Data extraction (coding)

Hilary Thomson
Data extraction

• Plan for this in protocol
  – List key fields that you plan to extract

• Data extraction “forms”
  – Sometimes paper but electronic more manageable
  – Likely to need changes/additions during review process

• Data extraction/coding
  – Coding assumes able to pre-specify all categories in advance
  – Rarely possible or appropriate for social interventions
Double extraction

- Extraction & assessment of study quality should be done by two reviewers independently and compared to minimise bias

- Disagreements resolved by discussion

- At the minimum study quality should be performed by two reviewers & checked as this may involve reviewer judgements

- Data extraction may be done by one reviewer and checked by a second reviewer for errors
Data extraction

- Population: age, sex, ethnic group etc
- Intervention
- Comparator
- Outcomes: can be multiple outcomes, may need to prioritise
- Context
- Study design

- Study quality items
Development of extraction form

• Types of Information to extract
  – Setting, study context, authors, publication date and type, etc.
  – Methods and method quality
  – Program/intervention
  – Participants/clients/sample
  – Outcomes
  – Findings, effect sizes
Types of Information to extract

• Setting, study context, authors, publications date and type, etc.
  – Multiple publications; “study” vs “report” “paper”
    • Assign study name (may cover >1 paper)
  – Geographical/national setting; language
  – Publication type
  – Decide re publication date vs study date
Study design: include/exclude?

• Randomised: cluster/individual

• Non-randomised designs
  – Agree consistent labels for designs
  – Do not assume authors use correct or same labelling system
  – (see earlier best available evidence session)
    • Retrospective/Prospective
    • Cohort/cross-sectional
    • Change in outcomes
    • Assignment/allocation of intervention
    • Selection of participants
Study design

• For experimental or quasi-experimental designs on interventions:
  – How were participants assigned/allocated to intervention?
  – Crossovers, dropouts, other changes to exposure to intervention
  – Nature of control condition
  – Multiple intervention and/or control groups
  – Initial comparability of groups prior to experiment
    • Baseline characteristics (socio-demographics)
    • Baseline outcomes
Example questions about RCTs

1. How were comparison/control groups formed?
   • Random assignment
   • Other (specify)

2. If random assignment, specify design:
   • Simple/systematic (individuals/families)
   • Stratified/blocked (identify stratifying variables)
   • Yoked pairs (created by timing of enrollment into the study)
   • Matched pairs (identify matching variables)
   • Cluster (group) randomized
   • Other (specify)
   • Can't tell
More on experimental research design

3. Who performed group assignment?
   - Research staff
   - Program staff
   - Can’t tell
   - Other (specify)

4. How was random assignment performed?
   - computer generated
   - random numbers table
   - coins or dice
   - other (describe)
   - can't tell

5. Was allocation sequence concealed from researchers?
Types of data to extract

• Methods: Other aspects
  – Issues depend on specific research area
  – Procedural, e.g.,
    • monitoring of implementation, fidelity
    • credentials, training of data collectors
  – Statistical, e.g.,
    • statistical controls for group differences
    • handling of missing data
Data extracting for complex interventions

- Use logic model to identify priorities for what data should be extracted
  - Additional outcomes- but remember you can’t extract everything!
Intervention details

• General program type

• Specific program elements (present/absent)

• Any treatment/contamination received by the comparison group

• Treatment implementation issues
  – Integrity: variation across the sample
  – amount, “dose”
Sample/population details

• Participants/clients/sample
  – Data are at aggregate level (are data reported for neighborhoods or schools?)
  – Mean age, age range
  – Gender mix
  – Racial/ethnic mix
Extracting outcome data

• Extract SE, sd, and confidence intervals, sample sizes, and variables adjusted for where reported

• Compute standardised effect sizes where possible
  – Relative Risk Ratio, Odds ratio, standardised mean difference

• May need to aggregate data or reconfigure findings
  – Compute weighted means of subgroups (e.g., boys and girls)
Extracting findings

• Extract all relevant outcome data

• Make sure consistent (e.g. exact p values or rounded up? Use of abbreviations & symbols *, **): a data extraction protocol is useful for this to ensure consistency over time and across reviewers

• Keep data raw where possible to avoid having to look at paper again

• Note where data unclear or not reported, e.g. only narrative reporting a finding

• Where data reported more than once (e.g. multiple papers) and it does not agree, make decision about what data to use (decide on this and record in protocol)
Example defaults

- Use ℞ where the figures for sample sizes do not tally across the tables or where sample sizes unclear

- ~estimated sample size/figure- use this where ORs calculated from data but where no indication of missing data by variable (only available by whole sample)

- Φ more data available (see fuller table/full data extraction/refer to paper)

- Time 0- baseline Time I- first follow-up etc

- Abbreviation: Int = intervention group; Cont= control group

- Report exact p values to three decimal places where available
<table>
<thead>
<tr>
<th>Author1</th>
<th>Int Summ COCHRANE</th>
<th>Author</th>
<th>Publication Year</th>
<th>Country</th>
<th>Description of Intervention</th>
<th>Summary FOLLOW-UP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wells</td>
<td>Rehousing or retrofitting with or without neighbourhood renewal (after 1995)</td>
<td>Wells N.</td>
<td>2000</td>
<td>USA</td>
<td>participants moved to newly constructed (n=29) or renovated homes (n=3). The new or renovated homes were in excellent condition and they had enough room for everyone (at most 2 spouses or siblings shared bedroom).</td>
<td>Twice: 5-12 months &amp; 2-3 years since intervention</td>
</tr>
<tr>
<td>Spiegel et al</td>
<td>Provision of basic housing needs/low or middle income country intervention</td>
<td>Spiegel, J., et al</td>
<td>2003</td>
<td>Cuba</td>
<td>External housing repair, voluntary internal repairs by residents (provision of cheap materials), Neighbourhood improvements and provision of social and cultural activities</td>
<td>Once: between 1-4 years since intervention, 5 years since baseline</td>
</tr>
<tr>
<td>Blackman and Harvey</td>
<td>Rehousing or retrofitting with or without neighbourhood renewal (after 1995)</td>
<td>Blackman</td>
<td>2001</td>
<td>UK</td>
<td>£5.5mll environmental improvements, external fabric repairs, refurbishment or demolition of void dwellings, discretionary renovation grants for individual dwellings, heating and security improvements, landscaping, environmental improvements, security and road safety measures (traffic calming), footpath improvement, work on walls. (p573/4)</td>
<td>Once: 5 years since intervention</td>
</tr>
<tr>
<td>Ambrose</td>
<td>Rehousing or retrofitting with or without neighbourhood renewal (after 1995)</td>
<td>Ambrose</td>
<td>1999</td>
<td>UK</td>
<td>Rehousing into better accommodation, or had existing accommodation improved plus SRB neighbourhood improvements plus other employment and education initiatives related to SRB</td>
<td>Once: ~4-4.5 years since baseline</td>
</tr>
<tr>
<td>Somerville et al</td>
<td>Warmth and energy efficiency improvements (after 1980)</td>
<td>Somerville et al</td>
<td>2000</td>
<td>UK</td>
<td>installation of gas central heating 28/59 (47%), electric storage heater 22/59 (37%), Solid fuel central heating 7759 (17%), oil-fired central heating 2/59 (4%), £2,500 to improve heating and reduce damp and mould growth in house- the improvements required were decided by individual housing officers, generally installation of central heating generally not gas central heating as mains gas not available in many areas.</td>
<td>Once: 3 months since intervention</td>
</tr>
<tr>
<td>Allen</td>
<td>Warmth and energy efficiency improvements (after 1980)</td>
<td>Allen (1st)</td>
<td>2005</td>
<td>UK</td>
<td>GP selected patients who had a serious heart condition and whose housing was likely to be poor, health and benefits advice given at baseline assessment. At baseline assessment house was assessed using HHSRS and propose required work.</td>
<td>Once: 3 years since intervention &amp; baseline</td>
</tr>
</tbody>
</table>
Data extracting for complex interventions

• Evaluations of complex interventions often reported in large reports rather than 2000 word journal paper

• Not neat randomised controlled trial reported in standard way
  – Often poorly reported. For example unclear sample size, time of follow-up, statistics incomplete, graphs with %
  – Conflicting data across the findings- in same paper or across multiple papers

• This can be very time consuming (and very confusing!)
  – Agree with co-reviewers how to deal with key issues
    • Exclude poorly reported studies?
    • Prioritise later publications of same study? Peer review publications?
Data extraction: summary

- Population: age, sex, ethnic group etc
- Intervention
- Comparator
- Outcomes: can be multiple outcomes, may need to prioritise
  - ITT or TOT: depend on your review question
  - ITT to assess effectiveness of programme
  - TOT to assess efficacy of active component
- Context

- Study quality items

- May need to summarise to reproduce in a table- think of the reader and purpose of the review