Research methodology for nurses

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

Contents

- Why we do research?
- How we do research?
- Effective literature search & management
- Choosing best study design
- Planning for data collection
- Planning for statistical test
- How to write for publication
The programme

Day 1
- Getting the research idea
- Literature review & managing them (using EndNote)
- Building conceptual framework
- Phrasing objective
- Determine best study design
- Sampling plan & sample size
- Plan for data collection – preparing data dictionary

The programme....

Day 2
- Plan for statistical analysis – preparing dummy table
- Basic statistical analysis
- Descriptive statistics – determine Normality, mean (SD), median (IQR), count (%)
- Analytical statistics – compare means, compare proportion, correlation
Why we do research?

- Answers our curiosity?
- Forced by our superior
- For promotion

Start with **end** in mind

*(Bijaksana)*
The 4 Critical Strategies

1. Clear conceptual **framework**
2. SMART **objectives**
3. Detail **data dictionary**
4. Expected **dummy table**
Organise your research idea

Jamalludin Ab Rahman MD MPH

Identify variables involved

- Main outcomes (dependant)
- Explanatory (exposures, factors) variables (independent)
- Confounding variables
Recent studies have suggested that the use of alcohol-containing mouthwashes may increase the risk of oropharyngeal cancer. Heavy alcohol intake and tobacco use are established causes of oropharyngeal cancer. Their use is associated with mouthwash use. In addition, alcohol and tobacco use both tend to be underreported. Here, the authors show that, under the hypothesis that mouthwash does not increase the risk of oropharyngeal cancer, confounding due to underascertainment exposure to alcohol and tobacco would result in a spuriously elevated odds ratio for mouthwash use. As a general principle, a null association becomes apparently positive if a confounding variable is incompletely ascertainment. A spurious association may be produced even in the absence of a difference in the extent of the underascertainment of the confounder among the comparison groups. Am J Epidemiol 1990;134:1001-5.

confounding factors (epidemiology); oropharyngeal neoplasms
Alcohol-based mouth wash

Smoking

Oral Cancer

Behavioiral risk factors
- Tobacco smoking
- Alcohol misuse
- Poor nutrition
- Physical inactivity

Biomedical risk factors
- High blood pressure
- High blood cholesterol
- Excess weight

Other factors
- Socio-environmental determinants
- Psychosocial factors
- Early life factors

Chronic diseases
Figure 4.4: Revised problem analysis diagram of factors contributing to high defaulter rate among TB patients

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.

Literature Review and Bibliographic Management

Jamalludin Ab Rahman MD MPH
Research add more information to the circle of knowledge.

Figure 1.1 The generation and communication of research knowledge and information.
Literature review is not

- A chronological catalogue of all material,
- A collection of quotes or paraphrases, or
- A sales pitch selling only the good side of a study.
Literature review is

- Organised
- Structured
- Evaluated (critically appraised)

research materials

The steps – 5S

1. Strategise (plan) \(^{\leftarrow}\) Most important step
   - Get your topic, problem statement & conceptual framework ready (& in-sync)
   - Identify components, variables, keywords & authors
2. Search – by keywords, domains of interest
3. Screen – title & abstract
4. Sort – organise & prioritise materials
5. Summarise – evaluate & re-organise
Strategise
- Based on problem statement & conceptual framework
- Identify causal relationship, variables, domains, authors

Search
- Online – Google, Google scholar, individual journal’s page, unpublished materials
- Use electronic bibliographic apps e.g. EndNote, Mendeley, Papers

Screen
- Screen the title, abstract

Sort
- Organise into the pre-planned structure or domain of interest
- Re-organise the information if needed

Summarise
- Credibility of articles (journal & authors)
- Identify study population, study design, measurement method used
- Check with the problem statement
- Annotate (e.g. by using EndNote)

Title: Relationship of OSA & HPT
Design: Case-control (HPT vs. No-HPT)
The number of articles found

Google

Altenative spellings

Obstructive sleep apnea

Time period

Since 2016
Since 2014
Since 2011

Number of citations by Google

6 September 2016
Research methodology for nurses

The same doc maybe available at multiple places

You can cite (import) into EndNote

No full text

Full text available
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. S. M. Gami, A. A.</td>
<td>Obstructive sleep apnea and other risk factors</td>
<td>The Journal of Sleep Disorders, 1993</td>
<td>1993</td>
</tr>
<tr>
<td>A. A. and M. S. Gami</td>
<td>Obstructive sleep apnea and other risk factors</td>
<td>The Journal of Sleep Disorders, 1993</td>
<td>1993</td>
</tr>
</tbody>
</table>

**Customise the header**

Include full text or any relevant documents

**Domains of interest: Can assign many domains for one article**

06/09/2016
Phrasing objectives

Jamalludin Ab Rahman  MD, MPH
SMART

- Specific – What you really want to do
- Measurable – Must be able to be measured
- Attainable – Achievable
- Relevant – Relevant to you
- Timely – Can be done within the time you have

**e.g.**

- To determine the prevalence of breastfeeding among mothers

Can rephrase to:

- To measure the prevalence of exclusive breastfeeding among working mothers attending post-natal clinic in Pahang
- To measure the prevalence of exclusive breastfeeding and factors related with it among working mothers attending post-natal clinic in Pahang
Choosing Best Research Designs

Jamalludin Ab Rahman MD MPH

Research design

Design

Observational

Cross sectional

Case-control

Cohort

Experimental

Animal Trial

Clinical Trial

Community Trial

Measure exposure & outcome & the same time

Fix the outcome. Measure the exposure

Fix the exposure. Measure the outcome
Cause & Effect

- Cause precedes effect
- Exposure:
  - Smoking
  - High fat diet
  - New drug
- Outcome:
  - Lung cancer
  - Obesity
  - Better prognosis

Cross sectional study

- Observe both Cause & Effect at the same time – No temporal association
- Exposure:
  - Smoking
  - High fat diet
  - New drug
- Outcome:
  - Lung cancer
  - Obesity
  - Better prognosis
Case-Control study

<table>
<thead>
<tr>
<th>Cause</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>Outcome</td>
</tr>
<tr>
<td>Smoking</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>High fat diet</td>
<td>Obesity</td>
</tr>
<tr>
<td>New drug</td>
<td>Better prognosis</td>
</tr>
</tbody>
</table>

Observe the Cause (PAST events)

Cohort study

<table>
<thead>
<tr>
<th>Cause</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>Outcome</td>
</tr>
<tr>
<td>Smoking</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>High fat diet</td>
<td>Obesity</td>
</tr>
<tr>
<td>New drug</td>
<td>Better prognosis</td>
</tr>
</tbody>
</table>

Observe the effect (NEW event)
Hierarchy of Evidence

- Meta-analysis
- Systematic review
- Randomised controlled trial
- Cohort
- Case control
- Case report
- Expert opinion

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

---

![Diagram](image_url)

Case control study

- Smoking Exposure → Lung cancer patients (Case)
- Non smoking No exposure
- Smoking Exposure
- Non smoking No exposure

History Now
### Lung Cancer vs. No Lung Cancer

<table>
<thead>
<tr>
<th></th>
<th>Lung Cancer</th>
<th>No Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>20 (18.2%)</td>
<td>90 (81.8%)</td>
</tr>
<tr>
<td>Not Smoking</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
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)

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.
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

<table>
<thead>
<tr>
<th>Basis</th>
<th>Cohort</th>
<th>Case-control</th>
<th>Cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare 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>
Sampling Plan
(Sampling technique & sample size)
Jamalludin Ab Rahman MD MPH
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

Random
Represent a population
Must have sampling frame

- Simple
- Systematic
- Stratified
- Cluster

Non-Random
Still valid if justified

- Convenience
- Purposive
- Snowball
- Volunteer
Simple

Systematic
Cluster

Stratified
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


Non representative. Convenience but with certain purpose. E.g. diabetic patients on single insulin therapy.

When the sampling stops after achieving certain size.

Sample by reference or recommendation.
Example – NHMS III 2006

Two stage stratified random sampling

- 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

Example – A clinical trial

- Target population: All diabetic patients
- Study population: Diabetic for at least 1 year on insulin therapy who attended MOPD from Jan-Dec 2014
- Sampling frame: N/A
- Sampling unit: Any diabetic who fulfilled the selection criteria
- Observation unit: Same as SU
Why we calculate sample size?

- Estimate sample required for external validity (inference) & logistic preparation
- Justify sample size for research to:
  - represent population
  - measure treatment effect (detect difference)
- We don’t do research haphazardly

Sample size is an estimate

- Best sample size for specific purpose *(for the main objective)*
- Never an exact value, always an estimate
- Calculated for certain expected estimates (outcomes) at certain degree of precision
- Expected values - estimated from previous studies or from intelligent ‘guess’
Plan with the end in mind

- No one formula fits all
- Depending on objective:
  1. Represent population
  2. Hypothesis testing
- Adjust for design effect (based on type of sampling), confidence level, alpha error, power, stratification and anticipated response rate

Pre-requisites

- Objective determined
- Sampling method known
- Estimate the outcome
- Precision required
- Statistical test used is known
- Set the power ($\beta$) and confidence level ($\alpha$)
- Anticipate non-response
Represent population

- Represent population e.g. state, district, village, institution etc
- Usually for prevalence study e.g. to measure prevalence of hypertension in Malaysia; or mean DMF among adult in Pahang

Single proportion

- \[ n = \frac{z_{\alpha/2}^2 p(1-p)}{d^2} \]
- Where \( z = \) value from standard normal distribution corresponding to desired confidence level, \( \alpha \) \( (z_{\alpha/2}=1.96 \) for 95%CI)
- \( p \) is expected true proportion
- \( d \) is desired precision
- For small populations \( n \) can be adjusted so that \( n_c = \frac{Nn}{N+n} \)
Example

- Study to measure prevalence of hypertension in a village of 1000 people
- Expected prevalence = 40%, Precision = 5%, CI=95%, expected non-response 20%
- \[ n = \frac{1.96^2 \times 0.4(1-0.4)}{0.05^2} = 369 \approx 370 \]
- Anticipating non response, \( n = 450 \)
Single mean

- Study to measure one sample mean
- \[ n = \left( \frac{Z_{\alpha/2} \cdot s}{d} \right)^2 \], where \( s \) = expected standard deviation (SD)
- We use smallest \( d \) to get largest \( n \) possible

Example

- Study to measure average DMF among adults in a village of 2000 people
- Estimated DMF = 11 (SD 10) with the precision of 2 at 95%CI
- \[ n = \left( \frac{1.96 \cdot 10}{2} \right)^2 = 96.04 \approx 100 \]
Hypothesis testing

- How one variable is different from the other
- E.g. different % of hypertension between male & female

Compare two proportions

\[ n = \frac{(z_\alpha + z_\beta)^2 \times (p_1(1-p_1) + p_2(1-p_2))}{(p_1-p_2)^2} \]

- Total sample size = 2*n
- Where \( p_1 \) and \( p_2 \) are the expected sample proportions of the two groups; \( z_{\alpha/2} \) is the critical value of the Normal distribution at \( \alpha/2 \) and \( z_\beta \) is the critical value of the Normal distribution at \( \beta \) (e.g. for a power of 80%, \( \beta \) is 0.2 and the critical value is 0.84)
Example

- A study to compare prevalence of diabetes mellitus between male and female. Expected prevalence are 30% vs. 40% respectively. Power 80% at 95% confidence level.

\[ n = (1.96 + 0.84)^2 \times \left( \frac{0.3(1-0.3)+0.4(1-0.4)}{(0.3-0.4)^2} \right) = 7.9 \times 45 = 355.5 \cong 360 \]
- Total sample size = 360 \times 2 = 720

Compare two means

\[ n = \left( \frac{2(z_{\alpha/2}+z_{\beta})^2 s^2}{d^2} \right) \]
- Where \( d \) = smallest means difference and \( s \) = standard deviation
- Total sample size = \( 2^n \)
Example

- A study to compare means of HbA1c between diabetic treated with new drug versus the standard drug (control).
  Expected difference of 10.5% vs. 11.5% with SD estimated of 5%
- \[ n = \frac{2 \times 5^2}{1^2} \times (1.96 + 0.84)^2 = 395 \]
- Total sample size = 790

Using applications

- Software or online calculators
- Make sure the formula is valid
- Suggestion:
  - PS: Power and Sample Size Calculation
    [http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize](http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize)
  - Epi Info [http://www.cdc.gov/epiinfo/7/](http://www.cdc.gov/epiinfo/7/)
  - OpenEpi
Epi Info – Population survey

Population survey or description study
For simple random sampling, leave design effect and clusters equal to 1.

<table>
<thead>
<tr>
<th>Population size</th>
<th>Confidence level</th>
<th>Cluster size</th>
<th>Total sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>95%</td>
<td>136</td>
<td>272</td>
</tr>
</tbody>
</table>

Epi Info – Cohort & Cross-sectional

Unmatched Cohort and Cross-Sectional Studies (Exposed and Unexposed)

<table>
<thead>
<tr>
<th>Two-sided confidence level</th>
<th>Power</th>
<th>Ratio (Exposed : Unexposed)</th>
<th>% outcome in exposed group</th>
<th>Odds ratio</th>
<th>% outcome in exposed group</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>80%</td>
<td>1</td>
<td>30%</td>
<td>1.33334</td>
<td>40%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>396</td>
<td>356</td>
<td>752</td>
<td></td>
</tr>
<tr>
<td>Unexposed</td>
<td>396</td>
<td>356</td>
<td>752</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>792</td>
<td>712</td>
<td>1504</td>
<td></td>
</tr>
</tbody>
</table>
Open Epi – Compare two means

Sample Size For Comparing Two Means

Input Data
- Confidence Interval (2-sided): 95%
- Power: 80%
- Ratio of sample size (Group 2/Group 1): 1

<table>
<thead>
<tr>
<th>Mean</th>
<th>Group 1</th>
<th>Group 2 Difference*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Variance</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Sample size of Group 1: 393
Sample size of Group 2: 393
Total sample size: 786

*Difference between the means

Results from OpenEpi, Version 3, open source calculator—SSMean
Print from the browser with ctrl-P
or select text to copy and paste to other programs.

PS - Dichotomous

Power and Sample Size Program: Main Window

Output
- What does one wish to test?
- Confidence Level: 95%

Input
- p1: 0.4
- p2: 0.4
- Power: 80%
- Calculate

Description
- We are planning a study of independent cases and controls with 1 control per case.
- Prior data indicate that the probability of exposure among controls is 0.4. Either the disease rate among exposed cases is 0.4 or the disease rate among controls is 0.5.002 is the criterion for rejection of the null hypothesis that the exposure is not related to the disease.
- p1 and p2 are equal with probability 0.5. The Type I error should be associated with this test of null hypothesis in 0.05. We will not use a continuity correction or squared statistic in Fisher’s exact test to evaluate this null hypothesis.
PS – Compare two means

Planning for Data Collection

Jamalludin Ab Rahman  MD MPH
You wanna collect now?

1. Objectives are clear
2. Variables identified & defined
3. Samples identified
4. Instrument valid & reliable
5. Instrument tested
6. Simulation done

Only then, you go and collect the data
Variables identified & defined

- Conceptual framework
- Identify variables
- Define variables

How to define variable?

- Definition
- Working definition
- How it will be measured
- Precision of measurement
- How it will be reported
Overweight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. (WHO 1998).

A crude population measure of obesity is the body mass index (BMI), a person's weight (in kilograms) divided by the square of his or her height (in metres). In this study, overweight and obesity are grouped together and labelled as Obesity. Any BMI 25 kg/m² or above is considered obese (WHO 1998).
Furnish the info for each variable e.g. Obesity

- Definition
- Working definition
- How it will be measured
- Precision of measurement

Weight is measured using calibrated Seca 762 with precision to the nearest of 1 kg
Height using Seca 206 with precision to the nearest 1 mm.

Make sure the instrument are valid & reliable

Fig. 1. Assessment scales: visual analogue scale (VAS), numerical rating scale (NRS) and verbal rating scale (VRS).
Validity and reliable enjoyment, social s

Yan Liang, Patrick W.C. *
* Department of Physical Education, Hong Kong
* National Institute for Health Innovation.
* LGI/UNR (Children's Nutrition Research

NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY III. Body Measurements (Anthropometry) 1998

Exhibit 3-1. SP position for standing height

LOOKING STRAIGHT AHEAD

SHOULDER RELAXED

ARMS AT SIDES

LEGS STRAIGHT AND KNEES TOGETHER

FEET PLAT TOGETHER
Furnish the info for each variable e.g. Obesity

- Definition
- Working definition
- How it will be measured
- Precision of measurement
- How it will be reported

Obesity will be described in count (N) and proportion (%).
Leading article

Defining datasets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology-driven approach

Simon de Lusignan MSc MD(Res) FHEA FRCS CTIP FRCP
Professor of Primary Care and Clinical Informatics, Department of Health Care Management and Policy, University of Surrey, Guildford, UK

Siew-Teng Liew PhD FRACGP FACHI
Professor of General Practice, University of New South Wales, Director, General Practice Unit, SW Sydney Local Health District, Sydney, Australia

Georgios Michalakakis MSc
Doctoral Student, Computing Department

Simon Jones MSc PhD
Research Professor, Department of Health Care Management and Policy
University of Surrey, Guildford, UK

The Data Dictionary Formats:

General Guidelines:

1. Please list or provide in a table format the following information:
   - File the data is coming from (e.g. NHIS 2000 person file)
   - Variable Names
   - Variable Descriptions

2. Please use the examples below to help you construct your data dictionary.

3. Highlight the variables in each dictionary that will be used in the merge. It is important that these variables be formatted consistently between data sets for the merge to go smoothly and most cost efficiently.
Then only you start to design the data collection form
Tools to create form

- Usual word processing (e.g. Word), spreadsheet (e.g. Excel)
- Better with Database (Microsoft Access)
- Google Form
- EpilInfo Make Questionnaire
Figure 1. How to Improve Header and Footer Designs of a CRF

Figure 4. Organizing CRF Data Fields

<table>
<thead>
<tr>
<th>Data Field Types</th>
<th>CRF Data Field Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CATEGORICAL</strong></td>
<td></td>
</tr>
<tr>
<td>• Single-Response</td>
<td>○ Yes ○ No</td>
</tr>
<tr>
<td>• Multiple-Response</td>
<td>□ Key(s) □ Button(s)</td>
</tr>
<tr>
<td>• Open-ended Multiple Response</td>
<td>□ Key(s) □ Button(s); If Button(s) is used, please explain: ___________</td>
</tr>
<tr>
<td>• Coded Response</td>
<td>○ 1 Mild 2 Moderate 3 Severe</td>
</tr>
<tr>
<td></td>
<td>(____) 1=Mild, 2= Moderate, 3=Severe</td>
</tr>
<tr>
<td><strong>TEXT</strong></td>
<td>Site (e.g. left shoulder): ______________________</td>
</tr>
<tr>
<td><strong>DATE</strong></td>
<td></td>
</tr>
<tr>
<td>• Standard Format</td>
<td>dd/mm/yyyy</td>
</tr>
<tr>
<td></td>
<td>DD MMM YY YY</td>
</tr>
<tr>
<td><strong>NUMERIC</strong></td>
<td>Number of Key(s) being used (____) (e.g. 0, 1, 2... etc)</td>
</tr>
</tbody>
</table>
Plan for Statistical Analysis

Jamalludin Ab Rahman MD MPH
Research idea

General objectives

Specific Objective

Analysis

Specific Objective

Analysis

Specific Objective

Analysis

Type of statistical analysis

Descriptive

Univariable

IV

DV

e.g. Describe socio-demographic characteristics - Age, Sex, Race etc.

e.g. Prevalence of hypertension.

Analytical

Bivariable

IV

DV

e.g. Compare demographic characteristics between two population – Compare age between male & female

e.g. Distribution of gender by hypertension status

Multivariable

IV

DV

IV

e.g. How demographic characteristics (more than one factors) explain hypertension
Steps for the plan

1. Focus on each specific objective
2. Decide whether it will be descriptive or analytical
3. For descriptive statistics: Mean (SD) or Median (IQR) or N(%) 
4. For analytical (present of hypothesis); state the Ho 
5. Select suitable statistical test – usually run Univariate first, then Multivariate
6. Present this in a dummy table

Dummy table - example

Objective
To compare blood glucose level between gender

Variables involved

<table>
<thead>
<tr>
<th>Variable label</th>
<th>Working definition (baseline 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>g</td>
<td>Interval</td>
<td></td>
<td></td>
<td>mmol/L</td>
<td>0.1</td>
</tr>
<tr>
<td>Gender</td>
<td>As reported</td>
<td>Independent</td>
<td>sex</td>
<td>Nominal</td>
<td>1 = Male, 2 = Female</td>
<td></td>
<td></td>
<td>None</td>
</tr>
</tbody>
</table>

Statistical analysis
1. Check normality of g
2. If g is Normal, run Independent sample t-test; if g is 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>mmol (int)</td>
<td>0.00</td>
<td>aa</td>
<td>0.00</td>
</tr>
<tr>
<td>Female</td>
<td>mmol (int)</td>
<td>0.00</td>
<td>aa</td>
<td>0.00</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

How to report the results

- Start with descriptive statistics – usually for baseline findings e.g. description of study population – socio-demographic or basic clinical characteristics
- Proceed with hypothesis testing (if any) – e.g. compare between groups/treatments
- State statistical tests used
- State data transformation done (if any)
- Declare missing value & how they were managed
Effects of Renal Sympathetic Denervation on Blood Pressure, Sleep Apnea Course, and Glycemic Control in Patients With Resistant Hypertension and Sleep Apnea

Statistical Analysis

The results thorough are presented as median and interquartile range. For comparison between measures before and after denervation, Wilcoxon signed-rank test was used. Comparison of the prevalence rates among groups was performed using the $\chi^2$ test. The degree of association between variables was assessed using Pearson correlation coefficients ($r$). $P<0.05$ was considered statistically significant. Data analysis was carried out using statistical software PASW Statistics 18 (SPSS Inc, Chicago, IL).

…

Statistical Analyses

After preliminary univariate and bivariate analyses, we used Cox proportional hazards models to examine the effect of sleep duration on the risk of being diagnosed with hypertension over the 8- to 10-year follow-up period. The time duration to diagnosis was determined from the baseline date to the first report of hypertension. Covariates in the first adjusted multivariate model (model 2) included daytime sleepiness (never/ rarely/ sometimes or often/ almost always), depression (yes or no), physical activity (6=high, 5, 4, 3, 2=low), alcohol consumption (0, >0 and <28, or >28 g/day), salt per day (continuous), current smoking status (0, 1 to 5, 6 to 10, 11 to 20, or >20 cigarettes per day), pulse rate, and gender. We chose to include the demographic variables of education (high school graduate or >high school graduate), age (5-year interval), and ethnicity (white or nonwhite, including black, Hispanic, Asian, and other) in a separate model (model 3), because we hypothesized that a significant proportion of their ability to influence the results would be based on the fact that they are associated with being overweight and obese. We included BMI (lean=<$25$, overweight=$>25$ and <$30$, or obesity=$>30$ and history of diabetes (yes or no) in the final model (model 4) to test whether these variables acted as partial mediators of the relationship between sleep duration and the incidence of hypertension. The significance of individual coefficients in the Cox proportional hazards models were determined by the 95% confidence limits for hazards ratios (HRs).

We tested for multiplicative interaction between age and duration of sleep with the log likelihood ratio test and found that age acted as an effect modifier between the number of hours of sleep per night and the incidence of hypertension ($P<0.05$). We then stratified the sample by 10-year age increments and performed Cox proportional hazards analyses with each. The HRs for the incidence of hypertension for sleeping $>7$ hours per night compared with the other sleep categories were similar for subjects who were in their 50s, 40s, and 50s at the time of the 1982–1984 follow-up. We chose to divide the sample into 2 age groups with subjects who, at the time of the 1982–1984 study, were between the ages of 32 and 59 years in one group and subjects who were between the ages of 60 and 86 years in another group. After excluding subjects who were deceased, who did not answer the sleep duration question, who had missing data on any of the covariates, and who had hypertension at or before the 1982–1984 survey, there were 3620 subjects between the ages of 32 and 59 years and 1190 subjects between the ages of 60 and 86 years for the analyses.

The NHANES I included weights to account for the complex sampling design and to allow approximations of the US population. We conducted nonweighted analyses using SAS software for 3 reasons. First, our objective was not to provide national estimates but to look at the relationship between sleep duration and the incidence of hypertension. Second, the baseline measures of our study were taken from the 1982–1984 follow-up to the NHANES I, so the weights created for baseline measures taken from the 1971–1974 NHANES I did not account for subjects who were lost to follow-up between the 2 waves. Third, there have been differences of opinion regarding the appropriateness of using the sample weights with the NHANES.25

was consistent with the hypothesis that these therefore, be a significant risk factor for hypertension, essential sleep
**Type of data**
*(Level of measurement)*

- **Categorical**
  - Nominal: e.g. Gender, Race
  - Ordinal: e.g. Cancer staging, Severity of CXR for PTB
- **Numerical**
  - Discrete: e.g. Parity, Gravida
  - Continuous: e.g. Hb, RBS, cholesterol.

**How to describe a data**

- **Data**
  - Categorical
    - Frequency (count) & Percentage
  - Numerical
    - Normal: Mean (SD)
    - Not Normal: Median (Range/IQR)
Bivariable analysis

- Compare numericals – compare means, compare ranks, correlation
- Compare proportions – chi-square

<table>
<thead>
<tr>
<th>Dependent (outcome) variable</th>
<th>Numerical</th>
<th>Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>Correlation</td>
<td>Independent sample t-test</td>
</tr>
<tr>
<td>Independent variable (factor)</td>
<td>One-way ANOVA</td>
<td>Chi-square</td>
</tr>
</tbody>
</table>
### Bivariable analysis - detail

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>Categorical</td>
<td>Chi-square</td>
</tr>
<tr>
<td>Categorical (2 pop)</td>
<td>Numerical (Normal)</td>
<td>Independent sample t-test</td>
</tr>
<tr>
<td>Categorical (2 pop)</td>
<td>Numerical (Not Normal)</td>
<td>Mann-Whitney U test</td>
</tr>
<tr>
<td>Categorical (&gt; 2 pop)</td>
<td>Numerical (Normal)</td>
<td>One-way ANOVA</td>
</tr>
<tr>
<td>Categorical (&gt; 2 pop)</td>
<td>Numerical (Not Normal)</td>
<td>Kruskal-Wallis test</td>
</tr>
<tr>
<td>Numerical (Normal)</td>
<td>Numerical (Not Normal)</td>
<td>Pearson Correlation Coefficient Test</td>
</tr>
<tr>
<td>Numerical (Normal/ Not Normal)</td>
<td>Numerical (Not Normal)</td>
<td>Spearman Correlation Coefficient Test</td>
</tr>
<tr>
<td>Numerical (Normal)</td>
<td>Numerical (Normal) – Paired</td>
<td>Paired t-test</td>
</tr>
<tr>
<td>Numerical (Not Normal)</td>
<td>Numerical (Not Normal) – Paired</td>
<td>Wilcoxon Signed Rank Test</td>
</tr>
</tbody>
</table>

### Multivariate test

<table>
<thead>
<tr>
<th>Dependant</th>
<th>Independent variables</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>All numerical</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Numerical</td>
<td>Numerical &amp; categorical</td>
<td>GLM</td>
</tr>
<tr>
<td>Repeated numericals</td>
<td>Numerical &amp; categorical</td>
<td>Repeated measures ANOVA</td>
</tr>
<tr>
<td>Categorical</td>
<td>Numerical &amp; categorical</td>
<td>Logistic regression</td>
</tr>
</tbody>
</table>
Specific test

- Survival analysis – time to event analysis
- Time series – observe trend over time & forecasting
- Factor analysis – data reduction
- Structural equation modelling – general tool to explore or confirm a ‘causal’ model (relationship between variables)

How to publish your research

Jamalludin Ab Rahman MD MPH
**How to write a research paper**

1. **Title**
   - State objectives
   - The backbone of the study

2. **Abstract**
   - How the study was done
   - Present based on objectives

3. **Introduction**
   - Finalise title
   - Justify research. State research gap. Phrase objective clearly

4. **Method**
   - Answer objectives. Discuss differences, recommendation & limitation

5. **Discussion & conclusion**
   - Present based on objectives

6. **Results**
   - Discuss differences, recommendation & limitation

**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
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