Southern African Journal of Anaesthesia and Analgesia. 2021;27(6 Suppl 1):S122-125
https://doi.org/10.36303/SAJAA.2021.27.6.S1.2690

South Afr J Anaesth Analg
ISSN 2220-1181 EISSN 2220-1173

FCA REFRESHER COURSE

Types of data and study design, and measures of disease frequency and association

TGA Leonard

Department of Anaesthesia, School of Clinical Medicine, Faculty of Health Sciences, Chris Hani Baragwanath Academic Hospital, University of the Witwatersrand, South Africa
Corresponding author, email: *tristan.leonard@wits.ac.za*

Keywords: measures of disease frequency, types of data, study design

Introduction

Knowledge of statistics is both examinable in the final college fellowship examination and is a key skill for all practising clinicians. To be able to conduct and critique research, it is important that registrars have a working understanding of statistical principles, tests, and models. This review will cover types of data and study design found in medical literature and will explain measures of disease frequency and association.

Types of data

The most important types of data that will be encountered are quantitative, categorical, and time-to-event variables.[1] Quantitative variables are numerical variables that can be added, subtracted, multiplied, or divided, such as age, body mass index, pulse, or blood pressure. These quantitative variables can be continuous or discrete. Continuous quantitative variables can theoretically take on any value within a given range and are limited only by how precisely they are measured, e.g. height of 174.995 cm. Discrete continuous variables can only take on a certain value such as a whole number, e.g. number of children.[1]

Categorical variables are not numerical in nature and can be binary, nominal, or ordinal.[1] Binary categorical variables have two categories such as dead or alive, treatment or placebo group. For data coding purposes, these can be coded as 0 or 1. Nominal categorical variables are unordered categories such as blood type, marital status, or occupation, whereas ordinal categorical variables are ordered categories such as cancer stage, birth order, or ratings on a Likert scale. Some continuous quantitative variables can be transferred into ordinal variables for data analysis, such as converting age into an age range category or converting the number of alcoholic beverages a person consumes into none, light, moderate, and heavy consumption.

The final variable that is encountered in medical research is the time-to-event variable. This is the time taken for an event to occur and is a hybrid variable with a continuous component (time) and a binary component (event: yes or no). Examples of these variables are time to death, time to recovery, or time to development of disease. These variables can only be described from studies that follow people over time, such as cohort studies or randomised trials.

Understanding the types of data is important because certain statistical tests and models are specific to each type of variable and knowledge of this will allow the reader to decide if the correct statistical tests have been applied in a research paper.

Types of study design

Understanding how a study is designed will allow the reader to interpret data in context and will allow for the correct conclusions to be made about frequencies and associations. Broadly, studies can be observational or experimental. Observational studies do not have an intervention and are easier and cheaper to perform. However, their biggest limitation is the risk of confounding. Confounding is when a variable influences both the dependent and independent variables and causes a spurious association.[2] It can be difficult to isolate a single risk factor in an observational study. Some techniques which can avoid or control for confounding are adequate randomisation or matching during the study design phase, and the use of multivariate regression in the analysis phase to adjust for confounding. Experimental studies have an intervention that is selected by the investigator and are better suited to identify associations and causation. Experimental studies should also use randomisation, which reduces the risk of confounding.[2] Below is a description of the most common types of observational and experimental studies.

Cross-sectional studies

A cross-sectional study aims to measure the prevalence of a disease and exposure to risk factors at one point in time and explore associations between them.[2] They are fairly inexpensive and easy to perform and can provide a snapshot of a population and provide information on the prevalence of risk factors and outcomes. They select participants from a population randomly and this helps to make the findings more generalisable. The limitation of these studies is that correlation does not equal causation, and it is difficult to know whether the exposure came before the outcome in question. There is also the risk of recall bias where participants that have the outcome or disease will

report their risk factors or exposures differently than those that have not had the outcome.[2]

Case-control studies

Case-control studies select participants based on them having the disease or outcome and because the participants already have the outcome are an ideal tool for investigating a rare disease. These participants are then asked retrospectively about exposures or risk factors. Subjects are selected who have the disease (case subject) and are compared with subjects who are similar to them and are disease-free (control subjects).[2] The greatest challenge in case-control studies is the selection of appropriate control subjects, these need to be similar enough to the case subjects so that associations can be made. Furthermore, the disease is already present in the cases and as such, causation cannot be concluded. Recall bias is also a risk as cases may wish to place more emphasis on their exposure and risk factors. Case-control studies cannot estimate prevalence or incidence because the percentages of cases in the sample do not reflect the percentage in the general population.[2] The only valid measure of disease association that can be determined from case-control studies is an odds ratio.

Prospective cohort studies

Prospective cohort studies measure risk factors on people who are disease-free at baseline and are then followed up over time until some of them develop the disease or outcome to calculate the rate or risk of developing a disease. The greatest strength of these studies is that participants are disease-free at baseline and as such temporal relationships can be determined. Additionally, they can be used to investigate the effects of rare exposures and can also identify multiple outcomes.[2] These studies sample from the general population and their results are often generalisable. The risk of recall bias is also eliminated. Prospective cohort studies can determine incidence and cumulative incidence. These studies are costly and time consuming. Large sample sizes and extended follow-up times may be required and losses to follow are a common occurrence.[2]

Retrospective cohort studies

A retrospective cohort study gathers a cohort after the outcome has occurred and uses stored data that was collected for some other purpose prior to the development of the outcome or disease. The key is that exposure data must have been collected before the outcome and researchers then attempt to link this data to outcomes.[2] As the data has already been collected, these studies are cheaper and faster to perform and can provide similar information as prospective cohort studies. The limitations lie in the quality of the data that was collected as this data was not specifically collected for the study and, as such, may not have been measured accurately, or in enough detail. There may also be loss to follow-up as everyone in the cohort may not have all the data points required.[2]

Nested case-control studies

In a nested case-control study, cases and controls are drawn from within a prospective cohort study. Cases that develop the outcome are compared with matched controls selected from the cohort who did not develop the outcome.[3] An example of a nested case-control study is the use of an expensive biomarker or assay. It is too expensive to run this test on the entire cohort and so researchers wait for enough subjects to develop the outcome, match these with controls from the same cohort, and then run the assay on this group only. The blood samples must have been collected prior to the development of the disease. These studies can also be done retrospectively with data or samples that were collected from a cohort and then a nest of cases is matched with controls.[3]

Randomised controlled trial

Well-designed randomised controlled trials (RCTs) are considered the gold standard for determining causality between risk factors and outcomes.[2] Participants are randomly assigned to various intervention or control arms and are then followed up over time. The addition of blinding and placebos also strengthens the findings from RCTs and further reduces the risk of confounding. While these studies are considered the gold standard, they are very expensive and cannot practically look at long-term outcomes. There may be ethical concerns with placebo groups or interventions that turn out to be harmful. Loss to follow-up can also affect the outcome of RCTs and researchers need to ensure the sample group is representative of the population for the findings to be generalisable.

Measures of disease frequency

The following are commonly used descriptions of the frequency of an outcome or disease. It is important to understand the definitions of these measures and to know what types of studies can accurately report incidence vs prevalence. Disease frequency cannot be described from case-control studies as participants already have the disease.

Incidence

This is the rate (over time) at which people are developing a disease. These are new cases of the disease or outcome.[4] To describe incidence, a study needs to follow subjects over time such as a cohort study or RCT. It is most often described as the number of events per person-year, e.g. five cases per 100 person-years. Person-year is the denominator most often used for incidence rates and is a summation of the total amount of follow-up time across all participants in the study.

$$\text{Incidence} = \frac{\text{Number of subjects developing the disease}}{\text{Total time at risk for the disease for all subjects followed}}$$

Cumulative risk or cumulative incidence

This is the proportion of people who develop a disease in a specified period of time. For example, 1% of smokers develop lung cancer in one year. This can also be described as the probability of a subject developing the disease or outcome over

a defined time.[4] This can be thought of as an estimation of the risk of disease in an individual person. These risks have to be presented with a defining time period.[4] Due to the follow-up over time required, this can only be described from a cohort study or RCT.

$$\text{Risk} = \frac{\text{Number of subjects developing the disease over a time period}}{\text{Total number of subjects followed over that time period}}$$

Prevalence

Prevalence is the proportion (percentage) of people who have a disease or outcome at a point in time, and it will contain both new and old cases. It is a measure of disease status in a population.[4] Prevalence can be described from a cross-sectional study.

$$\text{Prevalence} = \frac{\text{Number of subjects having the disease at a time point}}{\text{Total number of subjects in the population}}$$

Measures of association

Studies can describe association in absolute or relative terms. Journals will often choose to present relative risks as these numbers can appear more impactful but often will translate into a very minor change in absolute risk.[5]

Absolute differences in risks or rates

Absolute risk refers to the subtraction of the risk of disease in one group from the risk of disease in another group, often the placebo vs treatment group. This can be described using incidence rate, cumulative risk, or prevalence depending on the study design.[5] This is the overall difference in the outcome between two groups. A drug that decreases the incidence of upper gastrointestinal bleeding from four bleeds per 100 person-years to two bleeds per 100 person-years will have an absolute rate reduction of two events per 100 person-years. The same drug that decreases the risk of bleeding from 3% to 1% will have an absolute risk reduction of 2%.

Absolute risk reduction = control event rate – exposure event rate

Number-needed-to-treat (NNT), and number-needed-to-harm (NNH)

This is directly related to the absolute risk and tells us how many people would need to be treated to prevent one case of disease, the NNT, or how many people would need to be given an intervention to cause one case of disease or adverse outcome, the NNH.[5] Mathematically it is the inverse of the absolute risk difference between groups. Using the example above, the drug that reduced the incidence of bleeding by two events per 100 person-years would have an NNT of 100/2 = 50 people need to be treated to prevent one case of bleeding.

NNT = 1/absolute risk reduction

NNH = 1/absolute risk increase

Relative risk

Measure of relative risk includes the hazard ratio (ratio of two instantaneous incidence rates), risk ratio (ratio of two percentages), and odds ratio (ratio of two odds). Relative risk is calculated when the absolute risk of an outcome in the treatment or exposed group is divided by the absolute risk in the control or placebo group.[5] It can also be defined as the ratio of probability of an outcome in an exposed group to the probability of an outcome in the unexposed group.

- Rate ratio: % increase or % decrease in the rate of an outcome
- Risk ratio: % increase or % decrease in the risk of an outcome
- Odds ratio: % increase or % decrease in the odds of an outcome

Generally, 1.0 means no effect (null or no difference), < 1.0 is a protective effect (decreased risk) and > 1.0 is a harmful effect (increased risk).

Authors often choose to report relative risks because they are unitless, simpler, and appear more dramatic. Regression analysis techniques also produce relative risks.[5] Relative risks should be approached with caution when interpreting the medical literature; if a treatment reduces the risk of a disease from 2% to 1% the relative risk reduction will be reported as a 50% decrease when the absolute risk reduction is only 1%.

Table I: Relative risk

Exposure or treatment status	Event or outcome occurred	
	Yes	No
Exposed/treatment	**A**	**B**
Unexposed/control	**C**	**D**

$$\text{Relative risk} = \frac{(A/(A+B))}{(C/(C+D))}$$

A note on the odds ratio (OR)

From Table I:

$$\text{Odds ratio} = \frac{A/B}{C/D}$$

The odds ratio is the probability of an event happening divided by the probability of it not happening and is a ratio of two odds rather than risks.[6] The value of an odds ratio is that it is the only valid measure of disease association that can be reported from case-control studies and the use of logistic regression gives odds ratios. Logistic regression is used when a study has a binary outcome, and mathematically odds have better properties for this modelling.[6] Researchers can also adjust for confounding and examine multiple predictors simultaneously, as well as evaluate the effect of a continuous predictor.[6]

It is therefore important to be able to understand and interpret the odds ratio correctly as it can be misleading in some circumstances. As an example, if the odds of an outcome in an exposed group are 1 to 1 (odds of 50/50%) and the odds in another group are 1 to 3 (odds of 25/75%) then the odds ratio is 1.0 divided by 0.3 and the odds ratio is 3.0. This means that the odds of someone having the outcome when exposed are three

times greater than when someone is not exposed. This does not mean that the risk is tripled; in fact, the risk is only doubled. In the example, the outcome occurred in 50% of the exposed group and 25% of the unexposed group – a doubling of relative risk and a 25% increase in absolute risk.

Key to understanding this is that when an outcome is common, the odds ratio distorts the effects – in this example, the outcome in one group was 50% and the outcome in the other group was 25% – these would be classed as common outcomes. The odds ratio here distorts the effects. When an outcome is rare (< 10% in the reference group) the odds ratio is more similar to the risk ratio.[6] Using the same example as above, if the outcome in the exposed group had been 10%, then the odds are 1 to 9 (10/90%), and in the unexposed group 1%, the odds are 1 to 100 (1/99%), and the odds ratio is 0.1 divided by 0.01 which is 10, and this is similar to the actual risk ratio of 10. Therefore, for a rare outcome, the odds ratio can be interpreted as the risk ratio, but for a common outcome, it must be interpreted with caution.

When confronted with an odds ratio from a cohort study with a common outcome, there is a formula to convert the odds ratio into a relative risk.[7] It uses the odds ratio and the probability of the outcome in the reference group (Pref) to give an adjusted risk ratio.

$$\text{Risk ratio} = \frac{\text{OR}}{(1 - \text{Pref}) + (\text{Pref} \times \text{OR})}$$

ORCID

TGA Leonard https://orcid.org/0000-0003-4426-3972

References*

1. Thomas E. An introduction to medical statistics for health care professionals: describing and presenting data. Musculoskelet Care. 2004;2(4):218-28. https://doi.org/10.1002/msc.73.
2. Sainani KL, Popat RA. Understanding study design. PM&R. 2011;3:573-7. https://doi.org/10.1016/j.pmrj.2011.04.001.
3. Ernster VL. Nested case-control studies. Prev Med. 1994;23(5):587-90. https://doi.org/10.1006/pmed.1994.1093.
4. Noordzij M, Dekker FW, Zoccali C, Jager KJ. Measures of disease frequency: prevalence and incidence. Nephron Clin Pract. 2010;115(1):17-20. https://doi.org/10.1159/000286345.
5. Sainani KL. Communicating risks clearly: absolute risk and number needed to treat. PM&R. 2012;4:220-2. https://doi.org/10.1016/j.pmrj.2012.01.001.
6. Sainani KL. Understanding odds ratios. PM&R. 2011;3:263-7. https://doi.org/10.1016/j.pmrj.2011.01.009.
7. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. 1998;280(19):1690-1.https://doi.org/10.1001/jama.280.19.1690.

* The outline of these notes is taken from Medical Statistics I: Introduction to data analysis and descriptive statistics from the University of Stanford School of Medicine.