Research Methodology

Assoc. Prof. Dr. Jamalludin Ab Rahman
B.Med.Sc., MD, MPH
Department of Community Medicine
Kulliyyah of Medicine
International Islamic University Malaysia

The lectures will answer the following questions

► Why we do research?
► How to plan a scientific research?
► What is the best design for my research?
► How to select subjects for research?
► How to plan for data collection?
► How to manage & analyse data?
► How to report the findings?
Research Methodology Workshop
for Oral Health Division, J KNP
22-24th April 2014

Programme

- Day 1 - Reason, rationale & conceptual framework; Study design & Sampling plan
- Day 2 - Data collection & data dictionary; Statistical analysis & dummy table
- Day 3 - Writing proposal & report
Part 1 – Reasons & Rationale

Why we do research?

- To answer curiosity
- To solve problem
- To find alternative
- To fulfil academic requirements
- To get promotion
When the topic is important?

- When it fulfils the need
- Humankind need
- Country need
- Service need
- But never a personal need

Type of research

- **Descriptive** – e.g. prevalence study
- **Discovery** – e.g. new biomarker, new pathway (mostly lab study)
- **Proof causality (hypothesis testing)** – e.g. cause/causes of a disease or abnormality, reason for certain behaviour, clinical trial (a drug is better than others) etc
The steps

1. Organise research idea
2. Understand the problem → do literature review, construct conceptual framework
3. Formulate objective
4. Know your target population
5. Choose best study design
6. Choose best sampling method
7. Calculate sample size
8. Determine variables to collect → prepare data dictionary
9. Validate research instrument
10. Plan data collection, plan for quality control
11. Plan statistical analysis, prepare dummy table
12. Estimate & secure budget
13. Get approval
14. Collect data, monitor quality
15. Analyse data
16. Report finding

Literature review

► Search strategically
  ► Identify important keywords
  ► Identify authorities in the subject matter
  ► Observe the dates
► Manage bibliographic well – use software (e.g. EndNote, Mendeley, Papers)
Part 2 – Conceptual Framework

Build the concept

1. Main outcomes
2. Explanatory (exposures, factors) variables
3. Confounding variables
Conceptual idea

Example 1

Betel chewing — Oral Cancer — Alcohol based mouth wash

Family history — Smoking
Example 2

Radiation → Lung Cancer → Smoking

Mineral Dust

Peer Pressure

Figure 3 Structural Equation Model Showing the Relationship Between Family Processes, Child Characteristics, and Achievement for Girls Aged 6 to 11 years
Figure 1: A directed acyclic graph (DAG), characterising the direction of selected influences* on risk of developing cardiovascular disease. BP: Blood pressure; CVD: Cardiovascular disease. *Dark font variables are those which are observed, whereas grey font represent unobserved. Drug treatment variables, which we consider colliders, which introduce the influence of unobserved variables not directly linked with the outcome, are underlined.

Group Work #1 (2 hours)

1. Suggest the title of your research
2. State 3 reasons why your study is important
3. Identify the outcome, explanatory factors & the confounders (if applicable)
4. Build the conceptual framework
5. State the specific objectives of the study
Part 3 - Study design

Exposure & Outcome

- Smoking
- High carbo diet
- Oral mouth wash
- Lung cancer
- Diabetes mellitus
- Oral cancer

Population → Exposure → Outcome

History → Future

To proof causation, exposure must precedes outcome
Research design

Observational
- Cross-sectional: Measure exposure & outcome & the same time
- Case-control: Fix the outcome. Measure the exposure
- Cohort: Fix the exposure. Measure the outcome

Experimental
- Animal Trial
- Clinical Trial
- Community Trial

Cross-Sectional

Factor

Outcome
Case Control

Cohort
Experimental study

Hierarchy of Evidence

Population 
Selection criteria 
Population at risk 
Randomisation 
Intervention done 
Outcome observed 
No outcome 
No intervention 
Outcome observed 
No outcome 
No intervention 

Randomised controlled trial 
Cohort 
Case control 
Case report 
Expert opinion 

Meta-analysis 
Systematic review
Cross sectional study

- Population based - represent population
- Measure exposure & outcomes at the same point in time - No temporal association
- Impossible to infer causality
- Prevalence study - measure magnitude or burden of disease
- Descriptive
- Repeated cross-sectional study - pseudo-longitudinal e.g. British Association for the Study of Community Dentistry (BASCD) guidance on sampling for surveys of child dental health. A BASCD coordinated dental epidemiology programme quality standard (Pine et al. 1997)

Cross sectional study

Advantages
- Measure prevalence of a population
- Measure multiple exposures & outcomes
- Relatively inexpensive
- Relative shorter time

Disadvantages
- No temporal association - no inference to causality
- Prevalence-incidence bias (Nyman bias) e.g. if smokers die due to AMI faster, a cross-sectional study will reveal less smoker among AMI patients
- Health workers effect e.g. when survey done from house to house, only health respondent are available in their home/office
Cross sectional study - Example

- NHMS ~ Household study, all Malaysian (N=47,610 for 2006)
- NOHSA ~ Adult (>15) (N=14,444 for 2010)
- NMCS ~ GP vs. PHC (N~12,000)
- NHANES (US) - http://www.cdc.gov/nchs/nhanes.htm

Case control study

- Fix the outcomes, measure the exposures
- Longitudinal
- Retrospective
- Case = outcome of interest
- Control = comparing outcome
Case control study

<table>
<thead>
<tr>
<th></th>
<th>Smoking</th>
<th>Non smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>History</strong></td>
<td>Exposure</td>
<td>No exposure</td>
</tr>
<tr>
<td><strong>Case</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Now</strong></td>
<td>Exposure</td>
<td>No exposure</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy patients</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Lung Cancer</th>
<th>No Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>20 (18.2%)</td>
<td>90 (81.8%)</td>
</tr>
<tr>
<td>Not Smoker</td>
<td>5 (4.5%)</td>
<td>105 (95.5%)</td>
</tr>
</tbody>
</table>

\( \chi^2 (df=1) = 10.150, p = 0.001, OR = 4.7 \) (CI95% 1.7 – 13.0)

Because \( p < 0.05 \), we reject \( H_0 \). Therefore there is a different between smoker & non smoker.
### Cases

- Well defined
- Source – institution vs. population

### Control

- Matched vs. Unmatched
- Matching ~ controls resemble the cases with regard to certain characteristics (age, gender, SES etc)
- Individual vs. Group matching
- Source – institution vs. population
- Ratio to cases ~ up to 4:1

### Case control study

#### Advantages

- Good for rare conditions or diseases
- Less time needed to conduct the study because the condition or disease has already occurred
- Measure multiple risk factors
- Can establish an association

#### Disadvantages

- Recall bias
- Not good for evaluating diagnostic tests because it’s already clear that the cases have the condition and the controls do not
- It can be difficult to find a suitable control group
Example

- Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. (Teo 2006)

Cohort study

- Measure outcomes
- Compare incidence of a disease (or condition) among exposed and unexposed individuals over time
- Disease free at the onset (or inception)
- Repeated measurements ~ follow up
- Prospective vs. retrospective cohort
Cohort study

Define cohort

- Both exposed & not exposed groups have equal chance to:
  - Develop disease
  - Be followed-up

- Types:
  - Representative - low exposed subjects
  - Enriched - high exposed subjects
  - Specific group - occupational, institution etc
Measurements

- **Exposure**
  - Carefully defined in advance
  - Standard measurement for both E+ & E- groups

- **Outcome**
  - Primary vs. Secondary outcome

Follow-up

- Keep participation at >90%
- Equal likelihood to detect disease in all subjects
- Active vs. Passive follow-up
- Blinding
Example

Obesity as an Independent Risk Factor for Cardiovascular Disease: A 26-year Follow-up of Participants in the Framingham Heart Study (Hubert 1983)

Cohort study

**Advantages**
- Infer causality
- Measure multiple outcomes
- Study rare exposure
- Measure incidence

**Disadvantages**
- Costly
- Loss to follow up
- Large sample size for rare outcomes
- Selection bias

FIGURE 10. The relative odds of developing cardiovascular disease corresponding to degrees of change in Metropolitan Relative Weight between age 25 years and entry into the Framingham Study. The odds ratios reflect adjustments for the effects of relative weight at age 25 years and age and risk factor levels at exam 1.
### TABLE 2.3: CHOICE OF STRATEGY

<table>
<thead>
<tr>
<th>Basis</th>
<th>Cohort</th>
<th>Case-control</th>
<th>Cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw condition</td>
<td>Not practical</td>
<td>Bias</td>
<td>Not appropriate</td>
</tr>
<tr>
<td>To determine a precise risk</td>
<td>Best</td>
<td>Only estimate possible</td>
<td>Gives relative prevalence, not incidence</td>
</tr>
<tr>
<td>To determine whether exposure preceded disease</td>
<td>Best</td>
<td>Not appropriate</td>
<td>Not appropriate</td>
</tr>
<tr>
<td>For administrative purposes</td>
<td>Not appropriate</td>
<td>Not appropriate</td>
<td>Best</td>
</tr>
<tr>
<td>If attrition is a serious problem</td>
<td>Not appropriate</td>
<td>Attrition is usually minimal</td>
<td>Attrition may have occurred before the study</td>
</tr>
<tr>
<td>If selective survival is problem</td>
<td>Best</td>
<td>Not appropriate</td>
<td>Not appropriate</td>
</tr>
<tr>
<td>If all factors are not known</td>
<td>Best</td>
<td>Not appropriate</td>
<td>Less appropriate</td>
</tr>
<tr>
<td>Time and money</td>
<td>Most expensive</td>
<td>Least expensive</td>
<td>In between</td>
</tr>
</tbody>
</table>

#### Part 4 - Sample & sampling method
Sampling & sample size

- Type of sampling – Random vs. Non random
- Sample size – Based on objective & research design

Before we sample
Determine study place, duration & subjects

- Describe study place – especially if plan to represent a population
- State time & duration
- Who or what are the subjects – population, people, animal etc.
Subjects

- Target population
- Study population
- Sampling frame
- Sampling unit
- Observation unit

Example – NHMS III 2006

- Target population: All Malaysian
- Study population: Household up to strata 6
- Sampling frame: List of Enumeration Block & Living Quarters
- Sampling unit: Enumeration Block & Living Quarters
- Observation unit: All household in the selected Living Quarters
Sampling method

Random sampling

The ideal method. Randomly sample 10 students from a class of 50 students.

Follow certain pattern, order. Only first sample is random.

Study population divided into strata. All strata selected. Portion of sample in each strata sampled.

Study population divided into clusters. Assume all clusters are the same. Not all clusters selected. Only some will be sampled.

Several sampling techniques applied at different stage.
Non random sampling

- Purposive: Non representative. Convenience but with certain purpose. E.g. diabetic patients on single insulin therapy.
- Quota: When the sampling stop after achieving certain size.
- Snowball: Sample by reference or recommendation.

Example – NHMS III 2006

- Target population: All Malaysian
- Study population: Household up to strata 6
- Strata: State & location (urban or rural)
- Clusters: Enumeration Block & Living Quarters
- Sampling frame: List of Enumeration Block & Living Quarters
- Sampling unit: Enumeration Block & Living Quarters
- Observation unit: All household in the selected Living Quarters
- Sample distribution: Proportionate to size
Sample size

- How many sample required
- An estimate
- Adequate size to
  1. Represent population
  2. Test hypothesis
- Size affects duration & budget of the research
- Do not sample more than required

Sample size

- Sample size depends on
  1. Objective of the study
  2. Study design (design effect)
  3. Sampling method
  4. Expected (& precision) effect size
  5. Variability of sample
  6. Non-response rate
- Even the expected outcomes/effect size are estimates
Sample size – formula

- No single formula for all
- Depending on purpose – single proportion, compare two proportions, compare 3 proportions etc.
- Use of software or calculators e.g. PS Power and Sample Size, PASS etc.
- Anticipate non response, drop-out, loss to follow up, death (esp. animal study)

Example – Single proportion

- \[ N = \frac{z^2 p(1-p)}{d^2} \], where
- \( N \) is the sample size,
- \( z = z \) value for intended confidence interval,
- \( p \) is the estimated proportion (in decimal) &
- \( d \) is the precision of \( p \) (deviation from \( p \)) (in decimal)
Example – Single proportion

- Calculate sample size if you wish to do a study estimating the prevalence of DM as 20% with 5% variation at 95% CI (z for 95% CI is 1.96) anticipating 20% non-response.

- Answer: 246, ~250 then add 20% ~300 samples required.

Group Work #2

1. What is the best study design? Give one reason.
2. Describe the sample planned for your study – target, study, sampling frame, sampling unit & observation unit.
3. Calculate the sample size required.
Part 5 – Data collection

Plan for data collection

- Identify all variables of interest – produce detail data dictionary - declare all definitions & the measure types
- Must check validity & reliability of research tools – including questionnaires
- Training – ensure similar ways of collecting data
- Key word here - STANDARD
Research instruments

- Must be both valid and reliable
- Valid - content, face, criterion, construct etc
- Reliable - repeatability
- Must declare all tools used
- Including the questionnaire

Questionnaire or Record Form

- Face-to-face vs. self-filled (online or paper-based) vs. postal vs. telephone
- Specific questions (fulfill the objectives)
- Open vs. closed ended
- Structured usually closed-ended - e.g. dichotomous, Likert scale, multiple choice.
- Response code - e.g. 1=Male, 2=Female
Data dictionary

- List of variables
- Definitions & working definitions
  e.g. When you say a subject is diabetes? Hypertensive?
- Cut-of-point e.g. How many age category
- Outcome vs. factors
- Type of measures - dependent or independent
- Coding e.g. 1=Male, 2=Female

Data dictionary - Example
# Suggested details to include

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td><strong>Name</strong></td>
<td>The name normally required in computer such as in database and statistical analysis. Name can be in one short word e.g. agecat for Age Category</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td><strong>Label</strong></td>
<td>The name that can appear in table, graph or report</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td><strong>Definition</strong></td>
<td>The definition used in the research. It is advisable to include references used</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td><strong>Instrument used</strong></td>
<td>When relevant, we can describe the instrument used which include the brand and the method of calibration if relevant</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td><strong>Level of measurement</strong></td>
<td>Should specify either it is nominal, ordinal or continuous</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td><strong>Category option and code</strong></td>
<td>If the variable is categorical, the options should be specified e.g. Gender; Male=1, Female=2</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td><strong>Unit of measurement</strong></td>
<td>If the variable is numerical, we should specify its unit e.g. mmol/L, mg/dL</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td><strong>Precision of measurement</strong></td>
<td>How precise the variable is measured e.g. Age is measured to the nearest 1 year old. Income is measured to the nearest RM100</td>
</tr>
<tr>
<td><strong>9</strong></td>
<td><strong>Data linkage</strong></td>
<td>If this variable is related to other variable, we can specify here e.g. Missing value (Question on pregnancy) if respondent is Male (Question on Gender)</td>
</tr>
</tbody>
</table>

---

Data dictionary is very important!!
Part 6 – Statistical analysis plan

Statistical analysis plan

- Based on objective, especially the specific objectives – orderly manner
- Descriptive vs. analytical analysis
- The product is dummy table
- State all statistical tests planned to be used
- State significant level
- State software used
- Engage a statistician from the beginning!
Dummy table - example

Objective
To compare blood glucose level between gender

Variables involved

<table>
<thead>
<tr>
<th>Variable label</th>
<th>Working definition (study data)</th>
<th>Status</th>
<th>Variable name</th>
<th>Level of measurement</th>
<th>Category label (if relevant)</th>
<th>Variable Unit</th>
<th>Precision of measurement</th>
<th>Missing value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood glucose</td>
<td>As measured</td>
<td>Dependent</td>
<td>glu</td>
<td>interval</td>
<td>1 = Male, 2 = Female</td>
<td>mmol/L</td>
<td>0.1</td>
<td>999</td>
</tr>
<tr>
<td>Gender</td>
<td>As reported</td>
<td>Independent</td>
<td>sex</td>
<td>Nominal</td>
<td></td>
<td></td>
<td></td>
<td>None</td>
</tr>
</tbody>
</table>

Statistical analysis
1. Check normality of glu
2. If glu Normal, run Independent sample t-test; if glu not Normal, run Mann Whitney U-Test
3. Significance level = 0.05

Dummy table

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Statistics</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>10.6 (4.2)</td>
<td>nan</td>
<td>11</td>
<td>0.01</td>
</tr>
<tr>
<td>Female</td>
<td>10.6 (4.2)</td>
<td>nan</td>
<td>11</td>
<td>0.01</td>
</tr>
</tbody>
</table>

SD = Standard deviation
Data quality

- Valid value
  e.g. age > 200 years, weight > 500 kg, pregnant male etc
- No missing value
- Relevant skip response
  e.g. Not Applicable response for number of pregnancy for male respondent
- Declare method to ensure good data quality – e.g. double data entry

Group Work #3

- Prepare the data dictionary
- Describe what you will do with the data
- Prepare dummy table for each specific objective
Part 7 – Writing report

Before you write

▶ Decide the target audience
▶ Scientific publication or report
▶ Choose journal
▶ Study the format & requirement
▶ Separate text, table & graphics
The structure

Publication
- Title
- Abstract
- Keywords
- Introduction
- Method
- Results
- Discussion
- References

Report/thesis*
- Title
- Abstract
- Introduction
- Literature review
- Objective
- Methodology
- Results
- Discussion
- References

* Institution specific

The suggested sequence

1. Based on specific objective, analyse the data & produce planned tables
2. Interpret & describe the results in Result section
3. Discuss in Discussion section
4. Answer the research questions
5. Complete the method & introduction
6. Finally, write the abstract
Writing result

- Describe your result (no discussion)
- No reference (usually)
- Text vs. table vs. graphic (no redundancy)
- Text to summarise, Table for detail, Graphic to show trend
- May state relevant statistics done (if not mentioned in method)

Writing discussion

- Should answer the research questions mentioned in Introduction
- Discuss the result
- Do not repeat text as in Result
- May state limitation (but don’t go overboard)
- Recommend
- Conclude
Part 8 – The administration

Other things to plan

1. Ethical consideration – consent form, advisory committee
2. Budget
3. Approval
In summary, what are the critical information

1. Specific objectives
2. Conceptual framework
3. Data dictionary
4. Dummy table & analytics guidelines