Infectious disease epidemiology: slowly moving towards open science

Niel Hens

‘Embracing Data Management - Bridging the Gap Between Theory and Practice’, FOSTER, 4 June 2015, Brussels
Outline

1. Introduction

2. Infectious Disease Epidemiology & Statistics
   - Statistics
   - Infectious Disease Epidemiology

3. Discussion & Recommendations
Opinions are contagious, or not?

LaCour, M. and Green, D. Science, 2014 - retracted

When contact changes minds: an experiment on transmission of support for gay equality

→ **Outcome**: Contact with minorities coupled with discussion of issues pertinent to them is capable of producing a cascade of opinion change.

→ **Retraction because, among other issues**: ‘LaCour has not produced the original survey data from which someone else could independently confirm the validity of the reported findings.’

The Chronicle, 21 May 2015: How 2 Persistent Grad Students Upended a Blockbuster Study
Compliance with data sharing is challenging.

- **Allison, D. Science, 2009:** The movement toward open data has begun: NIH, Science, Nature journals.
- **Nature Genetics** requires authors to **publicly deposit microarray gene expression data:**
  - reproduction of published work for only 2 of 18 papers.
  - data were either not found or incomplete or original analysis descriptions were unclear
- **Public Library of Science** journals require data sharing
  - until recently this didn’t happen
  - now they insist
Open Science

‘Open Science - Prinzipien’ by Andreas E. Neuhold - Own work. Licensed under CC BY 3.0
Statistics: clinical trials

- Clinical trials are the best regulated studies
  - https://clinicaltrials.gov
  - https://www.clinicaltrialsregister.eu

- JAMA: clinical trials sponsored by drug companies have to be reanalyzed by ‘academic statisticians’ before acceptance for publication in the journal

- Keiding, N. Biostatistics, 2010
  - There is usually much more to ‘research’ than the statistical analysis.
    - A ‘reasonable choice of statistical model’

- Clinical trials have changed data management

- Unfortunately not all guidelines and recommendations make sense and require constant re-evaluation.
Statistics: sharing code

- Editorial board meeting @Biometrics:
  - issue: mandatory submission of R-package (stats software)
  - AEs voted against: burden & risk of packages with limited lifespan
- Sharing code: ‘available upon request’
- The Journal of Statistical Software: 1st rank
Infectious Disease Epidemiology: sharing data?

- Mostly **observational studies or surveys:**
  - outbreak data: SARS, H1N1, Ebola, MERS, ...
    - no clear data format
  - social contact survey (EU FP6 project POLYMOD)

- Emerging trend: **syndromic surveillance**
  - the **great influenza survey**
  - **funding** issues
  - little political support
Infectious Disease Epidemiology: sharing data?

The Good, the Bad and the Ugly
The good, ...

- **Benjamin Cowling** and colleagues: influenza research:
  - [https://sites.google.com/site/bencowling88/datasets](https://sites.google.com/site/bencowling88/datasets)
  - shares raw data & code according to NIH guidelines
    - promote reproducibility of results
    - allow others to conduct their own analyses
    - allow others to compare other data
    - allow others to plan their own studies

- **John Edmunds** and colleagues: Ebola
  - [http://cmmid.lshtm.ac.uk/research/ebola/](http://cmmid.lshtm.ac.uk/research/ebola/)
  - shares code and results in realtime on website
  - publications follow later

...
The good, . . .

- Jombaert et al. (Epidemics, 2014): OutbreakTools - Hackathon - Git repository

- Hens et al. (2012): ±9800 e-chapter downloads - all data & code available
the bad, ...

- violation of scientific integrity
- tax payer’s money
- human health
and the ugly.

- Sharing of raw data without documentation:

- Even with documentation it can go wrong . . .
Social Contact Survey

Goal:
Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents.

Objectives
- Disentangle contact behaviour from transmission process
- Get insights in predictiveness of social contact data
- Get new insights in the transmission process
Social Contact Survey

- Wallinga et al. (2006): Utrecht
- POLYMOD
  - pilot study: Beutels et al. (2006)
  - main study: Mossong et al. (2008)

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

Joël Mossong¹², Niel Hens³, Mark Jit⁴, Philippe Beutels⁵, Kari Auranen⁶, Rafael Mikolajczyk⁷, Marco Massari⁸, Stefania Salmaso⁹, Gianpaolo Scalia Tomba⁹, Jacco Wallinga¹⁰, Janneke Heijne¹⁰, Małgorzata Sadkowska-Todys¹¹, Magdalena Rosinska¹¹, W. John Edmunds⁴

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

(WoK: 447 citations)
Social Contact Survey

Belgian Contact Survey

- Part of POLYMOD project
- Period March - May 2006
- 750 participants, selected through random digit dialing
- Diary-based questionnaire
- Two main types of contact: non-close and close contacts
- Total of 12775 contacts (≈ 16 contacts per person per day)

Hens et al. (2009a,b)
Mixing patterns in EU
Mixing patterns in EU

- Data were shared, via research gate, the VENICE platform etc
- Data were misused
  - because they were made available
  - because researcher did not check the supplementary material
  - ‘supplementary’?
  - . . .?

Examples:

- hepatitis A
- impact of weather on social contact behaviour
- reciprocity
### Mixing patterns in EU

#### Supplementary Information Table S1. Details of survey methodology in each country

<table>
<thead>
<tr>
<th>Country</th>
<th>Belgium (BE)</th>
<th>Germany (DE)</th>
<th>Finland (FI)</th>
<th>Great Britain (GB)</th>
<th>Italy (IT)</th>
<th>Luxembourg (LU)</th>
<th>Netherlands (NL)</th>
<th>Poland (PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recruitment method</strong></td>
<td>Random digit dialling on fixed telephone lines</td>
<td>Door-to-door household visits</td>
<td>Stratified sample from population register, mobile or land lines</td>
<td>Door-to-door household visits</td>
<td>Random digit dialling on fixed telephone lines</td>
<td>Random digit dialling on fixed telephone lines</td>
<td>Sampling from a register</td>
<td>Household visits</td>
</tr>
<tr>
<td><strong>Financial incentives</strong></td>
<td>5€</td>
<td>None</td>
<td>None</td>
<td>£3</td>
<td>None</td>
<td>5€</td>
<td>10€ for participation in the larger study</td>
<td>5€</td>
</tr>
<tr>
<td><strong>Sampling method</strong></td>
<td>Quota sampling by age, sex, geographical region, rural/urban</td>
<td>Quota sampling by age</td>
<td>Quota sampling by age and sex</td>
<td>Quota by age, sex, geographical region, rural/urban and day of the week</td>
<td>Quota sampling by age and sex</td>
<td>Quota sampling by age and sex</td>
<td>Quota sampling by age, sex, geographical region, rural/urban</td>
<td></td>
</tr>
<tr>
<td><strong>How were the diaries delivered and collected?</strong></td>
<td>Delivered and collected by mail</td>
<td>Face-to-face</td>
<td>Delivered and collected by mail</td>
<td>Delivered by mail, collected in person</td>
<td>Delivered and collected by mail</td>
<td>Delivered and collected by mail</td>
<td>Delivered in person and collected by mail</td>
<td>Delivered and collected in person</td>
</tr>
</tbody>
</table>
## Mixing patterns in EU

<table>
<thead>
<tr>
<th>Country</th>
<th>Belgium (BE)</th>
<th>Germany (DE)</th>
<th>Finland (FI)</th>
<th>Great Britain (GB)</th>
<th>Italy (IT)</th>
<th>Luxembourg (LU)</th>
<th>Netherlands (NL)</th>
<th>Poland (PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Follow-up of participants</strong></td>
<td>RC, FU1, max. threeFU2</td>
<td>No*</td>
<td>FU1</td>
<td>RC, FU1</td>
<td>RC, FU1, FU2</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Was the survey part of a larger study?</strong></td>
<td>No*</td>
<td>Yes, participants of a large multi-theme survey</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Data quality checks</strong></td>
<td>All manually checked and double entered</td>
<td>100 diaries double entered</td>
<td>Not known</td>
<td>double entry</td>
<td>double entry</td>
<td>single entry, all diaries checked manually</td>
<td>10% double entry</td>
<td>single entry, all diaries checked manually</td>
</tr>
<tr>
<td><strong>Age groups used for different diary types:</strong> young children (parental proxy reporting), older children and teens (self-reporting), adults (self-reporting)</td>
<td>0-8 parental proxy reporting, 9-17 self-reporting children and teenagers, 18+ self reporting adults</td>
<td>0-13 parental proxy reporting, 14+ self-reporting</td>
<td>No younger children self-reporting</td>
<td>0-7 parental proxy reporting, 8-17 self-reporting children and teenagers, 18+ self reporting adults</td>
<td>0-8 parental proxy reporting, 9-17 self-reporting children and teenagers, 18+ self reporting adults</td>
<td>0-8 parental proxy reporting, 9-17 self-reporting children and teenagers, 18+ self reporting adults</td>
<td>0-8 parental proxy reporting, 9-17 self-reporting children and teenagers, 18+ self reporting adults</td>
<td></td>
</tr>
<tr>
<td><strong>Were participants instructed not to record professional contacts (eg, with clients) in the diary?</strong></td>
<td>Yes, if estimated at more than 20</td>
<td>Yes, if estimated at more than 10</td>
<td>Yes, if estimated at more than 10</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes, if estimated at more than 10</td>
<td>No</td>
</tr>
</tbody>
</table>
## Mixing patterns in EU

<table>
<thead>
<tr>
<th>Country</th>
<th>Belgium BE</th>
<th>Germany DE</th>
<th>Finland FI</th>
<th>Great Britain GB</th>
<th>Italy IT</th>
<th>Luxembourg LU</th>
<th>Netherlands NL</th>
<th>Poland PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum number of contact entries in the diary?</td>
<td>90</td>
<td>73</td>
<td>34</td>
<td>29</td>
<td>45</td>
<td>55</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>What languages were used?</td>
<td>French &amp; Dutch**</td>
<td>German</td>
<td>Finnish and Swedish</td>
<td>English</td>
<td>Italian</td>
<td>German, French &amp; Portuguese</td>
<td>Dutch</td>
<td>Polish</td>
</tr>
</tbody>
</table>

* Participants were assigned to collect diary data on two days of the week, one weekend and one workday per respondent. For the purpose of the current comparative multicountry analysis, a random selection was made from the Belgian data of one day per respondent, evenly distributed over all days of the week.

** Different contextual terminology between Belgium and The Netherlands (Dutch) or Luxembourg (French), mainly related to education, employment and migration
Mixing patterns in EU
Discussion & Recommendations

- **The Good, the Bad and the Ugly:**
  - trade-off: sharing raw data requires documenting the data and sharing code
  - worthwhile the effort

- There is hope: moving towards a unifying format for outbreaks (FF100, UK, 2009)

- Much more work needs to be done though!
Infectious disease epidemiology: slowly moving towards open science

Discussion & Recommendations
References
