Keep it clean: The importance of data preparation

Will Dixon
EULAR RODS meeting
Prague, Nov 14 and 15, 2013
Step by step

1. From the estate office, head towards the train station and then onwards along the canal until you reach the Standedge Tunnel visitor centre.

2. Head up to Waters Road. Walk up the drive beside the Tunnel End pub and then take the path behind it. Walk across fields through a green gate to a group of houses.

3. Go through another gate beside the house on the right to bring you out on a lane. Turn left and follow it past Berry Greave. When it starts to descend downhill, turn right and follow this lane past the first house and round to a path at the back of the next house. Walk across the fields and at the small footbridge keep to high ground before you drop sharply down to a stream and back up again to the right of the phone mast. Turn right down Blake Lee Lane past Lower Green Owlers to a road junction. Turn left and climb up the track to White Hall Farm. In front of the farm go through the gate leading on to the moor and through some rushes, following the path along the edge of Haigh Clough to the southern corner of March Haigh reservoir.

4. Here you have the option to head up to Buckstones car park for magnificent views across the estate. Cross the dam wall and take the path through the rocks to the car park. Once you’ve caught your breath, retrace your steps back to the dam wall.

5. Continue to head south to meet up with the Packhorse Trail, where you turn right and follow the trail across Willmer Clough and up to the top to one of the old marker stones. Go straight ahead until you reach Haigh Gutter.
Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration

Jan P. Vandenbroucke¹, Erik von Elm²,³, Douglas G. Altman⁴, Peter C. Gøtzsche⁵, Cynthia D. Mulrow⁶, Stuart J. Pocock⁷, Charles Poole⁸, James J. Schlesselman⁹, Matthias Egger²,¹⁰ for the STROBE Initiative

¹ Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands, ² Institute of Social & Preventive Medicine (ISPM), University of Bern, Bern, Switzerland, ³ Department of Medical Biometry and Medical Informatics, University Medical Centre, Freiburg, Germany, ⁴ Cancer Research UK/NHS Centre for Statistics in Medicine, Oxford, United Kingdom, ⁵ Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark, ⁶ University of Texas Health Science Center, San Antonio, United States of America, ⁷ Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, United Kingdom, ⁸ Department of Epidemiology, University of North Carolina School of Public Health, Chapel Hill, United States of America, ⁹ Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, and University of Pittsburgh Cancer Institute, Pittsburgh, United States of America, ¹⁰ Department of Social Medicine, University of Bristol, Bristol, United Kingdom

METHODS

### Study design

- Present key elements of study design early in the paper

### Setting

- Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection

### Participants

- **Cohort study**—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up
  - *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls
  - *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants
  - **Cohort study**—For matched studies, give matching criteria and number of exposed and unexposed
  - *Case-control study*—For matched studies, give matching criteria and the number of controls per case

### Variables

- Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable

### Data sources/measurement

- For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group

### Bias

- Describe any efforts to address potential sources of bias

### Study size

- Explain how the study size was arrived at

### Quantitative variables

- Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why

### Statistical methods

- **Cohort study**—If applicable, explain how to follow-up was addressed
- **Case-control study**—If applicable, explain how matching of cases and controls was addressed
- *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy
- Describe any sensitivity analyses
Example of checklist in action

- Variables
- Data sources/measurement
- Quantitative variables
- Statistical methods

Influence of Anti–Tumor Necrosis Factor Therapy on Cancer Incidence in Patients With Rheumatoid Arthritis Who Have Had a Prior Malignancy: Results From the British Society for Rheumatology Biologies Register

W. G. DIXON, K. D. WATSON, M. LUNT, L. K. MERCER, BRITISH SOCIETY FOR RHEUMATOLOGY BIOLOGICS REGISTER CONTROL CENTRE CONSORTIUM, K. L. HYRICH, AND D. P. M. SYMONDS, ON BEHALF OF THE BRITISH SOCIETY FOR RHEUMATOLOGY BIOLOGICS REGISTER

**Statistical analysis.** Followup time was calculated from the date of the first anti-TNF drug use for the anti-TNF–treated cohort, or from the registration date for the comparison cohort, to September 30, 2007 or the death date, whichever occurred first. Malignancy data were captured from all 3 sources until April 2008 to allow for delays in notification of incident malignancies to the BSRBR. Within the anti-TNF cohort, patients could switch between anti-TNF drugs. The anti-TNF cohort contributed person-years of followup even if the anti-TNF therapy was stopped. Malignancies were attributed to anti-TNF therapy irrespective of drug discontinuation. In other words, malignancies occurring both during active anti-TNF therapy and after stopping anti-TNF therapy were attributed to the anti-TNF cohort. Patients ever treated with a non–anti-TNF biologic drug (e.g., anakinra or rituximab) were excluded completely from the analysis. Patients initially registered in the comparison cohort who subsequently received an anti-TNF drug contributed person-years to the comparison cohort up to the date that the anti-TNF drug was started, and contributed subsequent followup to the anti-TNF cohort.

Patients could develop a second or subsequent malignancy following the date of their first incident malignancy. Followup was therefore not censored at the time of incident cancer diagnosis. However, a sensitivity analysis was conducted, excluding time and events after the first malignancy. The influence of time since prior malignancy was examined by stratifying the data according to time since prior cancer (more or less than 10 years).

Malignancy rates are presented as events/1,000 person-years with 95% confidence intervals (95% CIs). Incidence rate ratios (IRRs) were calculated using Cox regression, comparing between the anti-TNF cohort and the DMARD cohort. Adjustment was made for age and sex. A propensity score was calculated based on age, sex, disease duration, baseline DAS28 and HAQ scores, year of entry, and smoking status. Propensity-adjusted estimates were calculated by stratifying the propensity score into deciles. Multiple imputation was used to avoid bias caused by missing data (there were 3 subjects with missing disease duration, 5 with missing DAS28, and 32 with missing HAQ score). Nelson-Aalen cumulative incidence plots were generated to explore time-dependent incidence in the two cohorts.
### Raw Data

1. Does the patient have rheumatoid arthritis?  
   - Yes  
   - No

   If NO, can you specify the other diagnosis?

2. Please complete the following details:
   - Year of diagnosis
   - Year first seen by a rheumatologist

### Analysis package

<table>
<thead>
<tr>
<th>studyid</th>
<th>start</th>
<th>stop</th>
<th>serious</th>
<th>entry</th>
<th>biexem</th>
<th>pagen</th>
<th>discare</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12/01/2003</td>
<td>15/07/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>17/01/2003</td>
<td>14/07/2003</td>
<td>0</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14/07/2003</td>
<td>20/07/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16/07/2003</td>
<td>15/08/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15/08/2003</td>
<td>12/16/2003</td>
<td>0</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12/16/2003</td>
<td>20/16/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>12/19/2003</td>
<td>27/02/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>27/02/2003</td>
<td>17/03/2003</td>
<td>1</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12/07/2003</td>
<td>02/08/2004</td>
<td>0</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>02/08/2004</td>
<td>30/11/2004</td>
<td>0</td>
<td>1</td>
<td>female</td>
<td>3.9501263</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>15/11/2004</td>
<td>01/02/2005</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>01/02/2005</td>
<td>21/02/2005</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>12/07/2004</td>
<td>12/07/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>12/07/2004</td>
<td>21/02/2005</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>13/02/2005</td>
<td>30/11/2005</td>
<td>0</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>15/01/2003</td>
<td>15/01/2004</td>
<td>1</td>
<td>1</td>
<td>male</td>
<td>6.4471757</td>
<td>57</td>
<td></td>
</tr>
</tbody>
</table>
Beyond STROBE

Rothman: Modern Epidemiology

• Data editing (2/ 738 pages)
  – Impossible/ unusual values
  – “Correction of such errors where possible”

PubMed search

• (Data cleaning OR data preparation OR data editing) [tiab] AND epidemiology

• 38 results
  – 2 assessing the impact of editing/ cleaning
Importance of data cleaning

• “What is the association between oral glucocorticoid therapy and incident diabetes in patients with rheumatoid arthritis?”
Clinical Practice Research Datalink

- Previously known as General Practice Research Database (GPRD)
- Anonymised primary care electronic patient records from 8% UK population
- >900 international peer reviewed publications
Defining drug exposure

Raw Data

Data prepared for analysis: start and stop dates
Defining drug exposure

\[ \pi_{ij} = \sum_{k=1}^{K} \xi_{ik} \phi_k(z_{ij}). \]
Defining drug exposure

• Can identify all steroid prescriptions
• Have prescription date (start date)
• Several sources of duration of exposure
  – Coded information on duration
  – Dose duration from free text
    • “Take one daily for four weeks”
  – Calculated from total quantity/ quantity per day
Defining drug exposure

1. Selecting stop date

Stop date = Start date +
#1. numdays
or
#2. dose_duration
or
#3. quantity/ dailytabs

If 2 or 3 conflicting options, take midpoint if difference less than...

1 week
2 weeks
4 weeks
> 4 weeks, set as missing

Decision nodes
Defining drug exposure

1. Selecting stop date
2. Handling missing stop date

Drop prescription

Assume average duration from individual’s prescriptions

Assume average duration from population’s prescriptions

Decision nodes
Defining drug exposure

1. Selecting stop date
2. Handling missing stop date
3. Overlapping prescriptions

Decision nodes

- Ignore overlap
- Append overlap interval
Defining drug exposure

1. Selecting stop date
2. Handling missing stop date
3. Overlapping prescriptions
4. Small treatment gaps

Decision nodes

Keep gap

Assume continuous use if gap less than...

- 7 days
- 14 days
- 21 days

Weeks: 0 2 4 6 8 10 12

Prescription 1

Gap

Prescription 2
1. Selecting stop date
2. Handling missing stop date
3. Overlapping prescriptions
4. Small treatment gaps

Defining drug exposure

\[ \pi_{ij} = \sum_{k=1}^{K} \xi_{ik} \phi_k(z_{ij}) \]
Defining drug exposure

1. Selecting stop date
2. Handling missing stop date
3. Overlapping prescriptions
4. Small treatment gaps

192 different datasets

\[ \pi_y = \sum_{k=1}^{K} \xi_k \phi_k(t_{ij}) \]
Cohort

• 32,763 patients with RA
• 192,488 person years of follow-up
• 3,319 cases of incident diabetes
• 757,403 prescriptions for oral steroids
• 409,000 within study period
• 22,412 (68%) subjects with at least one steroid prescription
Defining drug exposure

Impact of different datasets on results

Relative Risk 1.17 to 3.68

Number needed to harm varies from 1 in 24 to 1 in 370

1. Selecting stop date
2. Handling missing stop date
3. Overlapping prescriptions
4. Small treatment gaps

Lunt et al. ICPE 2013
## Node one: selection of stop date

<table>
<thead>
<tr>
<th>Source of stop date</th>
<th>Person years of exposure</th>
<th>Rate in unexposed</th>
<th>Rate in exposed</th>
<th>Rate ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescription file #1: numdays</td>
<td>16</td>
<td>17.2</td>
<td>63.4</td>
<td>3.68</td>
</tr>
<tr>
<td>Text file #2: dose_duration</td>
<td>1,890</td>
<td>17.2</td>
<td>20.1</td>
<td>1.17</td>
</tr>
<tr>
<td>Calculation #3: quantity/dailytabs</td>
<td>18,934</td>
<td>16.6</td>
<td>23.3</td>
<td>1.41</td>
</tr>
<tr>
<td>Any of the above</td>
<td>19,131</td>
<td>16.5</td>
<td>23.7</td>
<td>1.44</td>
</tr>
</tbody>
</table>

Lunt et al. ICPE 2013
Summary

• Decisions in data preparation are important
  – Vast differences in RR, despite same raw data

• Need for greater transparency
Next steps (1)

• Developed code that allows selection of (currently) 15 decision nodes in CPRD
  – Testing in cohort of patients with diabetes
Next steps (2)
Incorporating free text

• CPRD: Free text transformed into coded data:
  – Number of tablets
  – Dose frequency
  – Dose intervals
## Next steps (2)

### Incorporating free text

<table>
<thead>
<tr>
<th></th>
<th>Dose number</th>
<th>Dose frequency</th>
<th>Dose interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take one 3 times a day with or without food</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Take one or two tablets when needed up to four times per day</td>
<td>1.5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Next steps (2)

Incorporating free text

• CPRD: Free text transformed into coded data:
  – Number of tablets
  – Dose frequency
  – Dose intervals

• Text mining to recode as
  – Dose number min & max
  – Dose frequency min & max
  – Dose interval min & max
Next steps (2)
Incorporating free text

<table>
<thead>
<tr>
<th>Description</th>
<th>Dose number (Min)</th>
<th>Dose number (Max)</th>
<th>Dose frequency (Min)</th>
<th>Dose frequency (Max)</th>
<th>Dose interval (Min)</th>
<th>Dose interval (Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take one 3 times a day with or without food</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Take one or two tablets when needed up to four times per day</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1-2 tabs 4-hourly pain max 8 tabs in 24 hours</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 pufftds</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Apply three or four times daily</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 a month</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Month</td>
<td>Month</td>
</tr>
</tbody>
</table>
All data preparation considerations

1. Defining prescription intervals
2. Defining dose (including text mining)

>50,000 different drug matrices
Next steps (3)

- Developed code that allows selection of (currently) 15 decision nodes in CPRD
- Incorporate text mining
- Developing means to share code in Health eResearch Centre
Conclusion

• Data preparation is important

• Analysis can only be reproduced if decisions are explicit

• Need to move towards transparency in data preparation (and means to do this)
**Step by step**

1. From the estate office, head towards the train station and then onwards along the canal until you reach the Standedge Tunnel visitor centre.

2. Head up to Waters Road. Walk up the drive beside the Tunnel End pub and then take the path behind it. Walk across fields through a green gate to a group of houses.

3. Go through another gate beside the house on the right to bring you out on a lane. Turn left and follow it past Berry Greave. When it starts to descend downhill, turn right and follow this lane past the first house and round to a path at the back of the next house. Walk across the fields and at the small footbridge keep to high ground before you drop sharply down to a stream and back up again to the right of the phone mast. Turn right down Blake Lee Lane past Lower Green Owlers to a road junction. Turn left and climb up the track to White Hall Farm. In front of the farm go through the gate leading on to the moor and through some rushes, following the path along the edge of Haigh Clough to the southern corner of March Haigh reservoir.

4. Here you have the option to head up to Buckstones car park for magnificent views across the estate. Cross the dam wall and take the path through the rocks to the car park. Once you’ve caught your breath, retrace your steps back to the dam wall.

5. Continue to head south to meet up with the Packhorse Trail, where you turn right and follow the trail across Willmer Clough and up to the top to one of the old marker stones. Go straight ahead until you reach Haigh Gutter.
Acknowledgements

• Mark Lunt
• Therese Sheppard
• Mohammad Movahedi
• Goran Nenadic

Funders

McGill

• Robyn Tamblyn
• Nadyne Girard