Data Sharing in Clinical Trials

Practical guidance on anonymizing trial datasets

C Tuck, S Lewis, G Milne, C Keerie
Today’s talk

• Background to data sharing

• Current guidance

• Our work on anonymization

• How we use Datashare
Data sharing (1)

Trial Dataset

- Study team
- Known collaborations
- Reg. Authorities
- Unspecified researchers
Data sharing (2)

• Usually at the end of the study

• Sharing data with unspecified secondary researchers

• Sharing data to the individual participant level
Anonymization (1)

- Anonymization and deidentification used interchangeably

- “information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable” [Recital 26]

- Allows for wider use of information
Anonymization (2)

<table>
<thead>
<tr>
<th>Trial ID</th>
<th>Date of Enrolment</th>
<th>Initials</th>
<th>Age at Enrolment</th>
<th>Postcode</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>04/05/2012</td>
<td>JB</td>
<td>67</td>
<td>EH1</td>
<td>M</td>
</tr>
<tr>
<td>1002</td>
<td>05/05/2012</td>
<td>AS</td>
<td>56</td>
<td>OX4</td>
<td>M</td>
</tr>
<tr>
<td>1003</td>
<td>12/05/2012</td>
<td>MM</td>
<td>72</td>
<td>IP2</td>
<td>M</td>
</tr>
<tr>
<td>1004</td>
<td>14/05/2012</td>
<td>EW</td>
<td>86</td>
<td>SW1A</td>
<td>F</td>
</tr>
<tr>
<td>1005</td>
<td>19/05/2012</td>
<td>MS</td>
<td>82</td>
<td>KT6</td>
<td>F</td>
</tr>
<tr>
<td>1006</td>
<td>19/05/2012</td>
<td>TR</td>
<td>79</td>
<td>EH4</td>
<td>M</td>
</tr>
<tr>
<td>1007</td>
<td>21/05/2012</td>
<td>AW</td>
<td>65</td>
<td>IP3</td>
<td>F</td>
</tr>
</tbody>
</table>
The balancing act

- Funders
- Publishers
- Other researchers
- Open Access Groups
- Privacy Groups
- Protection Act
- Sponsor

Researchers
MRC guidance (1)

- Aimed at CTUs
- Recommends a controlled access model
- Published April 2015

Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials.
MRC guidance (2)
The test case

- TOPPIC trial
- 240 participants

Mercaptopurine versus placebo to prevent recurrence of Crohn’s disease after surgical resection (TOPPIC): a multicentre, double-blind, randomised controlled trial


Summary

Background Up to 60% of patients with Crohn’s disease need intestinal resection within the first 10 years of diagnosis, and postoperative recurrence is common. We investigated whether mercaptopurine can prevent or delay postoperative clinical recurrence of Crohn’s disease.

Methods We did a randomised, placebo-controlled, double-blind trial at 29 UK secondary and tertiary hospitals of patients (aged >16 years in Scotland or >18 years in England and Wales) who had a confirmed diagnosis of Crohn's...
Preparing the dataset

• What to remove?

- 28 participant identifiers (Hrynaszkiewicz & colleagues)
- Both direct and indirect identifiers to consider
## Potential identifiers examples

<table>
<thead>
<tr>
<th>Direct Identifiers</th>
<th>Indirect Identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Place of treatment</td>
</tr>
<tr>
<td>Initials</td>
<td>Sex</td>
</tr>
<tr>
<td>Address, including full or partial postal code</td>
<td>Rare disease or treatment</td>
</tr>
<tr>
<td>Dates related to an individual (inc. date of birth)</td>
<td>Year of birth or age</td>
</tr>
<tr>
<td>Unique identifying numbers</td>
<td>Small denominators - population size of &lt;100</td>
</tr>
<tr>
<td>Medical device identifier</td>
<td>Very small numerators - event counts of &lt;3</td>
</tr>
</tbody>
</table>
Preparing the dataset

• What to remove?

• 28 participant identifiers (Hrynaszkiewicz & colleagues)

• Both direct and indirect identifiers to consider

• Also – remove superfluous data (e.g audit)
Anonymisation process

1st Pass
- Assess data variables & assign value
- Decide on anonymisation method for direct identifiers

2nd Pass
- Determine indirect identifier risk
- Decide on anonymisation method for high risk indirect identifiers

Test
- Check dataset for successful anonymisation and utility

Release
Anonymisation process

1st pass

• Variables coded and assigned value
• Direct identifiers 01-14 (+15, superfluous)
• Indirect identifiers A-N

1st pass of data dictionary

• Either direct/superfluous, indirect, or not requiring modification
• Direct identifiers – assign method of anonymization
• Indirect identifiers flagged
# 1st pass example

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
<th>Data type</th>
<th>Require anon (Y/N/?</th>
<th>Reason code</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubjectNo</td>
<td>Subject Number.</td>
<td>int</td>
<td>Y</td>
<td>06</td>
<td>Recode</td>
</tr>
<tr>
<td>AEDescription</td>
<td>description of the AE</td>
<td>int</td>
<td>?</td>
<td>C, N</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td></td>
<td>int</td>
<td>?</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>StartDD</td>
<td>Start Date (day)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>StartMMM</td>
<td>Start Date (month)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>StartYYYY</td>
<td>Start Date (year)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>AECategory ID</td>
<td></td>
<td>int</td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reason code 06 = Unique identifying number  
Reason code 14 = Dates related to an individual  
Reason code C = Rare disease or treatment  
Reason code N = Verbatim responses or transcripts
Study day modification

• Use date of randomisation as day 0

• All other dates relative to date of randomisation

• e.g. date of randomisation 15/01/2014, start date of AE 16/01/2014

• New study date of AE = 1
Anonymisation process 2\textsuperscript{nd} pass

- Indirect identifiers – decision to anonymise or leave alone

- Currently use a consensus model

- Some summarized to determine risk (event counts)

- May need medical input
### 2\textsuperscript{nd} pass example

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
<th>Data type</th>
<th>Require anon</th>
<th>Reason code</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubjectNo</td>
<td>Subject Number</td>
<td>int</td>
<td>Y</td>
<td>06</td>
<td>Recode</td>
</tr>
<tr>
<td>AEDescription</td>
<td>description of the AE</td>
<td>int</td>
<td>?</td>
<td>C, N</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td></td>
<td>int</td>
<td>?</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>StartDD</td>
<td>Start Date (day)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>StartMMM</td>
<td>Start Date (month)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>StartYYYY</td>
<td>Start Date (year)</td>
<td>varchar</td>
<td>Y</td>
<td>14</td>
<td>Study day</td>
</tr>
<tr>
<td>AECategory ID</td>
<td></td>
<td>int</td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reason code 06 = Unique identifying number  
Reason code 14 = Dates related to an individual  
Reason code C = Rare disease or treatment  
Reason code N = Verbatim responses or transcripts
Dataset release

• Is it anonymous?
  • Motivated intruder test
  • May not be required for every dataset (risk based approach)

• Is it useful?
  • Re-run analysis with modified dataset
  • Further QC checking might be required
Summary on anonymization

- Generic rules can be created for direct identifiers

- Decisions on indirect identifiers on a trial-by-trial basis

- Balance anonymisation, data utility and practicality
Using Datashare

Citation

Description
Anonymised TOPPIC trial dataset

- Protocol No MRC G060329 Version 12 03 October 2013.pdf (1.167Mb)
- TOPPIC Anonymised data dictionary.pdf (305.6Kb)
- DATASET IN CSV FILES (566.7Kb)
- TOPPIC -Annotated CRFs.pdf (3.644Mb)
Application process

- **Apply**
  - Standard form (credentials, research question)

- **Review**
  - Committee review (ECTU + Chief Investigator)

- **Contract**
  - Data Access Agreement

- **Release**
Questions?

• Thanks to: