Introduction to biostatistics: Basics and study design

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

17th August 2558

A little bit about me....

- Nearly 20 years experience as biostatistician having consulted in thousands of health and medical research projects (Pre-clinical->clinical->population-based)
- Have about 60 publications across a wide range of health-related disciplines
- Hold a joint appointment with Faculties of Public Health (Teaching appointment) and Medicine (Clinical biostatistician), KKU
- Main role (PH) is in the supervision of PhD students and (Med) consulting regarding clinical research
- Also teaching into courses (like this) in research methods and biostatistics
- Come from Queensland in North-eastern Australia (Australia's version of Isaan)

Conventions

Before we start, I will just point out a few conventions I will use:

Note:.....

Things to note given in a green box

Pitfalls:.....

Common mistakes and things to watch out for given in a red box

R SYNTAX:....

Important R syntax will be in purple boxes and be in `courier` font. This will help you find it easily when you have to refer back to these notes.

What we cover today(This session)

1. Introduction
   - Intro to study design
   - Intro to biostatistics
2. Study objectives and analytical approach
3. Our case study data: DMHT

Something to think about...

The US army draft during a certain war considered potential recruits doing a 100 question multiple choice exam.

- Those scoring ≥ 90 went into officer training
- Those scoring between $6 - 89$ went into basic training
- **Those scoring ≤ 5 went into officer training**

WHY????

Epidemiology Vs Biostatistics

- The disciplines of Epidemiology and Biostatistics are inextricably related.
- They both relate to the 'research' end of the health discipline
- ...as such, a basis in epi and biostats is essential for health and (bio)medical researchers

Definition:

Epidemiology is the study of the distribution and determinants of disease, health, or injury outcomes in human populations and the use of the knowledge we gain (from this study) to control the health problems

A bit of Epi101: Study design

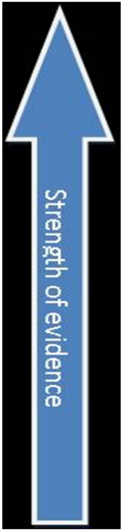
We will start by examining some of the (classical) health and medical study designs

- The main factors that govern our choice of study design are the: **1. research question**, **2. target population** (and ability to sample it) and (correspondingly) **3. our ability to control sources of bias**
- Fortunately (for you) most pre-clinical studies (especially those involving a 'treatment') tend to be up the stronger end of health study designs (more on this later)

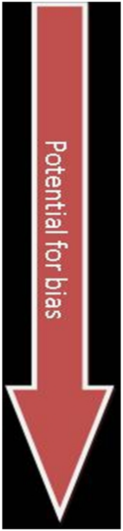
Study designs in pre-clinical studies

Most pre-clinical studies are conducted in the labaratory, if the study is experimental in nature (manipulation by the researcher), this results in studies with a high **strength of evidence**

Health and medical study designs



Study design
Randomized Controlled Trial
Quasi-experiment
Cohort
Cross-sectional
Case-Control
Ecological
Case series
Case study



Randomized controlled trials

- Widely considered to be the gold standard of study designs
- Provide the strongest evidence
- Double blinding and randomization attempt to minimize selection and confounding bias
- As a true experiment, RCTs always involve an intervention (or treatment)
- That is, experimental manipulation by the researcher
- Often not possible in (population-based) epidemiological studies, especially those involving the study of risk factors (unethical to impose a protocol of risky behaviour)

Quasi-experiments

- Many study don't (or can't) include a randomization process, but still involve an intervention
- For example, it may be unethical or impractical to randomize patients, or it might be impossible to avoid contamination between different experimental arms
- Classical example of a quasi experiment is (some) pre-post studies
- Example might be 2 treatments currently is use at a hopsital
  - Patients may be allocated teatment based on treatment availibility or physician preference
  - Here patients are not RANDOMIZED to treatment
  - What problems might this cause???

Cohort studies

- Cohort studies are considered the strongest (in terms of evidence) of the observational studies.
- **Observational designs**: Researcher has no control over group membership (i.e. non-experimental)
- Their main strength (above something like a cross-sectional study) is that the risk factors (exposures) are collected **before** the outcome (e.g. disease)
- cohort studies can be **retrospective**: data already exists and NOT collected for the purposes of the study (i.e. a secondary analysis or analysis of routine collected data); or
- **proscpective**: data collected specifically for answering the research question (i.e. the study generates the data collection)
- Also, cohort studies can involve a single collection of the endpoint (e.g. disease) ; or
- they can be **fully longitudinal**: where patients are repeatedly measure over time

Longitudinal cohort studies

- Generally the strongest of all 'observational' designs
- Repeated observation of participants over time
- Advantage is evidence of causality (exposure ⇒ Cohort study),
- Common problem is loss-to-follow-up(LTFU)
- LTFU especially a problem when it confounds with effects/outcomes of interest: LTFU does not occur at random
- Like other cohort studies longitudinal cohorts can be collected **retrospectively** or **prospectively** (data collected for study)

Reality check:

Some reviewers can be overly critical of retrospective studies. Their argument is that data were NOT collected to answer the question, so why should we believe is (e.g sample size/power is an issue). The logic is that a well-planned prospective study can be trusted. BUT in reality most 'planned' cohort studies are badly planned so it makes little difference (these issues can be adequately addressed in methods and discussions)

Cross-sectional studies

- Often called 'prevalence' studies as we can get an estimate of disease prevalence from our sample (or subgroups of the sample)
    - In contrast, see case-control studies below
- One of the most common study designs in population-based epidemiology as they are generally cheaper and more practical than longitudinal studies
- Main problem is that without strong contextual evidence, associations can not be assumed to be causal (only associative)
- Also rely strongly on a 'representative' sample of the target population (something harder to obtain than you might think)

Case-control studies

- Case-control studies are usually performed where there is a scarcity of participants with the 'condition' of interest (i.e. Low prevalence)
- Idea is to collect a group of cases (those with the outcome of interest) and a group that corresponds as much as possible without the 'condition' of interest (controls)
- In this respect the relative balance of cases to controls is an artefact of study design (so in no way reflects prevalence)
- Then (along with measuring observable traits) we ask participants about exposure history
- For this reason, case-control studies are prone to recall bias
- Important to note that (unlike any other health study design) it is the **outcome that defines group membership** (not exposure group)

Matched / stratified case-control studies

- There is also a variant of case-control designs, called matched case-control designs
- These designs attempt to control confounders by **matching** (**stratifying**) **individuals** (**groups of individuals**) into case-control strata based on particular 'known' confounders (e.g. Age, Gender, SES)
- For example, we may age-gender match individuals between the case and control groups
- These designs should be analysed using appropriate methods (e.g. conditional logistic regression) that account for the matched pairs\sets (strata)
- Notably, by adding strata we have to take these effects out of the model (as predictors) → we can't test hypotheses about them

Ecological studies

- Finally we come to the last of the 'data-based' study designs: Ecological studies
- Typically the outcomes from such studies represent aggregate counts (or rates) from different times (e.g. year) or spaces (e.g. countries or provinces)
- Often data are routinely collected (e.g. hospital or government departmental data)
- As the outcomes are often counts (or rates), Poisson regression (or related methods) are often used in these studies
- Probably the study design most prone to confounding, as there are always multiple sources of variation between 'observation' (→ ecological fallacy)

Impact of study design

- The study design can have a major impact on analytical planning (how we plan to analyze the data)
- I would like to clarify what I mean by **analytical planning**. I mean:
  - What is the likely nature of our outcome(s) and predictor(s)?
  - What are the practical issues in sampling (e.g. Random sample, stratified samples, convenience samples)?
  - Based on our study design, what statistical method \ modelling approach should we use?
  - Prospective powering: What sample size is appropriate?

Study design and srength of evidence

Study design + analytical planning = Scientific quality

Biostatistics in Biomedicical studies

Statistics is the science (and art) of dealing with (analyzing or modelling) variation in data to obtain reliable results and conclusions

Biostatistics is the application of statstical methods to the biomedical sciences, health and medicine

Biostatistics in the **public health** and **medical** contexts, is usually (but not always) the application of statistical models to **community-based** and **clinical** samples

Oh, why do we have to learn biostatistics

- Biostatistics represents the scientific framework underpinning most biomedical research
- An understanding of biostatistics is essential for both conducting and disseminating biomedical research

Warning: Why do grants, manuscripts and theses fail?

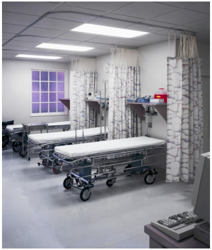
Most unsuccessful grants submitted to funding bodies, and unsuccessful manuscripts/thesis sent to reviewers/examiners, fail in their scientific quality (i.e. poor research design and/or analytical planning), **NOT** contextual significance (importance of the research question)

Case study 1: Epidemilogical study

Monitoring bed use in hospitals

Vast quantities of routinely collected hospital data (⇒ secondary analysis)

- Linking inpatient length of stay to diagnosis
- Summarizing data
- Presenting data
- Distributions (length of stay)
- Explore other (modifiable) factors associted with length of stay
- Costing models and consider possible interventions



Case study 2:

Keyhole surgery vs. traditional (slice and dice) surgery

Taking our (observational) length of stay data...

- Compare patients who underwent **keyhole** surgery with those undergoing**traditional** surgery (Medians: **9** days versus **11** days)
- Drawing conclusions: Real effect? Chance?
- Are there things that muddy the water (confounders: other explanations)
  - Keyhole patients ? younger?
  - Traditional surgery patients -developed wound infections?
  - Reasons (non-random) reason for type of surgery
  - Sample selection and sample size
- Comparing groups (t-test , ANOVA)

What type of study design is this????

Case study 3:
Support people who wish to stop smoking

Research question: Effectiveness of nicotine replacement therapy

(Pre-)Two equivalent groups → administer intervention

(A) **New treatment**

(B) **Existing treatment**

After some time: % who quit

(A) **New:** 63% quit

(B) **Existing:** 54% quit

Can we say the new treatment is superior ?
Can we say it is Equivalent (i.e. As good) or non-inferior?
Real ? Chance ? (chi-squared χ^2....Logistic regression)

Case study 4:
Exercise and Health

Research question:

1. **What are determinates of regular exercise?**
2. **What are true benefits (in terms of health outcomes)?**

- Identify connections
- Strength of these relationships (correlations / odds ratios)
- Strong enough to make predictions (regression models)

Case study 5:

Cervical cancer vaccination

- Already shown to severely reduce the risk of cervical cancer incidence.
- ...But low uptake of cervical vaccine compromise reduction in incidence.
  - Reasons? Many? , cultural values, age, family history, marital status, educational, access to transport, income etc..
  - Is there a high risk group who need heavier canvassing?
- Multivariable methods

Study objectives and analytical approach

- There are a number of objectives researchers might have in a project
- In my experience, a large majority of studies fall under three main categories:
  1. **"Predictive" Modelling**: Performing analysis for predictive purposes
     - For example: Diagnostic testing or prognostic models
  2. **Hypothesis testing**: Answering a previously posed (and very specific) research question
     - Includes most clinical trials, and epidemiological studies with a particular **study effect**
  3. **"Exploratory" modelling**: Trying to identify potential risk/protective factors (e.g. where disease or other health outcome epidemiology not yet well understood)

Predictive modelling

- 'Modelling' (in the classical sense) is about predictive modelling
- Much more common in clinical setting, than in population studies, particularly:
  1. Diagnostic tests
  2. Progostic models and predicting survival
  3. Continuous outcomes
- Predictive modelling relies on very strong relationships
- Need high R^2 for continuous outcome models and classification accuracy for binary outcome modelss (i.e. high sensitivity and/or specificity, high AUC in ROC curve).

Take home point: Preditive models

"Fit" to the data (rather than statistical significance) tis perhaps the most important factor in selecting the 'best' model in predictive studies

Addressing a specific research question

- Studies that consider a specific research question typically address a single hypothesis
- As such (and if sensible) they should be prospectively and formally powered (i.e. **Formal** sample size calculation)
- Examples of these types of studies include RCTs and population studies considering a particular and **well-understood** risk factor

Exploratory studies

- Exploratory studies are usually done when we don't have a strong (scientific) understanding of the system (e.g. disease or population) under consideration
- For this reason, these studies typically involve a large number of covariates which may represent potential risk factors, confounders and/or effect modifiers
- Limited knowledge (and lack of a specific research hypothesis) means we must be less formal in our powering of such studies. For example **rules of thumb** such as:
  - Continuous outcomes, t use $N = 60 + 5k$ where k is the number of covariate under consideration (more later)
  - Binary outcomes: 10 events (e.g. disease present) per covariate rule.

The Diabetes and Hypertension dataset

I will now introduce you to the Diabetes and Hypertension dataset that we will use throughout this subject:

- National Thai study documenting cases of Type 2 Diabetes Mellitus and Hypertension from 600 hospitals across Thailand (2554-2556)
- We will consider a very 'trimmed down' version of the dataset:
  - 2500 Diabetics (who may or may not also be hypertensive) - Hands-on data
  - 1000 randomly selected observations - Assignment data
  - Of the hundreds of variables measured in the study we will only consider a small subset (next page)
- Today we are just going to use this data to practice reading data in, basic graphics and some summary statistics

The Diabetes and Hypertension dataset

Variables measured

Outcomes	Study effect	Covariates
a1cyn	**ht**	**sex**
bpyn		age
ldlcyn		**religion**
all3yn		duradm
any3yn		bmi
a1c		**bmigroup**
ldlc		

Blue variables are categorical (all variables are binary) and grey variables are continuous

Variable description: Outcome

The outcomes we will consider represent the "ABC" clinical goals often used to assess the quality of diabetes care:

A a1cyn→ Haemoglobin A1C(yes: $< 7\%$; no $\geq 7\%$)

B bpyn→ Bloop pressure(yes: $< \frac{130}{80}$mmHG; no $\geq \frac{130}{80}$ mmHg)

C ldlcyn→ Low density lipoprotien-Cholesterol(yes: < 100mg/dL; no ≥ 100mg/dL)

Haemoglobin A1C (a1c) and LDL cholesterol ldlc are also in the dataset in their continuous form

We will also consider the 'aggregate' quality performance outcomes:

- all3yn: **All** three (ABC) clinical targets are met
- any3yn: **Any** of the three (ABC) clinical targets is met

Something to think about...

The US army draft during a certain war considered potential recruits doing a 100 question multiple choice exam.

- Those scoring ≥ 90 went into officer training
- Those scoring between $6 - 89$ went into basic training
- **Those scoring ≤ 5 went into officer training**

WHY????

Any questions??????

> *To call the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.*

Ronald A. Fisher

Thank-you!!!!!!

QUESTIONS???