw data, we thus obtained a list of interactions that contains 762,868 unique interactions between two motes for a duration of n consecutive beacon intervals (Datasets S1, S2, and S3). Because beacons are broadcast every 20 s, the number of beacons can be used as an approximate measure of contact duration (such that duration in minutes was approximately $n/3$).

This project was approved by the Stanford University Institutional Review Board on July 24, 2009.

1. Kazandjieva M, et al. (2010) Experiences in measuring a human contact network for epidemiology research. *HotEmNets '10: Proceedings of the ACM Workshop on Hot Topics in Embedded Networked Sensors,* (Association for Computing Machinery, Killarney, Ireland).

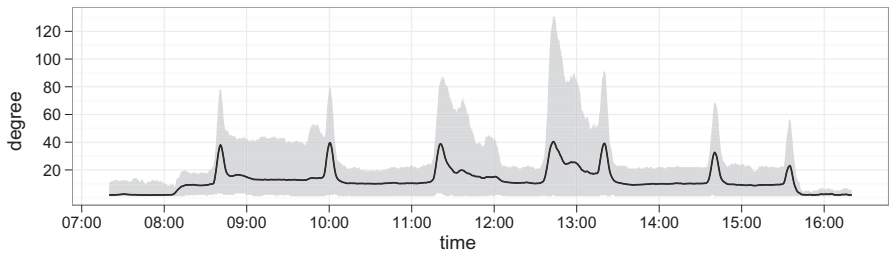


Fig. S1. Temporal dynamics of the average number of contacts (degree). Here, the degree of an individual is measured as the number of other individuals in close proximity during 5 min. Gray background spans the 2.5% and 97.5% percentiles of the degree distribution.

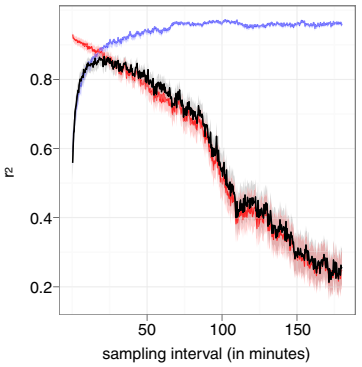


Fig. S2. Squared correlation (r^2) between outbreak size and degree of index case (black), outbreak size and strength of index case (red), and degree and strength of index case (blue) at various sampling rates. The left-most correlations are based on the full dataset (sampling interval of 1/3 min), and all others are based on subsampled datasets that would have been generated with the given sampling interval. The shaded area behind the line shows the 95% confidence interval of squared correlation.

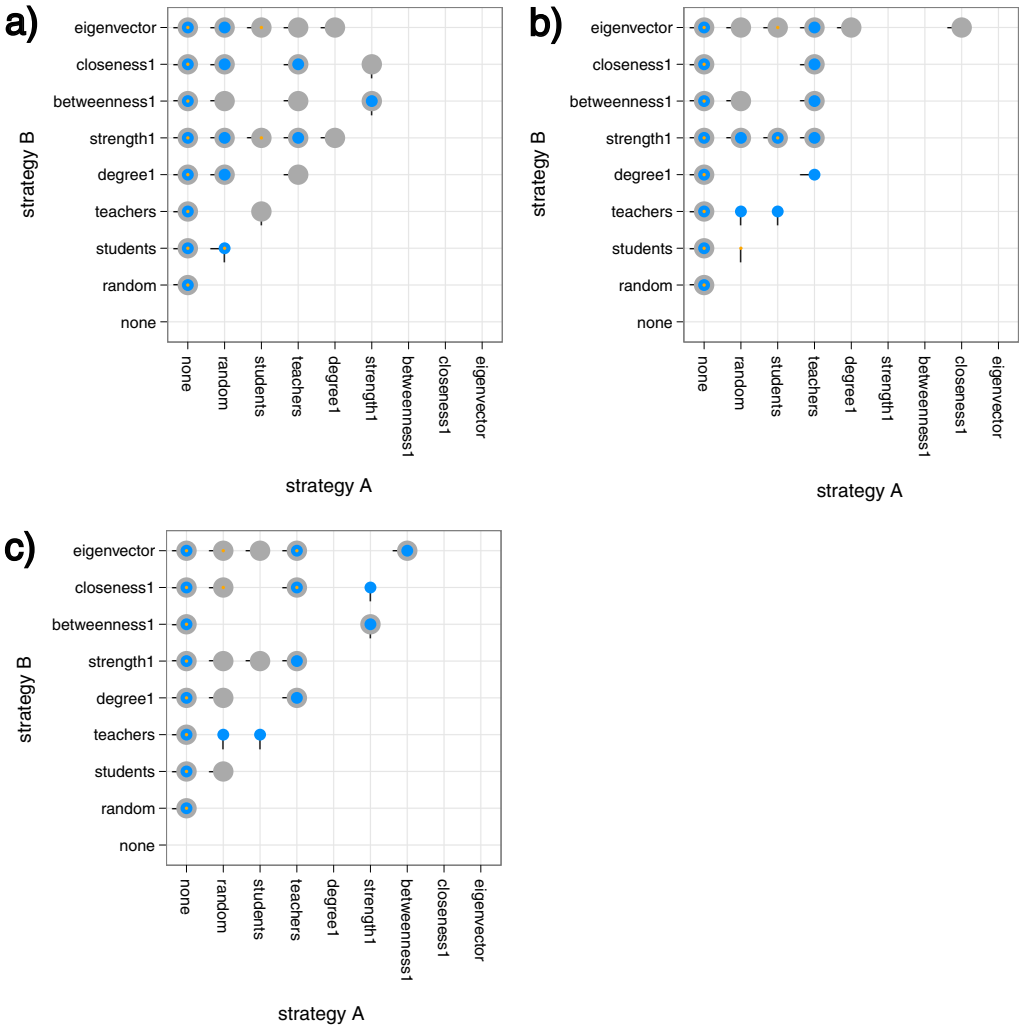


Fig. S3. Settings identical to those described in Fig. 4*B*, but the results are separated according to transmission probabilities per CPR used [0.002 (*A*), 0.003 (*B*), and 0.0045 (*C*)].

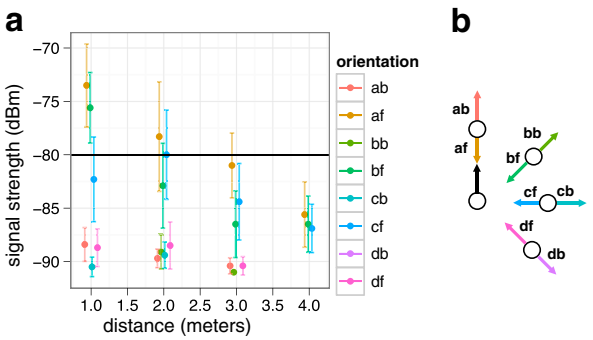


Fig. S4. Dependency of signal strength on distance (1, 2, 3, and 4 m), orientation (a, b, c, and d forward or backward), and angle (0°, 45°, 90°, and 135°). The black horizontal line shows the threshold value that was chosen for the data collection. (*A*) Points show the average signal strength, and bars represent the SD of a particular measurement. Some settings lack data because no packets were received. (A slight horizontal offset was added to the data points for visual clarity.) (*B*) Spatial setting of the four angles and two directions are shown, with reference to the main mote.

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)
Dataset S3 (TXT)

A high-resolution human contact network for infectious disease transmission

Marcel Salathé[a,1,2], Maria Kazandjieva[b], Jung Woo Lee[b], Philip Levis[b], Marcus W. Feldman[a], and James H. Jones[c,d]

Departments of [a]Biology, [b]Computer Sciences, and [c]Anthropology, and [d]Woods Institute for the Environment, Stanford University, Stanford, CA 94305-5020

Edited by Adrian Raftery, University of Washington, Seattle, WA, and approved November 8, 2010 (received for review June 25, 2010)

The most frequent infectious diseases in humans—and those with the highest potential for rapid pandemic spread—are usually transmitted via droplets during close proximity interactions (CPIs). Despite the importance of this transmission route, very little is known about the dynamic patterns of CPIs. Using wireless sensor network technology, we obtained high-resolution data of CPIs during a typical day at an American high school, permitting the reconstruction of the social network relevant for infectious disease transmission. At 94% coverage, we collected 762,868 CPIs at a maximal distance of 3 m among 788 individuals. The data revealed a high-density network with typical small-world properties and a relatively homogeneous distribution of both interaction time and interaction partners among subjects. Computer simulations of the spread of an influenza-like disease on the weighted contact graph are in good agreement with absentee data during the most recent influenza season. Analysis of targeted immunization strategies suggested that contact network data are required to design strategies that are significantly more effective than random immunization. Immunization strategies based on contact network data were most effective at high vaccination coverage.

disease dynamics | network topology | public health | human interactions

Pandemic spread of an infectious disease is one of the biggest threats to society because of the potentially high mortality and high economic costs associated with such an event (1, 2). Understanding the dynamics of infectious disease spread through human communities will facilitate the development of much needed mitigation strategies (3). Schools are particularly vulnerable to infectious disease spread because of the high frequency of close proximity interactions (CPIs) that most infectious disease transmission depends on (3, 4). Infections that are transmitted predominantly via the droplet route, such as influenza, common colds, whooping cough, severe acute respiratory syndrome (SARS), and many others, are among the most frequent infectious diseases. Droplets from an infected person can reach a susceptible person in close proximity, typically a distance of less than 3 m (5, 6), making CPIs highly relevant for disease spread. Very little is known about the dynamic patterns of CPIs in human communities, however [but see Cattuto et al. (7)]. Here, we present data collected with a wireless sensor network deployment using TelosB motes (Crossbow Technologies Inc.) (8) to detect high-resolution proximity (up to 3 m) between subjects in a U.S. high school. The dataset represents a high-resolution temporal contact network relevant to the spread of infectious diseases via droplet transmission in a school.

Previous attempts to capture the contact networks relevant for infectious disease transmission have mostly been based on data collection using surveys, sociotechnological networks, and mobile devices like cell phones. Each of these approaches has advantages and disadvantages. Surveys manage to capture the interactions relevant for disease transmission but are often limited by small sample sizes (9) and are subject to human error (10). Sociotechnological networks can provide large long-term datasets (11) but fail to capture the CPIs relevant for disease transmission. The use of mobile devices aware of their location (or of other mobile devices in proximity) represents a promising third alternative. Using mobile phones to detect spatial proximity of subjects is possible with repeated Bluetooth scans (10), but the resolution is too coarse for diseases that are transmitted through the close contact route. Our approach is free of human error, captures the vast majority (94%) of the community of interest, and allows us to collect high-resolution contact network data relevant for infectious disease transmission.

Most efforts to understand and mitigate the spread of pandemic diseases (influenza in particular) have made use of large-scale spatially explicit models parameterized with data from various sources, such as census data, traffic/migration data, and demographic data (3, 4, 12–15). The population is generally divided into communities of schools, workplaces, and households, but detailed data on mixing patterns in such communities are scarce. In particular, very little is known about the contact networks in schools (16) even though schools are known to play a crucially important role in pandemic spread, mainly owing to the intensity of CPIs at schools. In what follows, we describe and analyze the contact network observed at a U.S. high school during a typical school day. Using an SEIR (susceptible, exposed, infectious, and recovered) simulation model, we investigate the spread of influenza on the observed contact network and find that the results are in very good agreement with absentee data from the influenza A (H1N1) spread in the fall of 2009. Finally, we implement and test various immunization strategies to evaluate their efficacy in reducing disease spread within the school.

Results

The dataset covers CPIs of 94% of the entire school population, representing 655 students, 73 teachers, 55 staff, and 5 other persons, and it contains 2,148,991 unique close proximity records (CPRs). A CPR represents one close (maximum of 3 m) proximity detection event between two motes. An interaction is defined as a continuous sequence (≥1) of CPRs between the same two motes, and a contact is the sum of all interactions between these two motes. Thus, a contact exists between two motes if there is at least one interaction between them during the day, and the duration of the contact is the total duration of all interactions between these two motes. Because the beaconing frequency of a mote is 0.05 s^{-1}, an interaction of length 3 (in CPRs) corresponds to an interaction of about 1 min (*SI Text* and references therein). The entire dataset consists of 762,868 interactions with a mean duration of 2.8 CPRs (~1 min), or 118,291 contacts with mean duration of 18.1 CPRs (~6 min)

Author contributions: M.S., M.K., J.W.L., P.L., M.W.F., and J.H.J. designed research; M.S., M.K., J.W.L., and P.L. performed research; M.S. analyzed data; and M.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]Present address: Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA 16802.

[2]To whom correspondence should be addressed. E-mail: salathe@psu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1009094108/-/DCSupplemental.

(data available in *SI Methods*). Fig. 1*A* shows the frequency, *f*, of interactions and contacts of length *m* (in minutes) [*f*(*m*)]. The majority of interactions and contacts are very short (80th percentile of interactions at 3 CPRs, 80th percentile of contacts at 15 CPRs), and even though about 80% of the total time is spent in interactions that are shorter than 5 min, short contacts (<5 min) represent only about 10% of the total time (Fig. 1*B*).

The temporal mixing patterns observed are in accordance with the schedule of the school day [i.e., the average degree (number of contacts) peaks between classes and during lunch breaks] (Fig. S1). The aggregate network for the entire day can be represented by a weighted undirected graph, wherein nodes represent individuals and edges represent contacts (edges are weighted by contact duration). The topology of the contact network is an important determinant of infectious disease spread (17, 18). Traditional infectious disease models assume that all subjects have the same number of contacts, or that the contact network of subjects is described by a random graph with a binomial degree distribution. Many networks from a wide range of applications, including contact networks relevant for infectious disease transmission (19, 20), have been found to have highly heterogeneous degree distributions, however. Such heterogeneity is important because it directly affects the basic reproductive number, R_0, a crucially important indicator of how fast an infectious disease spreads and what fraction of the population will be infected. In particular, if ρ_0 is the incorrect estimate for R_0 in a heterogeneous network under the false assumption of a uniform degree distribution, the correct estimate is given by $R_0 = \rho_0 (1 + CV^2)$, where CV^2 is the squared coefficient of variation of the degree distribution (17, 21). Thus, the *CV* quantifies the extent to which contact heterogeneity affects disease dynamics.

The descriptive statistics of the school network with different definitions of contact are shown in Fig. 2. To account for the fact that the majority of the contacts are relatively short (Fig. 1*A*), we recalculated all statistics of the network with a minimum requirement for contact duration, c_m (i.e., all edges with weight $<c_m$ are removed from the graph). The network exhibits typical "small-world" properties (22), such as a relatively high transitivity (also known as clustering coefficient, which measures the ratio of triangles to connected triplets) and short average path length for all values of c_m. Assortativity, the tendency of nodes to associate with similar nodes with respect to a given property (23), was measured with respect to degree and role of the person (e.g., student, teacher). Interestingly, although both measures are relatively high, degree assortativity decreases and role assortativity increases with higher values of c_m. Because of the very high density of the contact network, a giant component exists for all values of c_m. Community structure (or modularity) is relatively high, increasingly so with higher values of c_m, indicating that more intense contacts tend to occur more often in subgroups and less often between such groups (24). We find a very homogeneous degree distribution with a $CV^2 = 0.118$ for the full network and slightly increased heterogeneity in the network with higher cutoff values c_m (Fig. 2*J*). The distributions of number of interactions, *c*, and the strength, *s* (the weighted equivalent of the degree) (25) are equally homogeneous (Fig. 3). Overall, the data suggest that the network topology is best described by a low-variance small-world network.

To understand infectious disease dynamics at the school, we used an SEIR simulation model (parameterized with data from influenza outbreaks; details presented in *SI Methods*), wherein an index case becomes infected outside of the school on a random day during the week and disease transmission at the school occurs during weekdays on the full contact network as described by the collected data. Each individual is chosen as an index case for 1,000 simulation runs, resulting in a total of 788,000 epidemic simulation runs. This simulation setting represents a base scenario, wherein a single infectious case introduces the disease into the school population. In reality, multiple introductions are to be expected if a disease spreads through a population, but the base scenario used here allows us to quantify the predictive power of graph-based properties of individuals on epidemic outcomes. We assume that symptomatic individuals remove themselves from the school population after a few hours. We find that in 67.7% of all simulations, no secondary infections occur and thus there is no outbreak, whereas in the remaining 32.3% of the simulations, outbreaks occur with an average attack rate of 3.87% (all simulations = 1.33%, maximum = 46.19%) and the average R_0, measured as the number of secondary infections caused by the index case, is 3.85 (all simulations = 1.24, maximum = 18). Recent work on disease spread on networks has identified the relationship between R_0, the network degree distribution, and the average probability that an infectious individual transmits the disease to a susceptible individual, T (18, 26). Based on this, R_0 would be valued at 4.52 (*SI Methods*). This value is higher than what we measure in the simulations because it is based on the assumption of continuous transmission, whereas the simulations exhibit discontinuous transmission attributable to weekends; during that time, the school is closed and the chain of transmission is effectively cut for 2 d. Finally, absentee data from the school during the fall of 2009 (i.e., during the second wave of H1N1 influenza in the northern hemisphere) are in good agreement with simulation data generated by the SEIR model running on the contact network (Fig. 4*A*).

A strong correlation exists between the size of an outbreak caused by index case individual *i* and the strength of the node representing individual *i* ($r^2 = 0.929$). The correlation between outbreak size and degree is substantially weaker ($r^2 = 0.525$) because at the high temporal resolution of the dataset, the de-

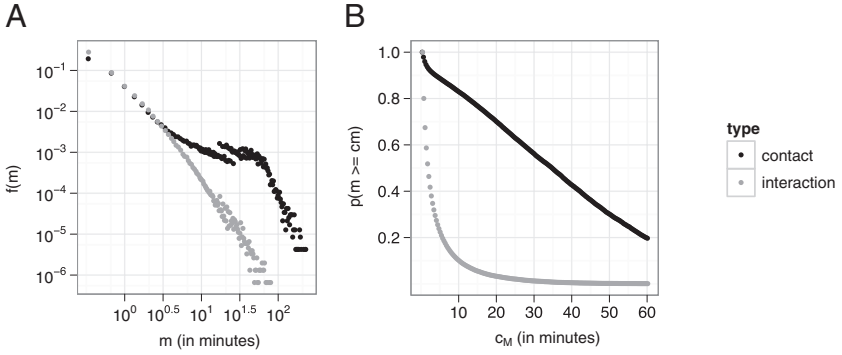


Fig. 1. (*A*) Normalized frequency, *f*, of interactions and contacts of duration *m* (in minutes) [*f*(*m*)] on a log-log scale. (*B*) Percentage, *p*, of total time of all CPIs by interactions and contacts with a minimum duration, c_m (in minutes). Most CPI time is spent in medium-duration contacts consisting of repeated short interactions.

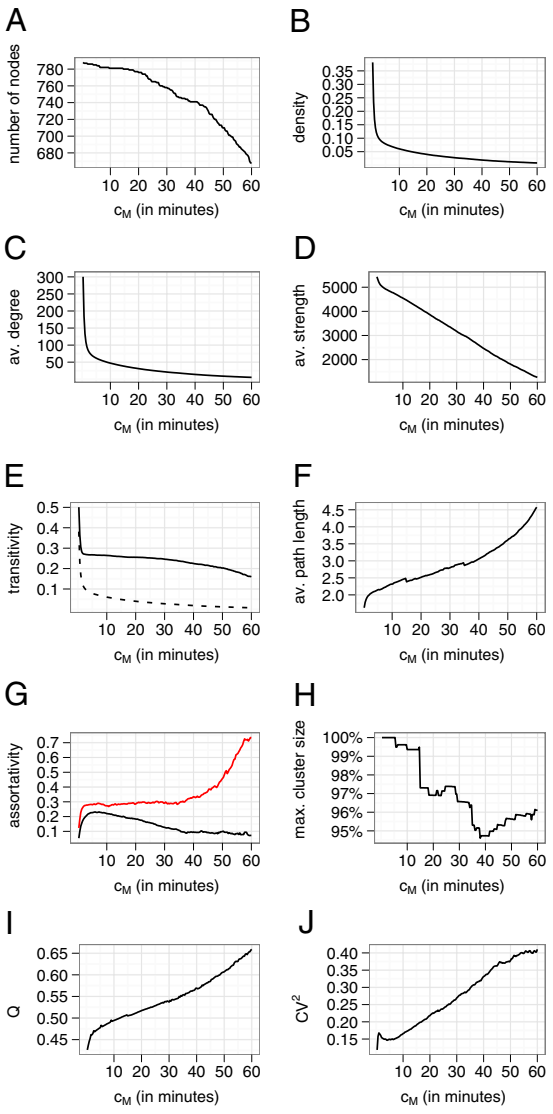


Fig. 2. Various statistics on the contact graph with minimum contact duration, c_m (i.e., the left-most point in each panel represents the full contact graph, the right-most point represents the contact graph that contains only contacts that are at least 60 min long). With increasing c_m, nodes drop out of the network if they have no contact that satisfies the minimum duration condition. (*A*) Hence, the reduction in the number, *V*, of nodes. (*B*) Density of the graph (*2E/[V*(V − 1)]*), where *E* is the number of edges. (C) Average (av.) degree. (*D*) av. strength, where the strength of a node is the total number of CPRs of the node. (*E*) Transitivity (i.e., cluster coefficient) as defined by Barrat et al. (25) and expected value (mean degree/*V*) in a random network (dashed line). (*F*) Average path length. (*G*) Assortativity (23) with respect to degree (black line) and role (red line). (*H*) Size of the largest component as a fraction of total network size. max., maximum. (*I*) Modularity, *Q*, as defined by Reichardt and Bornholdt (39). (*J*) CV^2 of degree.

gree contains many short-duration contacts whose impact on epidemic spread is minimal. To estimate the sampling rate at which degree has maximal predictive power, we systematically subsampled our original dataset to yield lower resolution datasets. Fig. S2 shows that sampling as infrequently as every 100 min would have resulted in the same predictive power for degree as sampling every 20 s, whereas the maximum predictive power for degree would have been attained at ∼20 min. At this sampling rate, the 95% confidence intervals for the correlation between degree and outbreak size and the correlation between strength and outbreak size start to overlap (because of the high correlation between degree and strength; Fig. S2, blue line). These results suggest that high-resolution sampling of network properties such as the degree of nodes might be highly misleading for prediction purposes if used in isolation (i.e., without the temporal information that allows for weighting).

To mitigate epidemic spread, targeted immunization interventions or social distancing interventions aim to prevent disease transmission from person to person. Finding the best immunization strategy is of great interest if only incomplete immunization is possible, as is often the case at the beginning of the spread of a novel virus. In recent years, the idea of protecting individuals based on their position in the contact network has received considerable attention (11, 27, 28). Graph-based properties, such as node degree and node betweenness centrality (29), have been proposed to help identify target nodes for control strategies, such as vaccination; however, because of the lack of empirical contact data on closed networks relevant for the spread of influenza-like diseases, such strategies could only be tested on purely theoretical networks [or on approximations from other empirical social networks that did not measure CPIs directly (11)]. To understand the effect of partial vaccination, we measured outbreak size for three different levels of vaccination coverage (5%, 10%, and 20%) and a number of different control strategies based on node degree, node strength, betweenness centrality, closeness centrality, and eigenvector centrality (so-called "graph-based strategies"). In addition, we tested vaccination strategies that do not require contact network data (random vaccination, preferential vaccination for teachers, and preferential vaccination for students; *SI Methods*). To ensure robustness of the results to variation in transmission probabilities, all simulations were tested with three different transmission probabilities (*Methods*). Ten thousand simulations for each combination of vaccination strategy, vaccination coverage, and transmission probability with a random index case per simulation were recorded (i.e., total of 810,000 simulations) to assess the effect of vaccination. Fig. 4*B* shows which strategies led to significantly ($P < 0.05$, two-sided Wilcoxon test) different outcomes at all transmission probability values (results separated by transmission probability are presented in Fig. S3). As expected, all strategies managed to reduce the final size of the epidemic significantly. Compared with the random strategy, graph-based strategies had an effect only at higher vaccination coverage. Graph-based strategies did not differ much in their efficacy; in general, strategy-based strategies were the most effective. Overall, two main results emerge: (*i*) in the absence of information on the contact network, all available strategies, including random immunization, performed equally well and (*ii*) in the presence of information on the contact network, high-resolution data support a strategy-based strategy, but there was no significant difference among the graph-based strategies.

Discussion

In summary, we present high-resolution data from the CPI network at a U.S. high school during a typical school day. Notably, the month of the experiment (January) is associated with the second highest percentage of influenza cases in the United States for the 1976–1977 through 2008–2009 influenza seasons (second only to February). The data suggest that the network relevant for disease transmission is best described as a small-world network with a very homogeneous contact structure in which short repeated interactions dominate. The low values of the coefficients of variation in degree, strength, and number of interactions (Fig. 3) suggest that the assumption of homogeneity in traditional disease models (21) might be sufficiently realistic for simulating the spread of influenza-like diseases in communities like high schools. Furthermore, we do not find any "fat tails" in the contact distribution of our dataset, corroborating the notion (9) that the current focus on networks with such distributions is not warranted for infectious disease spread within local communities.

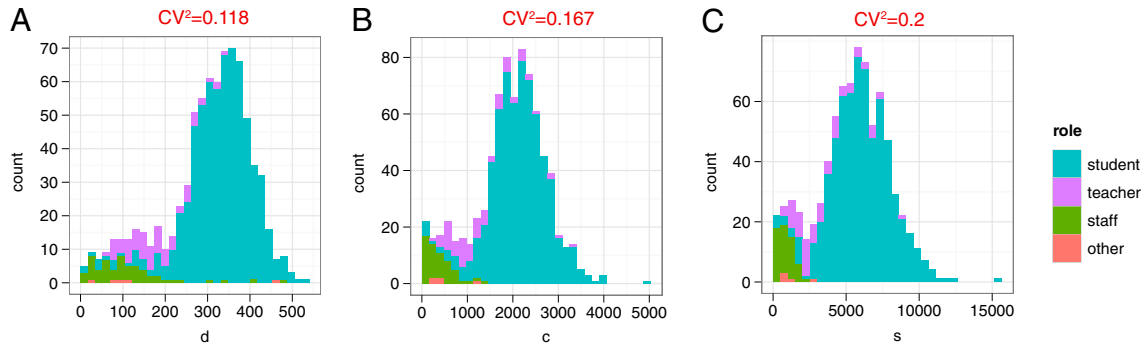


Fig. 3. Distribution and CV^2 of degree, *d* (*A*); number of interactions, *c* (*B*); and strength, *s* (*C*), based on the full contact network and colored by the role of individuals.

It is important to recognize the limitations of the data presented here, particularly in light of the fact that transmission of influenza-like diseases also occurs via other routes, for example, via contact with contaminated surfaces (30). Moreover, different pathogens as well as different strains of a particular pathogen might have different minimum requirements (both spatial and temporal) that need to be met for person-to-person transmission. At present, the data capture the contact network during a single day only. This is not an inherent shortcoming of the approach presented here, however, and long-term studies in the future could address how the large-scale structure of the contact network in a high school changes over time. Data collection at different schools with different demographic compositions would be helpful in clarifying if and how demographic compositions affect the properties of the network relevant for disease transmission. Wireless sensor network technology certainly allows further elucidation of the contact networks not only at different schools but in households, hospitals, workplaces, and other community settings.

With regard to immunization strategies, our simulation results suggest that contact network data are necessary to design strategies that are significantly more effective than random immunization to minimize the number of cases at the school caused by a single index case. Great care needs to be taken in interpreting these results for various reasons. First, the limitations of the data as discussed above mean that these results may not hold in other settings, underlining the need for further empirical network studies. Second, the simulations assume neither multiple introductions nor ongoing interactions of participants outside of the school. To what extent these assumptions, particularly the latter, are violated when a disease spreads through a community is unknown and remains to be measured. Third, future work needs to establish the robustness of the effect of vaccination strategies against errors in the measurement of graph-based properties. Fourth, and perhaps most importantly, a particular immunization strategy may be optimal for reducing the number of cases in one particular school but, at the same time, may not be optimal from the perspective of an entire community. Immunization strategies must also take into account medical, social, and ethical aspects (31). Thus, although we believe that data of the kind reported here can help to inform public health decisions, particularly as more data become available in the future, it is clear that one cannot derive public health recommendations at this stage directly from this study alone. We note, however, that our findings support the notion that graph-based immunization strategies could, in principle, help to mitigate disease outbreaks (11, 28).

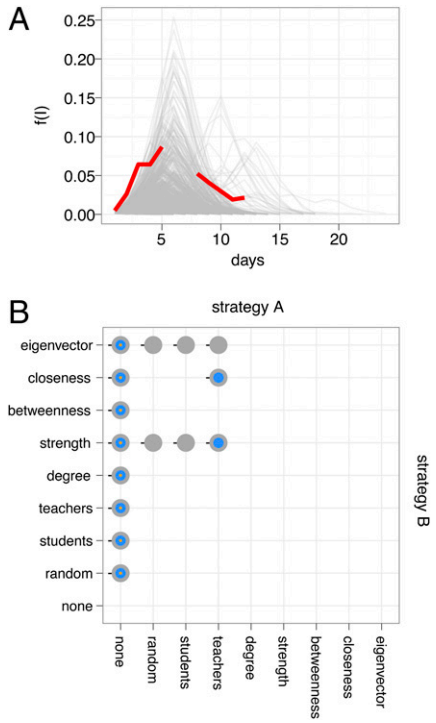


Fig. 4. (*A*) Absentee data (red) and data generated by the SEIR model (gray; 1,000 runs with R_0 >1 shown). Gray lines show frequency of infectious individuals, f(I); red lines show the combined frequency of students who reported, or were diagnosed with, a fever and teachers who were absent (gap in the line attributable to weekend). (*B*) Differences in effect of vaccination strategies. Colors represent vaccination coverage of 5% (orange), 10% (blue), and 20% (gray). A point at the intersection of strategy A and strategy B indicated that between those strategies, there was a significant difference ($P < 0.05$, two-sided Wilcoxon test) in the outbreak size at all transmission probability values at the given vaccination coverage. A black horizontal or vertical line points in the direction of the strategy that resulted in smaller outbreak sizes. Because of the symmetry of the grid, data points below the left bottom and top right diagonal line are not shown.

Methods

The mote deployment is described in detail in *SI Methods*.

Epidemic Simulations. To simulate the spread of an influenza-like disease, we used an SEIR simulation model parameterized with data from influenza outbreaks (12, 32, 33). In the following, we describe the model in detail.

Transmission occurs exclusively along the contacts of the graph as collected at the school. Each individual (i.e., node of the network) can be in one of four classes: susceptible, exposed, infectious, and recovered. Barring vaccination, all individuals are initially susceptible (more information on vaccination is presented below). At a random time step during the first week of the simulation, an individual is chosen as the index case and his or her status is set to exposed. A simulation is stopped after the number of both exposed and infectious individuals has gone back to 0 (i.e., all infected individuals have

recovered). Each time step represents 12 h and is divided into day and night. Transmission can occur only during the day and only on weekdays (i.e., apart from the initial infection of the index case, we do not consider any transmission outside of the school; although this assumption will not hold in reality, it allows us to focus exclusively on within-school transmission and to analyze the spread of a disease starting from a single infected case).

Transmission of disease from an infectious to a susceptible individual occurs with a probability of 0.003 per 20 s of contact (the interval between two beacons). This value has been chosen because it approximates the time-dependent attack rate observed in an outbreak of influenza aboard a commercial airliner (32). In particular, the probability of transmission per time step (12 h) from an infectious individual to a susceptible individual is $1 - (1 - 0.003)^w$, where w is the weight of the contact edge (in CPRs). On infection, an individual will move into the exposed class (infected but not infectious). After the incubation period, an exposed individual will become symptomatic and move into the infectious class. The incubation period distribution is modeled by a right-shifted Weibull distribution with a fixed offset of half a day [power parameter = 2.21, scale parameter = 1.10 (12)]. On the half day that the individual becomes infectious, the duration of all contacts of the infectious individual is reduced by 75%. This reduction ensures that if an individual becomes symptomatic and starts to feel ill during a school day, social contacts are reduced and the individual leaves the school or is dismissed from school after a few hours. In the following days, all contacts are reduced by 100% until recovery (i.e., the individual stays at home). Once an individual is infectious, recovery occurs with a probability of $1 - 0.95^t$ per time step, where t represents the number of time steps spent in the infectious state [in line with data from an outbreak of H1N1 at a New York City school (33)]. After 12 d in the infectious class, an individual will recover if recovery has not occurred before that time.

Based on these simulation settings and the finding that the average contact duration is 18.1 CPRs (*Results*), the transmissibility, T, as defined by Newman (18) and Meyers et al. (26), is $1 - (1 - 0.003)^{18.1*0.25} = 0.0135$. Furthermore, based on the framework established by Newman (18) and Meyers et al. (26), R_0 can be calculated as $R_0 = T \times \langle k_e \rangle$, where the average excess degree, $\langle k_e \rangle$, is $\langle k^2 \rangle / \langle k \rangle - 1 = 334.76$.

We assume that all exposed individuals developed symptoms. A high incidence of asymptomatic spread may affect infectious disease dynamics (34), but reports of asymptomatic individuals excreting high levels of influenza virus are rare (35). In addition, a recent community-based study investigating naturally acquired influenza virus infections found that only 14% of infections with detectable shedding at RT-PCR were asymptomatic and viral shedding was low in these cases (36), suggesting that the asymptomatic transmission plays a minor role. Similar patterns were observed for SARS-CoV, another virus with the potential for rapid pandemic spread: Asymptomatic cases were infrequent, and lack of transmission from asymptomatic cases was observed in several countries with SARS outbreaks (37).

Vaccination. The efficacy of vaccination strategies was tested by simulation. Vaccination occurs (if it occurs at all) before introduction of the disease by the index case. Vaccinated individuals are moved directly into the recovering class. We assume that the vaccine provides full protection during an epidemic.

Three vaccination strategies are implemented that do not require measuring graph-based properties; these strategies are called "random," "students," and "teachers."

Random. Individuals are chosen randomly until vaccination coverage is reached.

Students. Students only are chosen randomly until vaccination coverage is reached.

Teachers. Teachers only are chosen randomly until vaccination coverage is reached. If vaccination coverage is so high that all teachers get vaccinated before the coverage is reached, the strategy continues as the student strategy (see above) for the remaining vaccinations.

Five vaccination strategies are implemented that require measuring graph properties: These strategies are called "degree," "strength," "betweenness," "closeness," and "eigenvector." In all cases, individuals are ranked according to the specific graph property and chosen according to that ranking (in descending order) until vaccination coverage is reached.

Degree. Degree is calculated as the number of contacts during the day of measurement.

Strength. Strength is calculated as the total time exposed to others during the day of measurement.

Betweenness. Betweenness centrality, $C_B(i)$, of individual i is calculated as

$$C_B(i) = \sum_{s \neq i \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where s, t, and i are distinct individuals in the contact graph; σ_{st} is the total number of shortest paths between nodes s and t; and $\sigma_{st}(i)$ is the number of those shortest paths that go through node i (29). The shortest path is calculated using inverse weights.

Closeness. Closeness centrality, $C_C(i)$, of individual i is calculated as

$$C_C(i) = \frac{n-1}{\sum_{s \neq i} d_{si}}$$

where s and i are distinct individuals in the contact graph, d_{si} is the shortest path between nodes s and i, and n is the number of individuals in the graph (29). The shortest path is calculated using inverse weights.

Eigenvector. Calculation of eigenvector centrality is described by White and Smyth (38) through application of the page-rank algorithm with jumping probability 0. The measure captures the fraction of time that a random walk would spend at a given vertex during an infinite amount of time.

We tested three different levels of vaccination coverage: 5%, 10%, and 20%. These percentages apply to the entire population [i.e., a 10% vaccination coverage means that 10% of the entire school population is vaccinated, independent of the particular vaccination strategy (except for the strategy "none," which means no vaccinations occur]. In addition to the default transmission probability per CPR interval described above (i.e., 0.003), we tested lower (0.002) and higher (0.0045) transmission probability values.

ACKNOWLEDGMENTS. We thank Ignacio Cancino, Elena V. Jordan, Alison Brown, and Rahel Salathé as well as members of the M.W.F., P.L., and J.H.J. groups for help with the mote deployments; Marc Horowitz for providing a crucial link; and two anonymous referees for their valuable comments. We are particularly grateful to the staff members of the school who made this project possible. Special thanks to the creators and maintainers of the Java Universal Network/Graph Framework and the R package iGraph. This research was supported by a National Science Foundation award (BCS-0947132), a Branco Weiss fellowship (to M.S.), National Institute of Child Health and Human Development Award 1K01HD051494 (to J.H.J.), and National Institutes of Health Grant GM28016 (to M.W.F.).

1. Murray CJL, Lopez AD, Chin B, Feehan D, Hill KH (2006) Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918-20 pandemic: A quantitative analysis. *Lancet* 368:2211–2218.
2. Meltzer MI, Cox NJ, Fukuda K (1999) The economic impact of pandemic influenza in the United States: Priorities for intervention. *Emerg Infect Dis* 5:659–671.
3. Halloran ME, et al. (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *Proc Natl Acad Sci USA* 105:4639–4644.
4. Yang Y, et al. (2009) The transmissibility and control of pandemic influenza A (H1N1) virus. *Science* 326:729–733.
5. Fiore A, et al. (2008) Prevention and control of influenza: Recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep* 59(RR-8): 1–62, and erratum (2010) 59:1147.
6. Xie X, Li Y, Chwang AT, Ho PL, Seto WH (2007) How far droplets can move in indoor environments—Revisiting the Wells evaporation-falling curve. *Indoor Air* 17: 211–225.
7. Cattuto C, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* 5:e11596.
8. Polastre J, Szewczyk R, Culler D (2005) Telos: Enabling ultra-low power wireless research. *IPSN '05: Proceedings of the Fourth International Symposium on Information Processing in Sensor Networks*, (IEEE Press, Los Angeles).
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. Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106:15274–15278.
11. Salathé M, Jones JH (2010) Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol* 6:e1000736.
12. Ferguson NM, et al. (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437:209–214.
13. Ferguson NM, et al. (2006) Strategies for mitigating an influenza pandemic. *Nature* 442:448–452.
14. Longini IM, Jr., et al. (2005) Containing pandemic influenza at the source. *Science* 309: 1083–1087.
15. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. *Nature* 439:462–465.
16. Glass L, Glass R (2008) Social contact networks for the spread of pandemic influenza in children and teenagers *BMC Public Health* 8:61.
17. May RM (2006) Network structure and the biology of populations. *Trends Ecol Evol* 21:394–399.
18. Newman ME (2002) Spread of epidemic disease on networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 66:016128.

19. Liljeros F, Edling CR, Amaral LA, Stanley HE, Aberg Y (2001) The web of human sexual contacts. *Nature* 411:907–908.
20. Jones JH, Handcock MS (2003) Social networks: Sexual contacts and epidemic thresholds. *Nature* 423:605–606, discussion 606.
21. Anderson RM, May RM (1991) *Infectious Diseases of Humans, Dynamics and Control* (Oxford Science Publications, Oxford, UK).
22. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440–442.
23. Newman ME (2003) Mixing patterns in networks.. *Phys Rev E Stat Nonlin Soft Matter Phys* 67(Pt 2):026126.
24. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826.
25. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 101:3747–3752.
26. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC (2005) Network theory and SARS: Predicting outbreak diversity. *J Theor Biol* 232:71–81.
27. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355–359.
28. Chen YP, Paul G, Havlin S, Liljeros F, Stanley HE (2008) Finding a better immunization strategy. *Phys Rev Lett* 101:058701.
29. Freeman LC (1978) Centrality in social networks—Conceptual clarification. *Soc Networks* 1:215–239.
30. Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M (2007) Transmission of influenza A in human beings. *Lancet Infect Dis* 7:257–265.
31. Medlock J, Galvani AP (2009) Optimizing influenza vaccine distribution. *Science* 325: 1705–1708.
32. Moser MR, et al. (1979) An outbreak of influenza aboard a commercial airliner. *Am J Epidemiol* 110:1–6.
33. Lessler J, et al.; New York City Department of Health and Mental Hygiene Swine Influenza Investigation Team (2009) Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. *N Engl J Med* 361:2628–2636.
34. King AA, Ionides EL, Pascual M, Bouma MJ (2008) Inapparent infections and cholera dynamics. *Nature* 454:877–880.
35. Influenza Team, European Centre for Disease Prevention and Control (2007) Influenza transmission: Research needs for informing infection control policies and practice. *Euro Surveill* 12:E070510.
36. Lau LL, et al. (2010) Viral shedding and clinical illness in naturally acquired influenza virus infections. *J Infect Dis* 201:1509–1516.
37. Wilder-Smith A, et al. (2005) Asymptomatic SARS coronavirus infection among healthcare workers, Singapore. *Emerg Infect Dis* 11:1142–1145.
38. White S, Smyth P (2003) Algorithms for estimating relative importance in networks. *International Conference on Knowledge Discovery and Data Mining*,KDD '03 Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining (Association for Computing Machinery, New York), pp 266–275.
39. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys* 74:016110.

BMC Medicine

BioMed Central
The Open Access Publisher

A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks

BMC Medicine 2013, **11**:35 doi:10.1186/1741-7015-11-35

Timo Smieszek (timo.smieszek@daad-alumni.de)
Marcel Salathe (salathe@psu.edu)

ISSN 1741-7015

Article type Research article

Submission date 18 June 2012

Acceptance date 22 November 2012

Publication date 12 February 2013

Article URL http://www.biomedcentral.com/1741-7015/11/35

A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks

Timo Smieszek* and Marcel Salathé

Center for Infectious Disease Dynamics (CIDD), Department of Biology, The Pennsylvania State University, University Park, PA 16803, USA.

*Corresponding author

Email addresses:

TS: smieszek@psu.edu

MS: salathe@psu.edu

Abstract

Background

Infectious disease outbreaks in communities can be controlled by early detection and effective prevention measures. Assessing the relative importance of each individual community member with respect to these two processes requires detailed knowledge about the underlying social contact network on which the disease can spread. However, mapping social contact networks is typically too resource-intensive to be a practical possibility for most communities and institutions.

Methods

Here, we describe a simple, low-cost method - called collocation ranking - to assess individual importance for early detection and targeted intervention strategies that are easily implementable in practice. The method is based on knowledge about individual collocation which is readily available in many community settings such as schools, offices, hospitals, and so on. We computationally validate our method in a school setting by comparing the outcome of the method against computational predictions based on outbreak simulations on an empirical high-resolution contact network. We compare collocation ranking to other methods for assessing the epidemiological importance of the members of a population. To this end, we select subpopulations of the school population by applying these assessment methods to the population and adding individuals to the subpopulation according to their individual rank. Then, we assess how suited these subpopulations are for early detection and targeted intervention strategies.

Results

Likelihood and timing of infection during an outbreak are important features for early detection and targeted intervention strategies. Subpopulations selected by the collocation ranking method

show a substantially higher average infection probability and an earlier onset of symptoms than randomly selected subpopulations. Furthermore, these subpopulations selected by the collocation ranking method were close to the optimum.

Conclusions

Our results indicate that collocation ranking is a highly effective method to assess individual importance, providing critical low-cost information for the development of sentinel surveillance systems and prevention strategies.

See related commentary article here http://www.biomedcentral.com/1741-7015/11/36

Keywords: Sentinel surveillance; prevention; social network; influenza; collocation; SIR model.

Background

Social network analysis has become an important tool to assess infectious disease spread in communities [1-3]. In a social network model of disease spread, the network ties between individuals are considered to be relevant for the transmission of the disease whose spread is modeled. For many infectious diseases, including some of the diseases with the greatest pandemic potential, such as influenza, disease transmission is assumed to require spatio-temporal proximity of individuals. Spatio-temporal proximity is typically approximated by network ties, where it is assumed that social contacts (family, friends, co-workers, and so on) capture the majority of potential disease transmission events.

The predictive power of social network topology on the dynamics of infectious disease spread has been confirmed empirically [1,3-6]. It is well established that both high individual contact rates as well as a high dispersion in a population's contact distribution results in increased disease transmission within the population [7-9]. In many host-pathogen systems, it is a minority of individuals that contributes the most to infectious disease spread [8]. In addition, the position of individuals in a social network has been shown to correlate well with the likelihood and timing of infectious disease onset [1,3,10].

These findings immediately suggest an important role for social network analysis in the development of sentinel surveillance systems and targeted mitigation strategies. Sentinel surveillance is most efficient when disease outbreaks can be detected as early as possible. Mitigation strategies such as targeted vaccination are most efficient when each unit of resource (for example, vaccination dose) leads to the maximal case count reduction possible. Combined, these two methods can significantly mitigate infectious disease spread and thus reduce the morbidity and mortality associated with the disease.

Unfortunately, mapping social networks is very resource-intensive, and thus generally not a practical option in most communities. However, most communities do have information about the location of their members over time. For example, educational communities such as schools have detailed information about the location of their members in the form of rosters and schedules. From these readily available data sources, one can calculate the overall collocation time of each community member, that is, the cumulative time each individual is potentially exposed to other individuals. On a population-level, time-use surveys have been shown to be good proxies for contact data [11]. We suggest that also on a detailed community-level such a collocation measure can serve as a very good proxy indicator of the network measures that are associated with both increased infection likelihood and early infection during an outbreak. As a consequence, this method has the potential to be a simple, low-cost method to assess individual importance for early detection and targeted intervention strategies that are easily implementable in practice without the need to map social networks.

In this paper, we test how well collocation ranking can identify subpopulations suited for early detection and targeted intervention strategies. We compare the performance of the collocation ranking method (as defined by two benchmarks) to the performance achieved by other, partly network-based, methods. We further compare its performance to randomly selected subpopulations and the best possible subpopulations according to the two benchmarks.

Methods

We challenge various indicators for selecting subpopulations for early detection and targeted intervention with computational influenza outbreak simulations that are based on empirical high-

resolution contact and location data collected with wireless sensor technology at a US high school.

First, we describe the data that were used for our analyses. Then, we define two benchmarks according to which the proposed collocation ranking method and the other indicators are evaluated. Next, we describe all indicators that are tested in this paper. Finally, we describe the outbreak simulation model and how the performance tests were set up.

Both the empirical data and the simulation model are described in detail elsewhere [12]. Therefore, both are only described briefly here.

All simulations and analyses were coded in and executed by Python (Version 2.7.2, 32-bit, Enthought Python Distribution). Figures were created with R (Version 2.13.0) and the ggplot2 library.

Ethics statement

The data collection was approved by the Stanford University Institutional Review Board on 24 July 2009. Written informed consent was obtained from all participants.

Contact and location data

The data that we use in this paper were collected at a US high school during one school day with wireless sensor technology. A total of 789 individuals or 94% of the school population, including students, teachers, and staff, participated in the study. The participants wore small wireless sensors (so-called motes) that detect and log radio signals broadcast by other nearby motes. We refer to the motes that were worn by participants as mobile motes. Further, stationary motes were attached to fixed locations throughout the school campus to keep track of the participants' locations. As a consequence, the dataset contains two types of records. Close proximity interactions (CPIs) are records that indicate two participating individuals standing face-to-face

with a distance of less than three meters at a certain point in time. Location records are records that indicate the presence of an individual nearby a stationary mote (location information is at the level of rooms). A detailed description on how information and noise were separated in the data is provided in the supplementary material (see Additional file 1). Data were collected at time intervals of 20 seconds.

Schedule data

In many communities, full individual schedule data (that is, the schedule of each individual) is readily available to community health services. During an outbreak, this data could be used to calculate the collocation rank for all community members. For various reasons, it was not possible to obtain full individual schedule data from the school, but it was possible to obtain aggregated schedule data. We then reconstructed individual schedule data from a combination of the mote data and the aggregated schedule data. The aggregated schedule data file contains the following information about each class taught at the school during the mote deployment: (i) who taught the class, (ii) the room in which the class was taught, (iii) the period of the class, and (iv) the number of students who were signed up for the class. The aggregated schedule data combined with the high resolution location data allows us to reconstruct individual schedules with high fidelity. The algorithm for matching aggregated schedule and individual location data is further described in the supplementary material (see Additional file 1).

Benchmarks

The core idea is to identify simple indicators that allow the identification of subpopulations of the entire school population that are maximally relevant in prevention and surveillance efforts. Both prevention and surveillance efforts should target individuals who are more likely to become infected than others. In addition, these efforts should be targeted at individuals who become

infected early during an outbreak, allowing for early detection of outbreaks (surveillance) and early response (prevention).

We define two simple benchmarks to test the accuracy of any indicator to be evaluated:

1. The first benchmark B_1 is the average probability $\overline{P_i}$ of the individuals i of a subpopulation S to become infected. We use an empirical probability P_i that is defined as the ratio of the number of simulation runs n in which individual i in subpopulation S became infected and the total number of simulation runs N. Note that simulation runs in which i is the index case are ignored (see test setting section below). A subpopulation has been optimally identified if B_1 is maximal.

2. For the second benchmark, we calculate the ratio $\overline{t_i} / P_i$ for every individual i in subpopulation S, where $\overline{t_i}$ is the average simulation time step at which the individual became symptomatic. Then, the second benchmark B_2 is defined as the average of these ratios. The division by P_i is necessary to take into account that early detection of a symptomatic individual is more relevant when the infection probability of that individual is high. The time of the onset of symptoms has more practical relevance than the time of infection, because symptomatic cases can be identified more easily than pre- or asymptomatic carriers. A subpopulation has been optimally identified if B_2 is minimal.

Rank indicators

Several indicators are evaluated with respect to their ability to select subpopulations with optimal benchmark results. Thus, a good indicator would select subpopulations that have high B_1 values and low B_2 values. The basic principle of subpopulation formation is the same for all indicators: the individuals are ranked according to their individual respective indicator value (from high to

low values). Then, subpopulations are formed by selecting individuals from high to low ranks until the target subpopulation size is reached. We use the following rank indicators:

Presence

The presence indicator measures the total time an individual attends classes according to the schedule, and it is defined as $\sum_{p=1}^{7} t(p) \cdot T(p,i)$, where p is an index pointing to one of the seven periods of the surveyed high school day, $t(p)$ is the official duration of period p, and $T(p,i) = 1$ if individual *i* had a scheduled class during period p, and $T(p,i) = 0$ if not.

Collocation

The collocation indicator measures the cumulative time each individual is potentially exposed to other individuals during classes, and it is defined as $\sum_{p=1}^{7} t(p) \cdot \omega(p,i)$. Here, $\omega(p,i)$ denotes the number of students signed up for the class that individual *i* is taking during period p, and $t(p)$ is the official duration of period p. If i has no class during that period, $\omega(p,i) = 0$. The collocation indicator - like the presence indicator - is only based on schedule data.

Degree

We use the actor degree centrality $C_D(i)$ [13], which is one of the network indices that is frequently used in network epidemiology to identify the most important individuals in a transmission network [12,14-18]. The actor degree centrality of an individual i is defined as the number of contact partners of i - here determined by the presence of at least one CPI - during the measurement period.

Degree (>10 minutes)

The difference between this indicator and the previous one is that only contacts of more than 10 minutes of accumulated duration during the measurement period are considered. The cut-off of 10 minutes was chosen arbitrarily, but a sensitivity analysis shows that the indicator's performance changes only slightly when the cut-off is changed to 5, 15, or 20 minutes (see Additional file 1).

Strength

The strength of an individual i stands for the cumulative contact duration of i, and it is defined as $\sum_{j \in J^{\Delta}\{i\}} w(i,j)$. Thereby, $J^{\Delta}\{i\}$ is the set containing the entire school population except i, and $w(i,j)$ stands for the accumulated contact duration of individuals i and j. Strength is an enhancement of the degree concept and can be interpreted as a weighted degree [19].

There are other network measures which are frequently used to identify pivotal individuals in a social network, for instance closeness centrality or betweenness centrality. These measures, however, have been shown to be comparably good or even worse than the degree in indicating individuals who are important for disease spread [17,18]. For this reason we concentrate on the simpler, but still powerful, centrality indicators described above.

Model of influenza spread

We use an individual-based model of influenza spread to assess the importance of the members of the school population with respect to disease spread. The model is published and described in detail in [12], but briefly, we assume that the infection is introduced by one index case at the beginning of a simulation run and that all subsequent infections happen within the school population, that is, there are no further introductions from outside. The time step duration is half a day. The model is a susceptible, exposed, infectious, recovered (SEIR)-type model. The

probability to switch from the susceptible to the exposed state is $1-(1-0.003)^{w}$, where w is the accumulated contact time the susceptible individual has spent with infectious individuals while at school (in CPI records) [20]. The duration of the exposed state follows a Weibull distribution with an offset of half a day; the power parameter is 2.21, the scale parameter is 1.10 [21]. After that period in the exposed state, every individual will be in the infectious state for exactly one time step before turning into home confinement and, finally, recovering. To allow for the fact that the onset of influenza symptoms is typically sudden and that affected individuals will be dismissed quickly, we reduce the duration of contacts by 75% during the single time step at school.

Test setting

Each member of the school population could be the index case of an outbreak and introduce the infection from outside the school. Therefore, we initialize 100 independent runs for each member of the school population being the index case that introduces the infection. This results in a total of 78,900 simulation runs that build the basis of our analyses.

For all simulation runs, we keep track of which individuals got infected and when they became symptomatic during the course of the simulation run. This allows us to calculate the two benchmarks defined above.

Results

In order to assess the performance of the collocation indicator, we selected subpopulations of various sizes on the basis of the collocation indicator and compared their benchmarks to randomly selected subpopulations, optimal subpopulations, and subpopulations selected on the basis of the other indicators described in the Methods section. An optimal subpopulation stands

for a subpopulation selected in such a way that it achieves the best possible benchmark value for the given population.

First benchmark: average infection probability

The subpopulations that were selected on the basis of the collocation indicator constantly show a substantially higher average infection probability B_1 than randomly selected subpopulations of the same size (Figure 1a). Given a subpopulation of ten percent of the entire school population, collocation ranking resulted in 1.43 to 1.62 times better results than randomly composed subpopulations that were between the 10th and the 90th percentile. Given a subpopulation of twenty percent of the entire school population, collocation ranking resulted in 1.29 to 1.41 times better results than randomly composed subpopulations that were between the 10th and the 90th percentile.

Most subpopulations selected on the basis of rank indicators achieved consistently better benchmark results than random subpopulations, and all of them outcompeted random composition over a large range of subpopulation sizes. The performance of subpopulations selected on the basis of the collocation indicator was better than the performance of subpopulations selected on the basis of the presence and the degree indicator, but worse than the performance of subpopulations selected on the basis of the degree (>10 minutes) and the strength indicator (Figure 1b). For subpopulations smaller than 40% of the entire population, those selected on the basis of the collocation indicator achieved benchmark values that were only about 10% below the optimum.

Second benchmark: ratio of average infection time and probability

The qualitative picture for the second benchmark was very similar to that of the first benchmark. However, differences between the various subpopulations were more pronounced.

For subpopulations that represent between 2% and 90% of the entire population, the subpopulation selected on the basis of the collocation indicator performed consistently 2.5 or more times better than the median of the random subpopulations. For almost the entire range of subpopulation sizes, the benchmarks of the subpopulation selected on the basis of the collocation indicator were at least twice as low as the benchmarks of 90% of the random subpopulations (Figure 2a).

For small subpopulations up to approximately 22%, collocation ranking outcompeted ranking by degree. Further, collocation ranking was almost always better or as good as ranking by presence (Figure 2b).

Further analyses

Additional analyses, in particular how the role of members of the school population (that is, whether the individual is a student, a teacher, or a staff member) is related to individual importance, are provided in Additional file 1.

Discussion

Social networks have proven to be useful for understanding and predicting infectious disease dynamics. There is a discussion on how detailed network data must be in order to be useful in epidemiological applications [6,14,22]. However, even mapping low-detail social contact networks is typically too resource-intensive to be a practical possibility for most communities and institutions. What is needed instead are low-cost proxies for individual network properties that can serve as epidemiological predictors. Spatial distance measures, for example, have recently been found to be significant predictors of social ties (among other predictors) [23], and it is therefore reasonable to expect that spatial proxies can also serve as useful epidemiological

predictors. The collocation ranking method presented here is based on spatio-temporal considerations, and our results suggest that it may effectively identify subpopulations suited for sentinel surveillance systems and prevention strategies.

Current methods to identify subpopulations for sentinel surveillance systems and prevention strategies typically rely on demographic variables such as age (for example, children and young adults in influenza surveillance systems [24-26]) and geographic location (for example, administrative units in invasive meningococcal disease surveillance systems [27-29]). These methods work because there is sufficient variance of such demographic variables at the societal level. However, at the level of communities and institutions such as schools, there is often too little variance to make these methods applicable. Furthermore, because demographic variables are not direct proxies for transmission routes, they may fail to identify individuals with high transmission potential who fall outside of the targeted range of the demographic variable. In contrast, the collocation ranking indicator proposed here is a direct proxy of potential disease transmission events as given by the contact network.

Random selection serves as a null model method in the absence of epidemiologically relevant information about a population. The collocation ranking method significantly outcompetes the random method. As expected, some network indicators, such as the strength, were able to outcompete the collocation ranking method to identify subpopulations for early detection or targeted intervention strategies. This is not surprising because strength is essentially a direct measure of exposure, and it can thus serve as an indicator that can identify subpopulations which are almost identical to the optimal subpopulation. Nevertheless, measuring strength is resource-intensive, while collocation ranking is not.

Our research is not without limitations. The first limitation is that we rely on widely used computational simulation models of disease spread, rather than validating our method in an empirical setting. Our simulation model is based on high-resolution contact network data [12] as well as established disease transmission parameters [20,21], but ideally, any benchmark would be based on empirical outbreak data instead of simulated data. However, infection transmission is a highly stochastic process, requiring multiple outbreaks for a robust evaluation of the collocation ranking method presented above.

Limitations and uncertainties of our model are, in particular, the following: (i) There is still debate on the relative importance of the different potential pathways of influenza transmission [30-32]. Most models of influenza spread assume transmission by close contact, but there is the possibility that other transmission pathways are more important than currently thought. (ii) We model the spread between members of the school population during school hours, but we do not capture potentially infectious contacts between school members during their leisure time. (iii) We assumed that the probability of being an index case is homogeneous. In reality, this is most likely not the case. (iv) We also assumed that all individuals are fully susceptible. In reality, individuals differ in their serostatus and (partial) immunity is linked to patterns of previous exposure. (v) It might be that an ongoing epidemic changes the contact behavior not only of the symptomatic individuals, but also of the healthy ones who continue to attend school. Such potential behavior changes are not reflected in our model.

Another limitation is that the data to test our method were collected in one school only. Moreover, the data covers only one school day. While the method worked very well in this setting, the generalizability to other settings remains to be established.

Finally, we had to reconstruct individual schedules from aggregated schedules and mote data. Reconstructions may be incomplete (compare with Additional file 1), and the real course of a school day may differ from the scheduled sequence of classes. While it is important to recognize that we currently cannot conclusively validate our method, our simulation results indicate that the collocation method is an effective, low-cost tool that warrants further research.

Conclusions

Social networks have proven to be useful predictors of infectious disease outbreak dynamics. From a practical perspective, social network information can be highly valuable for the development of sentinel surveillance systems and prevention strategies because people's positions within the network correlate with their likelihood and timing of infection during an outbreak. The disadvantage of network-based approaches is that they are highly resource-intensive and, thus, can not be applied to every situation of interest. Hence, simple proxies, such as the collocation ranking method presented here, that fulfill the same purpose are needed. Subpopulations identified by the collocation ranking method are significantly better suited for sentinel surveillance systems and prevention strategies than randomly selected subpopulations. Some network-based ranking methods were slightly better for identifying such subpopulations than collocation ranking. The collocation ranking method, however, is a low-cost method that still manages to identify subpopulations that are very close to the optimum. The data requirement of the method is very low, and typically readily available in many community settings, such as schools, offices, hospitals, and so on in the form of rosters/schedules.

Our results suggest that the collocation ranking method may effectively identify subpopulations suited for sentinel surveillance systems and prevention strategies.

Abbreviations

CPI, close proximity interaction; SEIR, susceptible, exposed, infectious, recovered.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TS and MS conceived and designed the study. MS collected the data. TS performed the analyses. TS and MS contributed to the writing of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

This research was supported by a fellowship from the German Academic Exchange Service DAAD to Timo Smieszek (grant D/10/52328) and a Branco Weiss Fellowship to Marcel Salathé. Simulations were run on a computer cluster that was funded by the National Science Foundation through grant OCI-0821527 and the data collection was funded by the National Science Foundation through grant BCS-0947132. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We thank Victoria Barclay whose valuable comments helped us to improve the quality of the supplementary material. We are also grateful for the contributions of Maria Kazandjieva, Jung

Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. Finally, we thank the four reviewers for their thorough review and their valuable suggestions.

References

1. Christakis NA, Fowler JH: **Social network sensors for early detection of contagious outbreaks.** *PLoS ONE* 2010, **5:**e12948.

2. Gardy JL, Johnston JC, Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *New Engl J Med* 2011, **364:**730–739.

3. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow D, the Pennsylvania N1N1 working group: **Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza.** *Proc Natl Acad Sci U S A* 2011, **108:**2825–2830.

4. Goeyvaerts N, Hens N, Ogunjimi B, Aerts M, Shkedy Z, van Damme P, Beutels P: **Estimating infectious disease parameters from data on social contacts and serological status.** *J Roy Statist Soc Ser C* 2010, **59:**255–277.

5. Kretzschmar M, Teunis PFM, Pebody RG: **Incidence and reproduction numbers of pertussis: estimates from serological and social contact data in five European countries.** *PLoS Med* 2010, **7:**e1000291.

6. 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.

7. Woolhouse MEJ, Shaw DJ, Matthews L, Liu WC, Mellor DJ, Thomas MR: **Epidemiological implications of the contact network structure for cattle farms and the 20-80 rule.** *Biol Lett* 2005, **1:**350–352.

8. Woolhouse ME, Dye C, Etard JF, Smith T, Charlwood JD, Garnett GP, Hagan P, Hii JLK, Ndhlovu PD, Quinnell RJ, Watts CH, Chandiwana SK, Anderson RM: **Heterogeneities in the transmission of infectious agents: implications for the design of control programs.** *Proc Natl Acad Sci U S A* 1997, **94:**338–342.

9. Pastor-Satorras R, Vespignani A: **Epidemic spreading in scale-free networks.** *Phys Rev Lett* 2001, **86:**3200–3203.

10. Rea E, Laflèche 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.

11. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, Edmunds WJ: **Using time-use data to parameterize models for the spread of close-contact infectious diseases.** *Am J Epidemiol* 2008, **168:**1082–1090.

12. Salathé 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.

13. Wasserman S, Faust K: *Social Network Analysis.* Cambridge: Cambridge University Press; 1994:178.

14. Stehle J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L, Régis C, Pinton J-F, Khanafer N, van den Broeck W, Vanhems P: **Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees.** *BMC Med* 2011, **9:**87.

15. 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.

16. Meyers LA, Newman MEJ, Pourbohloul B: **Predicting epidemics on directed contact networks.** *J Theor Biol* 2006, **240:**400–418.

17. Christley R, Pinchbeck G, Bowers R, Clancy D, French N, Bennett R, Turner J: **Infection in social networks: using network analysis to identify high-risk individuals.** *Am J Epidemiol* 2005, **162:**1024–1031.

18. Bell D, Atkinson J: **Centrality measures for disease transmission networks.** *Soc Networks* 1999, **21:**1–21.

19. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A: **The architecture of complex weighted networks.** *Proc Natl Acad Sci U S A* 2004, **101:**3747–3752.

20. Moser MR, Bender TR, Margolis HS, Noble GR, Kendal AP, Ritter DG: **An outbreak of influenza aboard a commercial airliner.** *Am J Epidemiol* 1979, **110:**1–6.

21. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS: **Strategies for containing an emerging influenza pandemic in Southeast Asia.** *Nature* 2005, **437:**209–214.

22. Blower S, Go M-H: **The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy?** *BMC Med* 2011, **9:**88.

23. Sailer K, McCulloh I: **Social networks and spatial configuration—How office layouts drive social interaction.** *Soc Networks* 2012, **34:**47–58.

24. Poehling KA, Edwards KM, Weinberg GA, Szilagyi P, Staat MA, Iwane MK, Bridges CB, Grijalva CG, Zhu Y, Bernstein DI, Herrera G, Erdman D, Hall CB, Seither R, Griffin MR: **The underrecognized burden of influenza in young children.** *New Engl J Med* 2006, **355:**31–40.

25. Brownstein JS, Kleinman KP, Mandl KD: **Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system.** *Am J Epidemiol* 2005, **162:**686–693.

26. Iwane MK, Edwards KM, Szilagyi PG, Walker FJ, Griffin MR, Weinberg GA, Coulen C, Poehling KA, Shone LP, Balter S, Hall CB, Erdman DD, Wooten K, Schwartz B: **Population-based surveillance for hospitalizations associated with respiratory syncytial virus, influenza virus, and parainfluenza viruses among young children.** *Pediatrics* 2004, **113:**1758–1764.

27. Bertrand S, Van Meervenne E, De Baere T, Vanhoof R, Collard J-M, Ruckly C, Taha M, Carion F: **Detection of a geographical and endemic cluster of hyper-invasive meningococcal strains.** *Microbes Infect* 2011, **13:**684–690.

28. Reinhardt M, Elias J, Albert J, Frosch M, Harmsen D, Vogel U: **EpiScanGIS: an online geographic surveillance system for meningococcal disease.** *Int J Health Geogr* 2008, **7:**33.

29. Centers for Disease Control and Prevention: **Control and prevention of serogroup C meningococcal disease: evaluation and management of suspected outbreaks: recommendations of the Advisory Committee on Immunization Practices (ACIP).** *MMWR* 1997, **46:**13–21.

30. Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M: **Transmission of influenza A in human beings.** *Lancet Infect Dis* 2007, **7:**257–265.

31. Tellier R: **Review of aerosol transmission of influenza A virus.** *Emerg Infect Dis* 2006, **12:**1657–1662.

32. Teunis PF, Brienen N, Kretzschmar ME. **High infectivity and pathogenicity of influenza A virus via aerosol and droplet transmission.** *Epidemics* 2010, **2:**215–222.

Figure 1. Performance of collocation ranking: first benchmark. Subfigures 1**a** and 1**b** are based on the first benchmark, which is the average probability of individuals in a given subpopulation to become infected during an outbreak, B_1. The abscissa shows the percentage of the population selected for prevention or surveillance efforts. The ordinate shows the ratio of the B_1 of the collocation indicator and the B_1 of any other indicator, that is, ordinate values >1 indicate that the collocation indicator performs better than the other indicator it is compared to. Subfigure 1a compares the B_1 value of the 10th, 25th, 50th, 75th, and 90th percentile of 100,000 randomly selected subpopulations to the B_1 of subpopulations selected by the collocation indiciator. Subfigure 1b compares B_1 of all indicators defined in the Methods section, as well as the optimal B_1, to the B_1 of the collocation indicator.

Figure 2. Performance of collocation ranking: second benchmark. Subfigures 2**a** and 2**b** are based on the second benchmark. The abscissa shows the percentage of the population selected for prevention or surveillance efforts. The ordinate shows the ratio of the B_2 of a given indicator and the B_2 of the collocation indicator, that is, ordinate values >1 indicate that the collocation indicator performs better than the other indicator it is compared to. Subfigure 2a compares the

B_2 value of the 10th, 25th, 50th, 75th, and 90th percentile of 100,000 randomly selected subpopulations to the B_2 of subpopulations selected by the collocation indicator. Subfigure 2b compares B_2 of all indicators defined in the Methods section, as well as the optimal B_2, to the B_2 of the collocation indicator.

Additional files

Additional file 1

Title: Supplementary information.

Description: This additional file contains further information on (i) the data collection, (ii) how the locations of study participants were derived from the data, and (iii) how the individual schedules of students and teachers were reconstructed. The file further provides supplementary analyses which are not included in the main document. In particular, it contains figures that show (i) how well the five indicators define subpopulations according to a third benchmark (the average time to the onset of symptoms), (ii) how sensitive the outcome of the degree indicator reacts to various contact duration cut-offs, (iii) how predictive the role of an individual is for the likelihood and timing of infection, (iv) what the relationship between the five indicators is, and (v) how well the collocation indicator captures the number of infections that are induced by a certain index case.

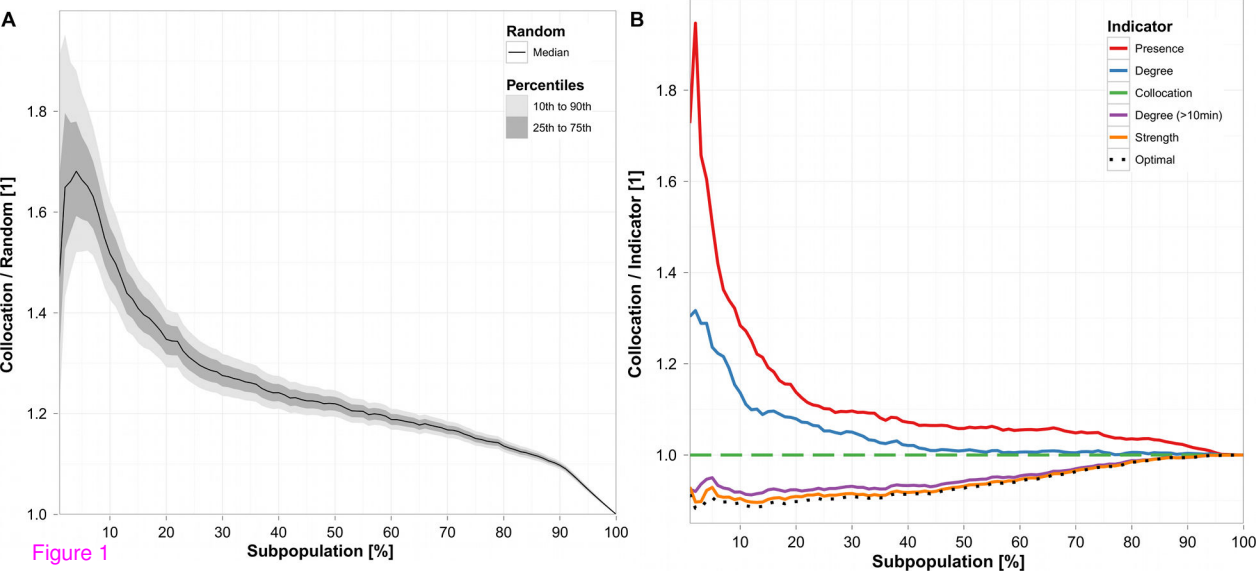

A
Random
Median
Percentiles
10th to 90th
25th to 75th
Collocation / Random [1]
1.8
1.6
1.4
1.2
1.0
10
20
30
40
50
60
70
80
90
100
Subpopulation [%]
B
Indicator
Presence
Degree
Collocation
Degree (>10min)
Strength
Optimal
Collocation / Indicator [1]
1.8
1.6
1.4
1.2
1.0
10
20
30
40
50
60
70
80
90
100
Subpopulation [%]

Figure 1

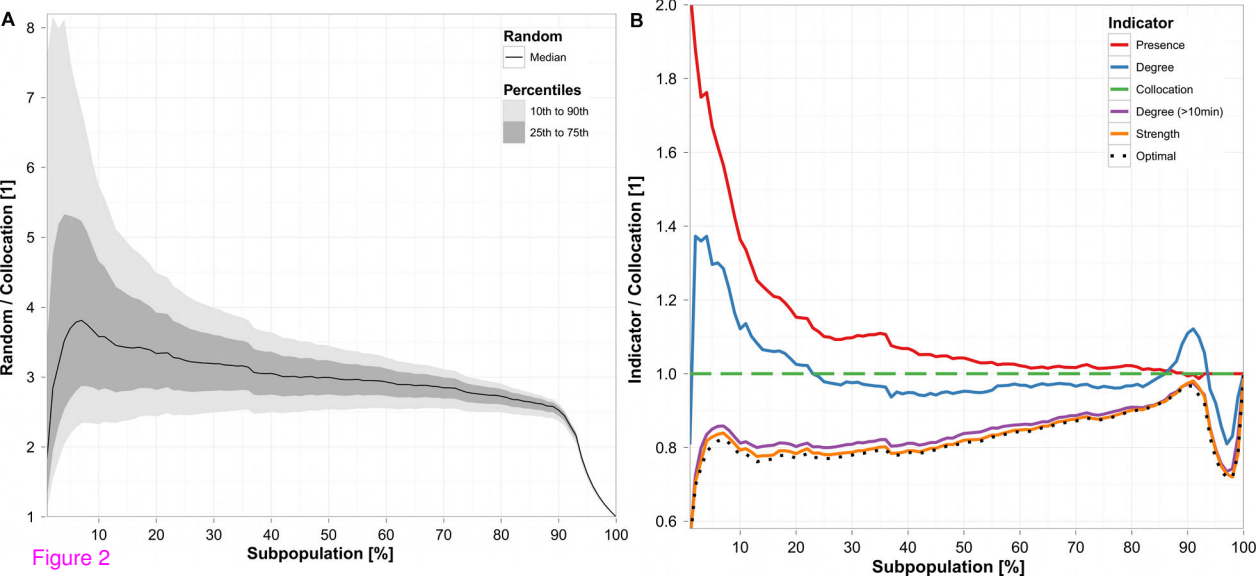

A
Random
Median
Percentiles
10th to 90th
25th to 75th
Random / Collocation [1]
Subpopulation [%]
B
Indicator
Presence
Degree
Collocation
Degree (>10min)
Strength
Optimal
Indicator / Collocation [1]
Subpopulation [%]

Figure 2

Additional files provided with this submission:

Additional file 1: sup1.pdf, 1061K
http://www.biomedcentral.com/imedia/1064245977866075/supp1.pdf

Epidemiol. Infect., Page 1 of 9. © Cambridge University Press 2011
doi:10.1017/S0950268811001130

Collecting close-contact social mixing data with contact diaries: reporting errors and biases

T. SMIESZEK*, E. U. BURRI, R. SCHERZINGER AND R. W. SCHOLZ

ETH Zurich, Institute for Environmental Decisions, Natural and Social Science Interface, Zurich, Switzerland

(*Accepted 22 May 2011*)

SUMMARY

The analysis of contact networks plays a major role to understanding the dynamics of disease spread. Empirical contact data is often collected using contact diaries. Such studies rely on self-reported perceptions of contacts, and arrangements for validation are usually not made. Our study was based on a complete network study design that allowed for the analysis of reporting accuracy in contact diary studies. We collected contact data of the employees of three research groups over a period of 1 work week. We found that more than one third of all reported contacts were only reported by one out of the two involved contact partners. Non-reporting is most frequent in cases of short, non-intense contact. We estimated that the probability of forgetting a contact of $\leqslant 5$ min duration is greater than 50%. Furthermore, the number of forgotten contacts appears to be proportional to the total number of contacts.

Key words: Contact diary, direct transmission, epidemiology, networks, respiratory infections.

INTRODUCTION

The topology of contacts in host organisms is known to be an important influencing factor in infectious disease dynamics. It has been argued theoretically that highly connected individuals play a pivotal role in disease spread and that they have a strong impact on both individual risks of infection as well as spread dynamics at the level of entire populations [1–3]. Furthermore, it has been shown that both the clustering of contact partners and repeated contact with the same person can slow down an outbreak compared to the dynamics of an otherwise identical random mixing model [4, 5].

* Author for correspondence: Dr T. Smieszek, ETH Zurich, Institute for Environmental Decisions, Natural and Social Science Interface, CHN J 70.1, Universitaetsstrasse 22, 8092 Zurich, Switzerland.
(Email: timo.smieszek@daad-alumni.de)

Empirical data on host-to-host contacts is needed to complement the theoretical knowledge concerning the importance of network topology for infectious disease dynamics. Methods have been developed to measure potentially contagious contacts in real-world settings. Currently, the dominant approach for measuring epidemiologically relevant contact data is contact diaries [6–12]. Empirical research on potentially contagious contacts, particularly the highly cited study by Mossong *et al.* [8], has influenced the discussion on the patterns and risk factors of disease spread and has informed infectious disease modelling [e.g. 13]. In addition, various studies have shown that empirical contact data can successfully be applied in epidemiological models to replicate serological data [14–16].

Despite the increasing use of diary-based contact data for understanding and explaining infectious disease dynamics, few studies have addressed the quality

and appropriateness of this methodological approach. One study compared retrospective and prospective study designs and found 'only minor differences in the number of contacts, with on average more contacts reported in the prospective survey' [7, p. 133]. Another study compared a web-based mode of data collection with a diary-based one and concluded that the diary-based approach is less demanding and better suited for collecting detailed data than the web-based approach [9]. A similar result was reported in a study that compared paper-based diaries with data collection via personal digital assistants (PDAs) [10]. Here, the classical diaries were also perceived to be easier to use. However, there is still a lack of research that aims to measure errors and biases related to the diary approach *directly*, and not only the differences between variations of the same method.

The goal of our research was to develop a study design that allows the measuring of reporting errors and biases related to contact diaries in a more encompassing and complete manner than previous studies. This paper provides first answers to the questions of (i) how important measurement errors related to the diary method are, (ii) how reporting errors are related to the duration of a contact, and (iii) how reporting errors are related to the total number of different contact partners during a day. Further, we analysed whether the participants showed fatigue during the later study days. We focused solely on contacts that are relevant for the spread of pathogens that are transmitted via direct, non-sexual contact between hosts.

METHODS

Study design and data collection

Typically, diary-based studies are designed as so-called egocentric network studies. That means, the participants are chosen randomly, or using any other appropriate sampling scheme, typically from a large population; the participants (egos) report information about their contact partners (alters), but these alters are not usually participants in the study. Thus, it is not possible to link up the participants of an egocentric network study with each other in order to achieve a complete network structure. Another drawback of the purely egocentric network design is that there are limited possibilities for validating the answers of the participants (e.g. by utilizing the symmetry condition for age-structured contact matrices, as done by Wallinga *et al.* [16]). Consequently, the participants' answers are usually taken for granted.

To overcome the methodological limitations of egocentric network studies and to be able to give answers to the posed research questions, we conducted an empirical network study with a complete network design (i.e. the alters of an ego are also participants in the study, and they can be linked). Our target population consisted of the employees of three research groups belonging to a single institute at ETH Zurich. In total, 50 employees agreed to participate and actually participated in our study. The data collection started on Monday, 17 May 2010, and ended on Friday, 21 May 2010.

The participants of our study were asked to report only potentially contagious contacts they had with other participants of this study. A potentially contagious contact was defined as (i) a conversation held at <2 m distance and with more than ten words spoken, or as (ii) any sort of physical contact with a person. When a contact event in keeping with this definition occurred with any other participant of the study, both involved participants were asked to note the respective alter's name in their diaries and an estimation of the total time of contact during the entire day (in 5-min intervals).

All participants were asked to complete their diaries independently and not to communicate with the other participants about the contents. Thus, if all participants perceived and recalled all contacts correctly, there would be a mirror-inverted – but otherwise totally identical – match for every reported contact in the database. As a consequence, our study design allows investigation of the accuracy with which contact diaries measure potentially contagious contacts, because every deviation from the aforementioned ideal indicates a reporting error.

Analyses of errors and biases

Although the chosen study design allows the investigation of reporting errors in contact diary studies, even this design results in unidentified contacts whenever both involved participants do not report a common contact that actually took place. However, with few assumptions it is possible to approximate the number of completely unreported contacts as well as the probability of reporting a contact or of forgetting to report a contact in a particular setting. In the following text we present a mathematical approach for doing so, and describe how we assess the

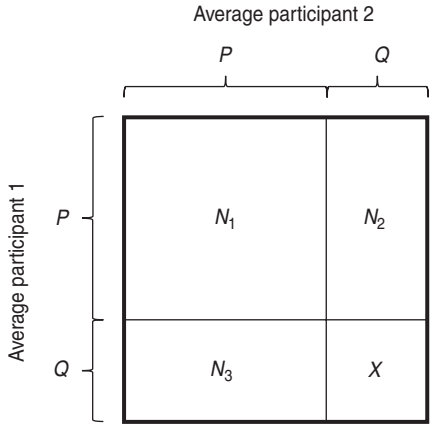


Fig. 1. Unit square representation of all possible combinations of reporting behaviour. P is the probability of reporting a specific contact (assumed to be equal for all participants). Q is the probability of not reporting the contact. N_1 is the number of contacts that were reported by both involved participants. N_2 and N_3 stand for the those contacts that were reported only by one participant. X is the number of contacts that were reported by none of the involved participants.

uncertainty of these approximations by means of bootstrapping.

The probability of forgetting to report a contact most likely depends on many factors, such as the duration and the intensity of the contact, the traits and the intra-individual variation of the motivation of the involved participant, as well as the context in which the contact takes place. Controlling and investigating all of these factors requires large datasets and complex study designs, which makes it difficult to convince target groups to participate. Thus, we concentrate on one of the supposedly most influential factors, i.e. contact duration, and analyse how reporting behaviour depends on a contact's duration.

We introduce the following simplifying assumptions and conventions as a prerequisite for approximating the probability of reporting a contact of a certain duration, P, as well as the number of completely unreported contacts: (i) the recall bias depends only on the duration of the contact and not on the characteristics of the involved participants or the context; (ii) the reports of the participants are stochastically independent; (iii) in any matching pair of contact reports, the duration with the higher value is assumed to be the true duration; (iv) contacts can be forgotten, but no contacts are reported that did not occur in reality.

Under these assumptions, the problem can be represented by a unit square (see Fig. 1) for all four duration categories. In this unit square, N_1 is the number of contacts with the duration of interest that were reported by both participants. N_2 is the number of contacts reported by participant 1, but not by participant 2. N_3 is the number of contacts reported by participant 2, but not by participant 1. We assumed here that all participants report contacts of a certain duration with the same probability [assumption (i)]. Accordingly, N_2 and N_3 can be derived from the total number of contacts reported by just one participant, N_{2+3}, by using the relation $N_{2+3} = 2N_2 = 2N_3$. X is the unknown number of contacts that were reported neither by participant 1 nor by participant 2. Due to assumptions (i) and (ii), the probability of reporting a contact, P, is defined as $P = N_1/(N_1+N_2) = N_1/(N_1+N_3)$ and the probability of forgetting to report a contact is given by the complementary probability $Q = 1-P$.

We assessed the uncertainty of our approximations by bootstrapping. To this end, 1000 resamples were constructed from the original sample and the probabilities P and Q were calculated for each of these resamples. Therefore, for all resampled participants, we added up (i) the numbers of contacts reported mutually by all egos and their alters, as well as (ii) the numbers of contacts that were only reported by the alters. Then, P is defined as the sum of all mutually reported contacts divided by the total of both sums. We used the mean, the 0·025 quantile (referred to as lower quantile) and the 0·975 quantile (referred to as upper quantile) as indices for describing the distribution and uncertainty of our approximations.

Statistical relationships between different variables were analysed with standard statistical tools such as the χ^2 test and linear regression analysis.

RESULTS

Descriptive characterization of the contact data

A total of 623 instances of contact were reported: 405 (65·0 %) of which were reported by both involved participants and 218 (35·0 %) were reported by only one participant and, thus, had no match (a list of all reported contacts is provided in the Supplementary online material, contact_data.csv). The cumulative distribution of contact duration is as follows: for 31·1 % of all individual contact reports, a duration of $\leqslant$5 min was listed; for 51·6 % of reports, $\leqslant$15 min was

Table 1. *Cross-tabulation of pairs of duration estimates*

Reported duration: lower value*	Reported duration: higher value				
	1–5 min	6–15 min	16–60 min	61–480 min	Total count
Not valid†	**0**	**0**	**3**	**2**	**5 + 4**‡
Not reported	**123** (57·5 %) (67·6 %)	**39** (18·2 %) (32·0 %)	**43** (20·1 %) (20·1 %)	**9** (4·2 %) (9·4 %)	**214** (100·0 %) (34·9 %)
1–5 min	**59** (43·7 %) (32·4 %)	**52** (38·5 %) (42·6 %)	**18** (13·3 %) (8·4 %)	**6** (4·4 %) (6·3 %)	**135** (100·0 %) (22·0 %)
6–15 min		**31** (35·6 %) (25·4 %)	**45** (51·7 %) (21·0 %)	**11** (12·6 %) (11·5 %)	**87** (100·0 %) (14·2 %)
16–60 min			**108** (74·5 %) (50·5 %)	**37** (25·5 %) (38·5 %)	**145** (100·0 %) (23·6 %)
61–480 min				**33** (100·0 %) (34·4 %)	**33** (100·0 %) (5·4 %)
Total count	**182** (29·6 %) (100·0 %)	**122** (19·9 %) (100·0 %)	**214** (34·9 %) (100·0 %)	**96** (15·6 %) (100·0 %)	**614** (100·0 %) (100·0 %)

* For every contact that was reported in this study, there is information regarding the existence and duration of this respective contact from two participants. This table shows a cross-tabulation of the higher contact duration estimate *vs.* the lower duration estimate of every reported contact. If just one participant reported the contact, then the lower value is set to 'not reported'.

† 'Not valid' indicates that the contact was reported, but no information or not-interpretable information about the duration was provided by one participant.

‡ There were four contacts that were reported only by one involved participant, but without information on the duration.

listed; for 69·2 %, $\leqslant$30 min; for 75·4 %, $\leqslant$45 min; and for 87·1 %, $\leqslant$1 h. The longest reported contact duration was 8 h. Most (90·0 %) of all valid reports asserted that the respective contact with a certain alter was only conversational. Only 10·0 % of all individual contact reports included physical contact.

Congruence between contact reports

For every matching pair of reported contacts, Table 1 shows whether or not the respective estimates of the contact duration were in accord with one another. For Table 1, we recoded the duration estimates of the participants into the time categories used by Mossong *et al.* [8], Mikolajczyk *et al.* [6], Horby *et al.* [12], and Smieszek [11]. In this table, the higher duration estimate (columns) was cross-tabulated against the lower duration estimate (rows). In the case of contacts that were only reported by one contact partner, we took the existing duration estimate as the higher estimate and introduced missing second reports of contact as the lowest category for the lower duration estimate. When analysing the correspondence of the duration categories of all matching pairs of contact reports, we see that not only 57·8 % of all reports were recoded into the same duration category and that 33·5 % of all pairs were allocated to adjacent duration categories, but also that 8·8 % differed by two or more time categories.

Table 2 shows a cross-tabulation of the kinds of contact for matching pairs of reported contacts. We classified contact events including physical contact as more intense than purely conversational contacts – regardless of the contact's duration. Table 3 has the same layout as Table 1; however, it includes only those contacts that were reported, at least by one of the involved participants, to have included physical contact. As the number of reports including physical contact is very low, we decided not to further analyse the impact of the reported kind of contact on the reporting behaviour.

Reporting behaviour by duration category

The descriptive data shown in Table 1 suggests that problems recalling contacts occur more often in the case of short encounters than in the case of long-lasting interactions. This is further confirmed by the results of a χ^2 test for independence between contact duration (four categories as defined in Table 1) and reporting

Table 2. *Cross-tabulation of pairs of reports on kind of contact*

	Kind of contact: more intense		
Kind of contact: less intense*	Only conversational	Including physical	Total count
Not valid†	**50**	**2**	**52 + 11**‡
Not reported	**192** (92·8 %) (39·2 %)	**15** (7·2 %) (21·4 %)	**207** (100·0 %) (37·0 %)
Only conversational	**298** (90·9 %) (60·8 %)	**30** (9·1 %) (42·9 %)	**328** (100·0 %) (58·6 %)
Including physical		**25** (100·0 %) (35·7 %)	**25** (100·0 %) (4·5 %)
Total count	**490** (87·5 %) (100·0 %)	**70** (12·5 %) (100·0 %)	**560** (100·0 %) (100·0 %)

* This table shows a cross-tabulation of the more intense contact report *vs*. the less intense report. If just one participant reported the contact, then the lower value is set to 'not reported'.
† 'Not valid' indicates that the contact was reported, but no information or not-interpretable information about the intensity of the contact was provided by at least one involved participant.
‡ There were 11 contacts that were reported by only one participant, but without information on the intensity of the contact.

Table 3. *Cross-tabulation of pairs of duration estimates* (*only events including physical contact*)

	Reported duration: higher value				
Reported duration: lower value*	1–5 min	6–15 min	16–60 min	61–480 min	Total count
Not valid	**0**	**0**	**0**	**0**	**0**
Not reported	**8** (53·3 %) (72·7 %)	**1** (6·7 %) (7·1 %)	**5** (33·3 %) (20·8 %)	**1** (6·7 %) (4·3 %)	**15** (100·0 %) (20·8 %)
1–5 min	**3** (14·3 %) (27·3 %)	**10** (47·6 %) (71·4 %)	**5** (23·8 %) (20·8 %)	**3** (14·3 %) (13·0 %)	**21** (100·0 %) (29·2 %)
6–15 min		**3** (37·5 %) (21·4 %)	**3** (37·5 %) (12·5 %)	**2** (25·0 %) (8·7 %)	**8** (100·0 %) (11·1 %)
16–60 min			**11** (61·1 %) (45·8 %)	**7** (38·9 %) (30·4 %)	**18** (100·0 %) (25·0 %)
61–480 min				**10** (100·0 %) (43·5 %)	**10** (100·0 %) (13·9 %)
Total count	**11** (15·3 %) (100·0 %)	**14** (19·4 %) (100·0 %)	**24** (33·3 %) (100·0 %)	**23** (31·9 %) (100·0 %)	**72** (100·0 %) (100·0 %)

* This table shows a cross-tabulation of the higher contact duration estimate *vs*. the lower duration estimate of every reported contact, but only those contact reports are included for which at least one participant stated that physical contact took place. If just one participant reported the contact, then the lower value is set to 'not reported'.

behaviour (contact reports by both contact partners *vs*. just by one contact partner), which rejects the null hypothesis that there is no relationship between these two variables with $\chi^2(3) = 134{\cdot}3$ ($P < 0{\cdot}001$).

According to our calculations, the probability P of reporting a contact is 49·0 % [bootstrapping inter-quantile interval (BIQI) 39·8–58·3] if contact duration is reported to be between 1 and 5 min; 81·0 %

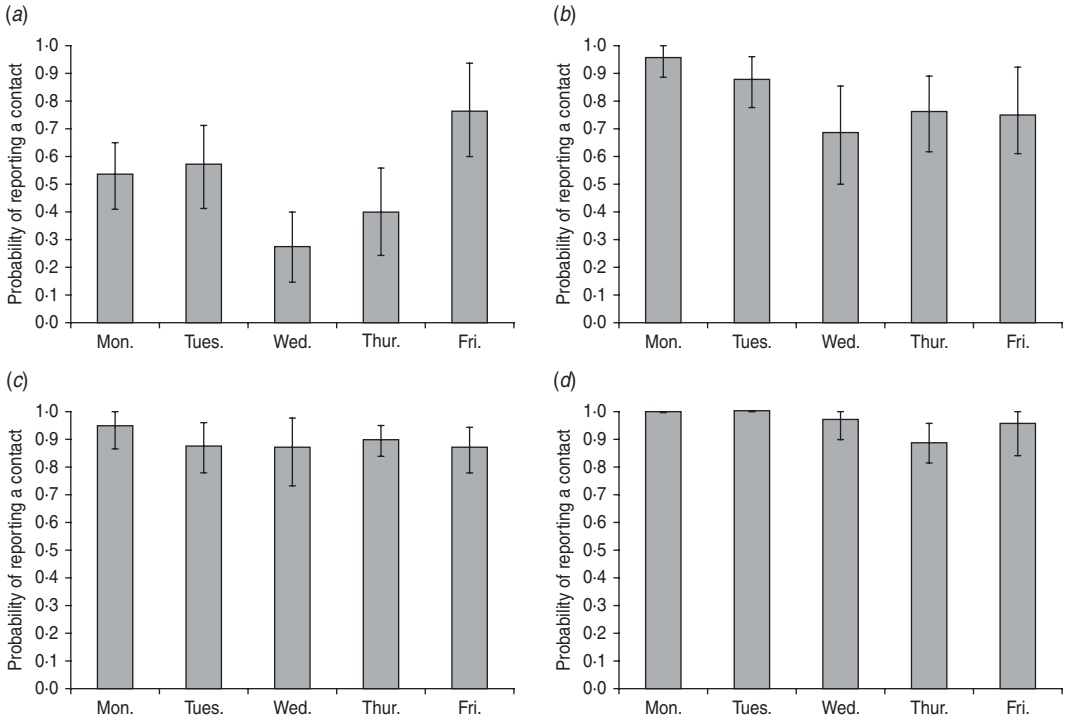


Fig. 2. Mean (grey bars), and upper and lower quantiles (whiskers) of the probabilities of reporting a contact by day of the week (calculated by bootstrapping). Indices for contacts of duration of (*a*) $\leqslant$5 min; (*b*) 6–15 min; (*c*) 16–60 min; (*d*) $>$1 h.

(BIQI 75·4–88·8) for 6–15 min; 89·0% (BIQI 84·6–93·1) for 16–60 min; and 95·2% (BIQI 92·0–97·9) for contacts $>$1 h. Thus, we expected that more than one quarter of contacts lasting $\leqslant$5 min were not reported at all, and less than 4% of contacts lasting between 6–15 min (Supplementary online material, section 1).

Self-reported *vs.* total number of contacts

We further analysed the relationship between the total number of contact partners attributed to a participant during the course of the study week (i.e. the number of set elements in the union of the contacts reported by an ego or its alters; $N_1+N_2+N_3$ in Fig. 1) and the actual number of contact partners reported by this participant (N_1+N_2). The relationship can be well described with a linear model: a linear regression analysis with the total reported number of contact partners as the independent variable, the self-reported number of contact partners as the dependent variable, and a forced intercept of zero (i.e. the regression line had to go through the origin) resulted in a slope of 0·83 with an explained variance $R^2=97{\cdot}7$ (the regression diagnostics are shown in the Supplementary online material, section 3).

Fatigue effects

Figure 2 shows the mean, the lower and the upper quantile for the probabilities of reporting a contact, P, calculated separately for all four duration categories and for all 5 days of the working week by means of bootstrapping. A decline in the reporting accuracy over time can be caused by fatigue. In the case of short contacts (1–5 min), the average P is between 50% and 60% on Monday and Tuesday; it drops below 40% on Wednesday and Thursday; however, the highest average P is 76·7% on Friday. In the case of all other duration categories, there appears to be a trend that P declines over the course of the week.

DISCUSSION

Interpretation of the results

On the basis of our analyses and the feedback we received from our participants, we interpret and discuss the results as follows:

(1) The overall level of reporting errors using the diary approach is rather high. More than one third of all reported contacts were only reported

by one participant. While our study design allows us to reconstruct those – presumably forgotten – contacts of an ego which are reported by the alter, in the common egocentric study design, this information is lost.

(2) We found the number of contact partners reported by a certain ego (N_1+N_2 in Fig. 1) to be approximately proportionally related to its total reported number of contact partners ($N_1+N_2+N_3$). This finding is in accord with other research on recall bias in network research [7, 17, 18] and with our other datasets (T. Smieszek, J. Maag and L. Muggler unpublished findings). That means that there is higher underreporting for highly connected individuals than for rather isolated individuals. While for some research questions and methodologies this bias might be unproblematical, other findings might be highly affected by it. For instance, Mikolajczyk & Kretzschmar [7] argue that for models based purely on the *relative* average contact frequency differences between age groups, this bias is irrelevant (see discussion on p. 133 of their paper). However, their argument is only correct if age is not correlated with other predictors for reporting errors, such as the duration of the contacts.

(3) It is likely that the proportional relationship between the total and the self-reported number of contacts we found only holds true for a limited range of contact partners. The maximum number of contact partners at work during one day reported in this study was 16. It is plausible to assume in cases of much higher contact numbers (e.g. from a train conductor or flight attendant), that individuals would either deny their participation or would report disproportionally fewer contact partners. Furthermore, there is evidence that the proportion of short and non-intense contacts increases with the total number of contact partners [11]. If highly connected individuals show disproportionately high numbers of short contacts, they are also likely to particularly suffer from difficulty recalling the contacts they had.

(4) The underreporting of contacts in diary-based datasets is highly correlated with the duration of a certain contact. We estimate that the probability of forgetting a contact that lasts $\leqslant 5$ min is more than 50%. In contrast, contacts that last >1 h have an estimated probability of about 5% of going unreported. This finding, that deficient recall depends on measures of contact intensity, is intuitively plausible: short encounters are, in many cases, accidental and of rather low importance for the involved individuals. Humans tend to remember events that have a high emotional or resource involvement better than they do short and unimportant occurrences. This systematic bias might particularly affect research that builds upon intensity-differentiated contact data [e.g. 11].

(5) Finally, in longitudinal studies like ours, fatigue effects might occur and can be a relevant influence factor on the number and kind of reporting errors. McCaw *et al*. searched for fatigue effects in their contact data with two different analyses: they found no evidence that the sequence of the different modes of data collection influenced the reporting quality, but within a particular mode the number of reported contacts declined with time [10]. It is difficult to interpret our data with respect to fatigue effects as – due to the study design – it is inherently impossible to distinguish the effects of the specific peculiarities of a certain study day from fatigue: it seems plausible to us that the pronounced fall in reporting accuracy on Wednesday was caused by a particularly strenuous workload for one research group on that day, while the fact that many study participants work at home on Fridays might explain that day's above-average accuracy in reporting contacts lasting between 1 and 5 min. Considering that it was not possible to control for the impact of the particular study day, the decline of the probabilities towards the end of the week still suggests that there might be a slight fatigue effect.

Limitations of the study

Caution should be exercised when generalizing our findings because they are based on a small, specific group of participants (academically trained people) within a specific setting (scientists working for a university). Although the office setting found in a university is typical of many professions, the results of an analogous study with other participants and another setting might differ. Although we deem it plausible that the general effects found in this study are also true for other groups, more studies on different groups are needed to achieve a more robust picture on the errors in diary-based contact data.

Furthermore, our data did not allow us to analyse and to control for all potentially relevant

determinants of reporting behaviour. We assumed, for instance, that participants in a contact study do not differ in their reporting probabilities. In reality, participants in such studies differ in their motivation as well as in their cognitive abilities. In principle, it is possible to calculate the individual probabilities of reporting a contact by applying the unit square (Fig. 1) to all possible combinations of individuals (Supplementary online material, section 2). However, the theoretical maximum of reported contacts per pair of participants is specified by the number of study days, because the usual contact definition relies on the accumulated time of interaction during an *entire* day. In our study, there are at maximum five contact reports per pair of individuals. On one hand, such low numbers do not allow robust estimates of P_1 and P_2. On the other hand, it is not feasible to conduct longitudinal contact diary studies that last much longer, because in that case many people would refuse to participate.

We believe that most unmatched contact reports are the result of underreporting. In principle, it is also possible that contacts are reported that have either not occurred or that do not fall under the given definition of a potentially contagious contact. Some participants mentioned difficulties in deciding whether a certain interaction occurred at a distance of less than or more than 2 m. They mentioned particular difficulties with accurately reporting interactions that took place during meetings or social gatherings. It is further possible that participants of such a study do not understand the contact definition correctly, which also might result in over- or underreporting of contacts.

CONCLUSION

To conclude, it can be stated that diary-based contact data is more appropriate for certain types of analyses and for certain host–pathogen systems than it is for others. The contact diary approach is probably problematical for detailed investigations of the spread dynamics of highly contagious diseases (e.g. typical childhood diseases such as *Bordetella pertussis*). In the case of such host–pathogen systems, even minor contact is sufficient to transmit infection. Since such contacts are particularly affected by the described biases, it is likely that a large proportion of important contact information is missing in diary-based datasets.

The opposite is true for host–pathogen systems in which transmission takes place through long and intense interaction (e.g. *Neisseria meningitidis* or *Staphylococcus aureus*) and which often achieve only low to medium basic reproduction numbers. Here, the contact topology greatly influences spread dynamics [4] and, at the same time, contact diary-based data is likely to be more accurate than in the case of highly contagious infections.

We only recommend applying the contact diary method either when the planned analyses are robust against the expected reporting errors and biases, or when the relevant contacts are so intense that the expected level of reporting accuracy is sufficient. When possible, diary-based approaches should be complemented with other approaches, like measurements made with wearable sensor badges that precisely record close spatial co-location [19–21]. Such complementary measurements allow data cross-validation and provide more robust insights into a system's contact topology.

NOTE

Supplementary material accompanies this paper on the Journal's website (http://journals.cambridge.org/hyg).

ACKNOWLEDGEMENTS

This research was funded by the Swiss National Science Foundation (project 32003B_127548). The cooperation and commitment of the 50 participants made this study possible. Mirjam Kretzschmar, Lena Fiebig, Corinne Moser, Andrea Ulrich, Anna Drewek, and Jan Hattendorf helped to improve the quality of this paper with their valuable comments. We thank two anonymous referees for their thorough review of our manuscript that further advanced our work. Sandro Bösch helped with the final layout of the figures. The manuscript was copyedited by EditMyEnglish.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Bansal S, Grenfell BT, Meyers LA.** When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface* 2007; **4**: 879–891.

2. **Duerr HP, *et al.*** The impact of contact structure on infectious disease control: influenza and antiviral agents. *Epidemiology and Infection* 2007; **135**: 1124–1132.
3. **Pastor-Satorras R, Vespignani A.** Epidemic spreading in scale-free networks. *Physical Review Letters* 2001; **86**: 3200–3203.
4. **Smieszek T, Fiebig L, Scholz RW.** Models of epidemics: when contact repetition and clustering should be included. *Theoretical Biology and Medical Modelling* 2009; **6**: 11.
5. **Szendrói B, Csányi G.** Polynomial epidemics and clustering in contact networks. *Proceedings of the Royal Society of London, Series B: Biological Science* 2004; **271**: S364–S366.
6. **Mikolajczyk RT, *et al.*** Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology and Infection* 2008; **136**: 813–822.
7. **Mikolajczyk RT, Kretzschmar M.** Collecting social contact data in the context of disease transmission: prospective and retrospective study designs. *Social Networks* 2008; **30**: 127–135.
8. **Mossong J, *et al.*** Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* 2008; **5**: e74.
9. **Beutels P, *et al.*** Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiology and Infection* 2006; **134**: 1158–1166.
10. **McCaw JM, *et al.*** Comparison of three methods for ascertainment of contact information relevant to respiratory pathogen transmission in encounter networks. *BMC Infectious Diseases* 2010; **10**: 166.
11. **Smieszek T.** A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. *Theoretical Biology and Medical Modelling* 2009; **6**: 25.
12. **Horby P, *et al.*** Social contact patterns in Vietnam and implications for the control of infectious diseases. *PLoS One* 2011; **6**: e16965.
13. **Smieszek T, *et al.*** Reconstructing the 2003/2004 H3N2 influenza epidemic in Switzerland with a spatially explicit, individual-based model. *BMC Infectious Diseases* 2011; **11**: 115.
14. **Goeyvaerts N, *et al.*** Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society: Series C* (*Applied Statistics*) 2010; **59**: 255–277.
15. **Kretzschmar M, Teunis PFM, Pebody RG.** Incidence and reproduction numbers of Pertussis: estimates from serological and social contact data in five European countries. *PLoS Medicine* 2010; **7**: e1000291.
16. **Wallinga J, Teunis PFM, Kretzschmar M.** Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology* 2006; **164**: 936–944.
17. **Brewer DD, Webster CM.** Forgetting of friends and its effect on measuring friendship networks. *Social Networks* 1999; **21**: 361–373.
18. **Brewer DD, Garrett SB, Kulasingam S.** Forgetting as a cause of incomplete reporting of sexual and drug injection partners. *Sexually Transmitted Diseases* 1999; **26**: 166–176.
19. **Salathé M, *et al.*** A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences USA* 2010; **107**: 22020–22025.
20. **Cattuto C, *et al.*** Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS One* 2010; **5**: e11596.
21. **Pentland A.** Automatic mapping and modeling of human networks. *Physica A* 2007; **378**: 59–67.

BMC Medicine

BioMed Central
The Open Access Publisher

A practical method to target individuals for outbreak detection and control

BMC Medicine 2013, **11**:36 doi:10.1186/1741-7015-11-36

Gerardo Chowell (gchowell@asu.edu)
Cecile Viboud (viboudc@mail.nih.gov)

ISSN 1741-7015

Article type Commentary

Submission date 30 January 2013

Acceptance date 30 January 2013

Publication date 12 February 2013

Article URL http://www.biomedcentral.com/1741-7015/11/36

A practical method to target individuals for outbreak detection and control

Gerardo Chowell[1,2*] and Cécile Viboud[2]

[1]Mathematical and Computational Modeling Sciences Center, School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

[2]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

[*]Corresponding author

Email addresses:

GC: gchowell@asu.edu

CV: viboudc@mail.nih.gov

Abstract

Identification of individuals or subpopulations that contribute the most to disease transmission is key to target surveillance and control efforts. In a recent study in *BMC Medicine*, Smieszek and Salathé introduced a novel method based on readily available information about spatial proximity in high schools, to help identify individuals at higher risk of infection and those more likely to be infected early in the outbreak. By combining simulation models for influenza transmission with high-resolution data on school contact patterns, the authors showed that their proximity method compares favorably to more sophisticated methods using detailed contact tracing information. The proximity method is simple and promising, but further research is warranted to confront this method against real influenza outbreak data, and to assess the generalizability of the approach to other important transmission units, such as work, households, and transportation systems.

See related research article here http://www.biomedcentral.com/1741-7015/11/35

Keywords: contact network; hotspot; dynamic network; contact pattern; wireless sensing devices; collocation ranking; class schedule; high school; influenza; disease transmission.

Background

The transmission potential of an infectious disease is directly related to the characteristics of the infectious agent, its host population and the local environment [1]. The contribution of these factors can be encapsulated in a single parameter that is key for disease control, namely, the 'reproduction number', which quantifies the average number of secondary cases generated by an infectious individual during the early epidemic phase [1]. Identification of individuals or subpopulations associated with high transmission potential is particularly useful to guide surveillance and control strategies, especially when resources are limited [2].

Understanding the complexity of dynamic human interactions and contact networks is crucial to identifying hotspots of disease transmission during an outbreak [3]. The dynamic social contact networks relevant for disease spread depends on a number of factors, including individual host characteristics (e.g, age, prior immunity, number of contacts), pathogen characteristics (transmission mode), characteristics of the space in which individuals interact (for example, confined versus open setting, room capacity), and the duration and proximity of human interactions.

Recent technological advances in miniature wireless sensing devices have allowed unobtrusive and unsupervised quantification of the dynamic network of human interactions in various settings, including schools [4-6], conferences [7], and hospitals [8]. In particular, these innovative technologies have increased our understanding of face-to-face contact patterns relevant for the spread of rapidly transmitted infectious agents [4, 9]. Given the large amount of costly information captured by these devices, there is active debate on the minimum level of data that is required to capture the essence of disease transmission and to be sufficient to inform disease control [7, 10].

Performances of various indicators of social connectivity

A recent study by Smieszek and Salathé, published in *BMC Medicine* [11], used high-resolution contact-network data collected by wireless sensing devices during a 1-day period at a high school in the USA, combined with extensive epidemic simulations, to evaluate the effectiveness of several metrics to identify individuals who play a significant role in outbreak dissemination. The consolidated network dataset was limited to close proximity interactions, based on records indicating face-to-face contacts within a distance of less than 3 m at a certain point in time. The dataset also included location records indicating the presence of an individual in a specific classroom.

The authors then quantified the performances of a variety of indicators of social connectivity, which required different levels of information on the high-school contact network to identify individuals with high transmission potential. In particular, the authors introduced a low-cost indicator of social connectivity, based on the 'collocation-ranking method', which relies on the cumulative amount of time that an individual spends with other individuals in the same room, modulated by class size. Such information does not rely on the detailed structure of the high-school contact network, and can be retrieved from schedule data alone. The Smieszek and Salathé study relied on simulations of influenza transmission on the detailed high-school contact network to assess the performances of the different indicators, in terms of their ability to identify individuals at higher risk of infection and those with early disease onset.

Findings and potential applications

Epidemic simulations showed that the simple schedule-based collocation ranking indicator clearly outperformed methods selecting individuals at random, and compared favorably with more data-hungry indicators. Because collecting reliable data about individual-level interactions is cumbersome and expensive to obtain at the community level, the authors

proposed that their low-cost collocation method can be exploited for the design of sentinel surveillance systems, with the potential to quickly detect the onset of an infectious disease outbreak, and thereby optimize mitigation and prevention strategies. In particular, sentinel high-school students could be selected from those with high collocation ranking, and these could then be monitored for their infection status throughout the influenza season, and/or be prioritized for vaccination in the case of vaccine shortage, in an effort to stamp out an emerging outbreak.

Limitations and future directions

This interesting proof-of-concept study by Smieszek and Salathé addressed social interactions within a high school, which is an important focus for seasonal and pandemic influenza transmission [12]. As acknowledged by the authors, a key limitation of this study is the lack of validation against epidemiological data from real school outbreaks. The simulation model used to evaluate the performances of the method is a conceptualized version of disease transmission, and although it is driven by real contact information, it remains one step removed from the actual disease-transmission process. A previous study combining outbreak data in an elementary school with contact-network information highlighted the importance of gender on influenza transmission, with children of the same gender infecting each other more frequently (reflecting assortative mixing) [4], an issue that was not considered by Smieszek and Salathé. Interestingly, school outbreak data have also shown that the exact location of children within the classroom does not matter, which supports the use of simple class-schedule information as proposed by Smieszek and Salathé [11] rather than the use of more detailed seating charts. Although there has been good progress overall in elucidating social interactions among school-age children, more studies are needed to address whether contact patterns, and hence transmission links, might differ between elementary and high schools.

Another limitation of the school-based study by Smieszek and Salathé [11] relates to the contribution of other units to disease transmission. About one-third of all influenza secondary-transmission events are believed to occur within households [13], whereas only 7 to 20% are thought to occur in schools [14]. Hence, estimating the relative infection risk of individuals in a variety of settings relevant for disease transmission, including schools, households, conferences, and transportation systems, will be important in future research. It is not clear how the method proposed by Smieszek and Salathé [11] could be generalized to household and work environments, where systematic 'schedules' are more difficult to obtain.

As noted by the authors, the transmission mode of influenza and other respiratory pathogens is not clearly understood, but probably involves a combination of direct contact and transmission by fomites and aerosols, which makes it difficult to capture the social network relevant for disease transmission. Because the transmissibility of influenza has been shown to be associated with environmental conditions [15, 16], actual transmission rates could vary within the same school, house, or office building, owing to local differences in the environment. In the future, more elaborate studies should collect local environmental variables such as room ventilation rates to better quantify influenza transmission potential in confined settings [17].

In summary, Smieszek and Salathé [11] have introduced a promising and practical method to identify individuals with high infection potential who can be targeted for outbreak detection and control. Future studies should employ consistent methodological approaches to measure contact networks in different settings, in parallel with careful disease monitoring. Technological advances in contact-network sensing devices and pathogen identification methods (for example, multiplex PCR), combined with innovative approaches for disease surveillance (for example, web-based and smart-phone technologies [18]), have huge

potential to increase our understanding of infectious disease transmission and to suggest novel ways of detecting and controlling outbreaks.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Both authors contributed to the writing and editing of this commentary. Both authors read and approved the final manuscript.

Authors' information

GC is an associate professor in the School of Human Evolution and Social Change at Arizona State University and a research fellow at the Fogarty International Center, US National Institutes of Health. Research interests include mathematical and statistical modeling of infectious disease transmission and control interventions, with a focus on seasonal and pandemic influenza and the quantitative characterization of past influenza pandemics.

CV is a senior research scientist at the Fogarty International Center, US National Institutes of Health, focusing on the transmission dynamics and health burden of acute viral infections, at the interface between mathematical modeling, epidemiology, evolutionary genetics, and public health.

Acknowledgments

We are grateful for financial support from MISMS (Multinational Influenza Seasonal Mortality Study), an ongoing international collaborative effort to understand influenza epidemiological and evolutionary patterns, led by the Fogarty International Center, National Institutes of Health (http://www.origem.info/misms/index.php). The MISMS study is funded by the International Influenza Unit, Office of Global Health Affairs, Department of Health and Human Services.

References

1. Anderson RM, May RM: *Infectious diseases of humans.* Oxford: Oxford University Press; 1991.

2. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM: **Superspreading and the effect of individual variation on disease emergence**. *Nature* 2005, **438**:355-359.

3. Chowell G, Nishiura H, Viboud C: **Modeling rapidly disseminating infectious disease during mass gatherings**. *BMC Med* 2012, **10**:159.

4. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow D: **Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza**. *Proc Natl Acad Sci U S A* 2011, **108**:2825-2830.

5. Stehle J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton JF, Quaggiotto M, Van den Broeck W, Regis C, Lina B, Vanhems P: **High-resolution measurements of face-to-face contact patterns in a primary school**. *PLoS One* 2011, **6**:e23176.

6. 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,* **107**:22020-22025.

7. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L, Régis C, Pinton JF, Khanafer N, Van den Broeck W, Vanhems P: **Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees**. *BMC Med* 2011, **9**:87.

8. Isella L, Romano M, Barrat A, Cattuto C, Colizza V, Van den Broeck W, Gesualdo F, Pandolfi E, Ravà L, Rizzo C, Tozzi AE: **Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors**. *PLoS One* 2011, **6**:e17144.

9. 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**:e11596.

10. Blower S, Go MH: **The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy?** *BMC Med* 2011, **9**:88.

11. Smieszek T, Salathé M: **A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks**. *BMC Med* 2013.

12. Cauchemez S, Ferguson NM, Wachtel C, Tegnell A, Saour G, Duncan B, Nicoll A: **Closure of schools during an influenza pandemic**. *Lancet Infect Dis* 2009, **9**:473-481.

13. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS: **Strategies for containing an emerging influenza pandemic in Southeast Asia**. *Nature* 2005, **437**:209-214.

14. Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson NM: **Estimating the impact of school closure on influenza transmission from Sentinel data**. *Nature* 2008, **452**:750-754.

15. Steel J, Staeheli P, Mubareka S, Garcia-Sastre A, Palese P, Lowen AC: **Transmission of pandemic H1N1 influenza virus and impact of prior exposure to seasonal strains or interferon treatment**. *J Virol* 2011, **84**:21-26.

16. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M: **Absolute humidity and the seasonal onset of influenza in the continental United States**. *PLoS Biol* 2010, **8**:e1000316.

17. Fabian P, McDevitt JJ, DeHaan WH, Fung RO, Cowling BJ, Chan KH, Leung GM, Milton DK: **Influenza virus in human exhaled breath: an observational study**. *PLoS One* 2008, **3**:e2691.

18. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, Buckeridge DL: **Information technology and global surveillance of cases of 2009 H1N1 influenza**. *N Engl J Med* 2010, **362**:1731-1735.