Introduction to Research Methods

Session 4: Tools to run a Research Project

Prof. David Taylor
Director of Emergency and General Medicine Research, Austin Health
Chair, Austin Health Human Research Ethics Committee
Overview:

1. Sources of Data
2. Retrospective data collection
3. Case Report Forms
4. Surveys
5. Databases and spreadsheets
6. Data validation and cleaning
1. Sources of Data:

- Essentially, biomedical research revolves around the collection and analysis of data.
- Pre-existing sources of data are frequently overlooked.
- Databases can be a wealth of information and may obviate the:
  - time
  - expense
  - complications associated with data collection.
1. Sources of Data:

Existing Databases

- World Health Organization (WHO) statistics
- Australian Bureau of Statistics
- Victorian Emergency Minimum Dataset (VEMD)
- Victorian Admissions Minimum Dataset (VAMD)
- Government agencies
- Coroners’ reports
- Cancer registries
- Victoria Police statistics
- Insurance company statistics
1. Sources of Data:

Informal sources of data:

• Medical Records
• Ambulance case notes
• MCG ticket sales records
• Television viewing ratings
1. Sources of Data:

Strengths of existing databases

• They are usually very specific to topic
  eg. for mortality data use the Death Registry

• Some databases are necessarily complete
  eg. Coroners’ reports must be complete by law

• They are often readily accessible to authorized personnel

• The data has already been collected and may save
  considerable time and expense
1. Sources of Data:

Weaknesses of existing databases

- incomplete, missing data, or not actually collected
- data is available as summary statistics only
- data may be quite inaccurate
  - depends upon the vigilance/training of those extracting the data/entering
- data not collected for research purposes and biased
  - eg. measurement, selection, interviewer biases
- some databases limit accessibility for confidentiality, business interest or even politics!
  - eg. skiing injuries
1. Sources of Data:

• The limitations with existing databases should not preclude their use

• Such data is often most useful in the:
  • identification of clinical or social problems
  • generation of research questions

• The next step is to design and undertake more powerful research studies that aim to answer the questions
2. Retrospective data collection:

- Medical records abound with clinical data
- Complexities involved in obtaining high quality data often overlooked
- Data subject to:
  - Ambiguity
  - Variability in clinician interpretation
  - Omission
  - Error
  - Illegibility
2. Retrospective data collection:

• In turn, data from medical records subject to:
  • selection bias
  • variability in interpretation and handling of uncertain or missing data
  • errors in transcription
  • chart availability

• Some difficulties minimised by adopting methodological strategies to improve the quality
2. Retrospective data collection:

Chart Acquisition:

- Determine the procedure for selecting the records
- Should align with inclusion criteria
e.g. use EDIS to list all ED presentations with:
  - ICD 10 codes S42 and S52
  - from 1/1/09 to 31/12/09
Preliminary Review:

Review a sample to:

• determine if the desired data are available

• to identify difficulties that might arise
  
  e.g. follow up data may be difficult to collect if some patients follow-up in private
2. Retrospective data collection:

Develop a data collection form:

- Consultation with a statistician
- Trial and revise
- Ease of entry of data both *in* and *out* of the form
Define variables precisely:

- **Thresholds**
  e.g. fever = temperature $> 37.5^\circ C$

- **Precision**
  e.g. $25^\circ C$ or $25.6^\circ C$

- **Source designation**
  e.g. vital signs on arrival from nursing not medical notes

- **Time frame**
  e.g. lowest BP between triage and discharge, *not* from ambulance

- **List of acceptable synonyms**
  e.g. altered consciousness may appear as GCS $<15$, confusion, disorientation
Determine uniform methods:

- Especially for handling of data that is conflicting, ambiguous, missing, or unknown

- Differentiate between “not recorded” and “not present” e.g. fever should not be entered as “Yes” or “No”
Data collector training

- Components:
  - orientation to the sources of information in the chart
  - chart review protocol
- Opportunity for practice and feedback
- Maximum of four abstractors
  - avoids complex inter-rater reliability checks
- Describe the qualifications of the abstractors
2. Retrospective data collection:

Pilot test the chart review protocol

• Include all phases of the extracting process
• Assists by confirming:
  • data availability
  • appropriateness of the procedures
  • data collector performance
• Provides feedback from the data collectors which may identify:
  • confusing and misleading questions
  • unforeseen ambiguities or conflicts
  • inefficiencies in data collection form layout
Blinding of Data

Collectors should be blinded to:

• the study hypothesis
• the patient’s group assignment
  -if groups are being compared

2. Retrospective data collection:
2. Retrospective data collection:

Inter-rater reliability

- The percentage of agreement when multiple people collect data from the same charts
- Blinded to the information obtained by the other data collectors
- Report the $\kappa$-statistic
  - a measure of inter-rater agreement accounting for agreement expected by chance alone
- When only one abstractor, abstract a number of charts twice
Many a study has foundered because wrong data was collected or important data was not collected e.g. weight not collected in a lipid study.

Data usually grouped as:
- Identification data
  - personal information needed to link to the appropriate patient
- Research data
  - information analysed to answer the study question
- Administrative data
  - initials of the data collector, the study centre, etc.
3. Case Report Forms:

- Collect only the research data that is essential to answer the study objective.

- Avoid collecting data that will not be of use:
  - Time-consuming
  - Expensive
  - May detract from the quality of the remaining data.
3. Case Report Forms:

Other points:

- Separate identification and research data
- Patient ID code on every page
- Provide special instructions where necessary
- Use self-coding forms where necessary
  
  e.g. What is the sex of the patient?  
  1=Male  
  2=Female  
  9=Unknown  

- Character separators, pre-print decimal points and units
  
  e.g. height: .□□□ metres
Important:

• Check completeness of data collection
• Usually one chance to collect data from study subjects e.g. questionnaire
• Imagine sitting down to analyse 1000 questionnaires only to realise that 30% of the subjects had not realised that half of the questions were on the reverse side of the page!! Yes, this has happened!
3. Case Report Forms:

Data confidentiality:

• Information collected must be used only for that study

• Remove of personal identifying information from questionnaires and computer files

• Master List:
  • links study ID codes with individuals
  • stored separately from research data

• No identifiable data should be published
4. Surveys:

Survey Design

• One of the most difficult steps in clinical research

Principles:
• Begin with a clear idea of the research questions
• Design questions based on principles of survey design
• Keep in mind how the responses to specific questions will be used in analysis of the survey
Think:

- Is a survey the best way to collect the data that will address your question?
- List the major areas of interest (domains) to your study
- In writing each question, consider:
  - How will you code the answer?
  - How will you use it in analysis?
  - How does it relate to the other questions you are considering?
Cover Letter:

All questionnaires should have one:

• introduce and explain the survey
• the first thing the respondent sees!
• keep it simple
• no longer than ONE page
• anticipate respondents’ questions
• provide information in your covering letter accordingly
4. Surveys:

Cover Letter format:

• 1st paragraph: purpose and importance of the study
• 2nd paragraph: explain confidentiality will be maintained
• 3rd paragraph: re-emphasise study reason/justification
• 4th paragraph: explain if results will be made available
• Finally: THANK THE PARTICIPANT!
4. Surveys:

- Incorporate *commonly asked questions* from standard surveys
  - Australian Bureau of Statistics surveys
  - measures on quality of life
  - functional ability
  - disease-specific symptoms

- The scientific literature is a good start

- Previously used questionnaires for similar topics:
  - are very helpful
  - often can be used directly
  - reliability and validity are established

- Particularly important for comparing your results with others
4. Surveys:

Open-ended questions

- no answer choices to select from
- the respondent must do more work in formulating answers

Examples:
- How satisfied are you with the care provided in this hospital?
- What are the barriers to good care in this hospital’s ED?
Open-ended questions

**Advantages:**
- Stimulates free thought
- Most useful in exploratory (also qualitative) studies

**Disadvantages:**
- Creating answers may be difficult
- Answers may be incomplete, irrelevant
- May be difficult to analyse using quantitative techniques
Closed-ended with Ordered Choices

- answers provided
- may have gradations of a dimension of thought or behaviour
- need to find answer that most closely matches their beliefs

Example:
- How often have you come to this ED in the last six months?
  - once
  - twice
  - etc
Closed-ended with Ordered Choices

**Advantages:**
- Readily analysable
- Useful for assessing gradations of intensity (feelings, pain)

**Disadvantages:**
- May limit range of answers to a single dimension
- Forces answers in a specific way
4. Surveys:

Closed-ended with Unordered Choices

Example:

• During the past week on which of the following have you spent the most time?
  • Doing exercises prescribed by the physiotherapist
  • Resting in bed
  • Walking within the house
  • Walking outside the house
Closed-ended with Unordered Choices

**Advantages:**
- not limited to choosing among gradations of a single dimension
- useful for selecting priorities (policy or decision making)

**Disadvantages:**
- more difficult than ordered choices
- must select an answer when more than one may be appropriate
4. Surveys:

Partially closed Questions

*Example:*

- Which of the following areas of expenditure do you want to have the highest priority for improvement in this hospital?
  - Emergency Department
  - X-Ray Department
  - Laboratories
  - other (please specify) .................................
Partially closed Questions

Advantages:
• Provides an ‘out’ if certain responses are overlooked
• If ‘other’ is used frequently, may signal a faulty question (or responses!)

Disadvantages:
• ‘other’ may not add much
• bias toward answering with one of the given options
Pointers for questionnaire design:

• Questions flow in a logical, conversational tone
• Go from:
  • general to specific
  • impersonal to personal
  • easy to difficult
• Wording should facilitate recall and keep respondent interested
• Avoid questions that are:
  • difficult to answer
  • time consuming
  • embarrassing
  • threatening to answer
• Respect the respondent and their willingness to participate!
4. Surveys:

Pointers for questionnaire layout:

• Neatly balance the layout
• Ensure that it is easy for you to use, code and keep together
• Leave SPACE between items
• Use A4 size paper
• Number the pages consecutively
• Do not split questions on two pages
• Use COLOUR!
• Ensure instructions distinguishable from questions e.g. italics
• Use arrows, boxes to keep interesting
• Consider pre-coding questions or scanability
4. Surveys:

Pre-testing of a questionnaire is most important:

• Assess face validity of all questions
• Is the wording clear?
• Do different people have similar interpretations of questions?
• Do closed-ended questions have appropriate possible answers?
• Does the questionnaire give a positive impression?
• Is there any bias in the questions?
4. Surveys:

Pre-testing

• Check with your colleagues to ascertain if the questionnaire will answer the study question

• Trial on a cross-section of potential respondents of differing reading levels and background
  • amazing how many surprises you get
4. Surveys:

Quality of the Instrument: Validity

- Validity is the ability of the questions (interview or questionnaire) to capture the underlying concepts being evaluated.

- Does the test measure what it is intended to measure? (akin to sensitivity)

- What can we conclude about the person who produced a particular score on a test?
4. Surveys:

- **Face Validity**
  Do the questions make sense?

- **Content Validity**
  Do the questions express the underlying concept they were designed to reflect?

- **Criterion Validity**
  Do the responses to the questions agree with an objective criterion or gold standard for the underlying concepts?

- **Construct Validity**
  Are the hypotheses concerning the relationships between the underlying concepts borne out by the responses?
Quality of the Instrument: Reliability

Various estimates of reliability are used:

• **Test-retest:**
  Does the same question have the same response over time?

• **Inter-rater:**
  Do 2 interviewers with the same questionnaire get the same response?

• **Internal consistency:**
  Do questions designed to evaluate the same concept obtain equivalent responses?
Data management deals with the whole process of what you do with data once collected

How it is:

• stored electronically
• ‘cleaned’
• manipulated (appropriately)
• analysed
Coding conventions

• use standard conventions if possible
  • familiar to data processors, analysers and statisticians
  • assist in comparing your data with that of others ie. standardisation of data

Examples:

• Unique patient ID e.g. UR number
• Code categorical data using numerical values
  eg. 1=yes, 2=no, 9=unknown
• Use standard codes
  eg. ICD-10, drug codes
The Database

• a specific collection of data that is organised in a structured fashion

• A way of organising the data systematically

• Good database design will:
  • reduce repetitiveness
    eg. entering in an address or age for a patient many times
  • have data in a convenient form for analysis
Database Software

- Some merely allow data to be stored and manipulated
- Others allow sophisticated statistical analysis

Examples:

- Excel – simply a spreadsheet, useful for simple stats
- ACCESS – facilitates data entry and tabulation
- SAS – generally useful for statistical analysis
- SPSS – incorporates data spreadsheets and powerful statistical techniques
Data Entry

• entry of data into the electronic database
• final data set may be inaccurate if the data entry process was inadequate
• relates particularly to manually entered data where mistakes are bound to happen
6. Data validation and cleaning:

Types of data entry:

• Manual entry
  • Single entry: one person only enters all the data
  • Double entry: two independent people enter the same forms and any differences between the two are reconciled.
    - a form of double-checking
    - more time consuming and, therefore, expensive

• Direct data entry
  • Data collection forms on a computer screen allow direct entry
  • Via an Internet page
  • Forms can be fed into a scanner
6. Data validation and cleaning:

Data Validation

- Quality assurance processes that confirm the accuracy of the data during its various phases of the study

- Validation should occur at the following stages:
  - Prior to data entry
  - During data entry
  - Post data entry
6. Data validation and cleaning:

Data validation methods:

- **Visual review**
  - eg. matching data on questionnaire with medical records (source data)

- **Value range checks**
  - that is, do the numbers in the database make sense?
  - eg. cholesterol levels should be > 0 and < 20 mmol/L

- **Field type checks**
  - eg. text should not be entered into numerical field

- **Logical checks (IF, THEN)**
  - eg. if classed as a non-smoker, then cigarettes/day should be 0
6. Data validation and cleaning:

Data management procedures

- Questionnaires kept securely in locked cabinets and rooms
- Identification data should be stored separately
- The database should be password-protected
- Adequate database backup procedures should be in place
  - keep the discs separate from your computer
  - a PhD student who lost 5 years of work when his computer bag was stolen – computer and files and the only backup discs he had!!
- Long-term storage of data:
  - 7 years for observational studies
  - 15 years for clinical trials
Summary

- Remember existing sources of data
- Don’t collect what is already available
- Expend energy before collecting your first data item
- Borrow other people’s instruments
- Questionnaire design is difficult
- Consider and address expected data inaccuracies
- Good data management strategies
- Confidentiality
"Why, thank you. ... Thank you very much!"