



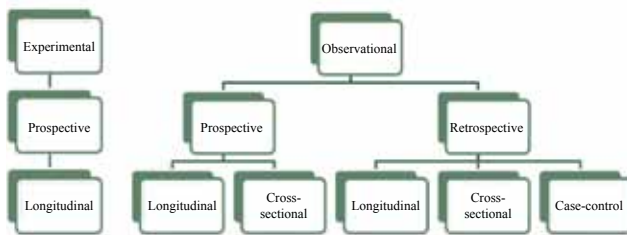
Statistics of observational studies: Basic concepts for medical writers

by Andrea Rossi

Research can be crudely divided into observational and experimental studies. In observational studies the researcher collects information about one or more groups of subjects but does nothing to affect them [1]. While experimental trials have great internal validity, meaning that they can be replicated with a good probability of having the same results under the same experimental conditions, observational studies have a higher external validity, meaning that they are more likely to be applicable to everyday clinical practice. For this reason, results of well designed, conducted, analysed and disclosed observational studies can have an immediate consequence on daily practice.

Study design will not be fully covered; a simple scheme summarising the main types of observational studies and differences with experimental studies is available in Figure 1.

Figure 1: Types of research design



An observational study must not influence usual clinical practice and the design has to reflect this requirement. Furthermore, observational studies are not randomised or blinded as randomisation and blinding are not part of daily practice. For this reason, patients receive the treatments and assessments they normally would. Once patients are included in an observational studies, they may be grouped into a ‘cohort’ of patients who, for instance, were prescribed the same drug. The sample size of an observational studies is usually higher than the one used for an experiment answering the same question, because bias and variability expected in an unselected population is higher than in a selected population included in any experimental trial and, consequently, a higher sample size is needed to assess the same difference with the same power and level of probability.

Statistical analyses of observational studies

The first step in the analysis of a set of data is to describe basic data, including demographic and clinical characteristics of subjects being studied. Standard statistics such as number of observations, means, standard deviations and

frequencies are the most commonly used. In observational studies these statistical descriptors are usually enough to properly describe the study population [1]. But how much is the observed population representative of the whole population? How do we compare sets of data, especially in view of the desire to generalise the findings? Some statistical methods are peculiarly created and used in observational studies; the ones that are most commonly used are described below.

Prevalence and incidence

In epidemiology, the prevalence of a disease in a population is defined as the total number of cases of the disease at a given time, or the total number of cases in the population, divided by the number of individuals in the population. It is used as an estimate of how common a condition is within a population.

Lifetime prevalence (LTP) is the number of individuals that at some point in their life (up to the time of assessment) have experienced a ‘case’ (e.g., a disorder), compared to the total number of individuals (i.e. it is expressed as a ratio or percentage) in the studied population. Often, a 12-month prevalence (or some other type of ‘period prevalence’) is used in conjunction with lifetime prevalence. There is also point prevalence, the prevalence of disorder at a more specific (a month or less) point in time. There is also a related figure lifetime morbid risk—the theoretical prevalence at any point in life for anyone, regardless of time of assessment.

Incidence is a measure of the risk of developing some new condition within a specified period of time. Although sometimes loosely expressed simply as the number of new cases during some time period, it is better expressed as a proportion or a rate with a denominator [1].

Incidence proportion (also known as cumulative incidence) is the number of new cases within a specified time period divided by the size of the population initially at risk. For example, if a population initially contains 1,000 non-diseased persons and 64 develop a condition over 4 years of observation, the incidence proportion is 64 cases per 1,000 persons, i.e. 6.4%.

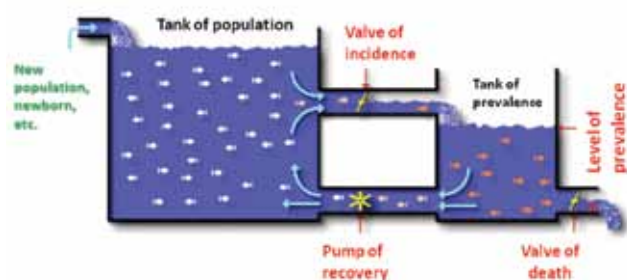
The incidence rate is the number of new cases per unit of person-time at risk. In the same example as above, the incidence rate is 16 cases per 1,000 person-years, because the incidence proportion (64 per 1,000) is divided by the number of years (4). Using person-time rather than just time handles situations where the amount of observation time differs between people, or when the population at risk var-

Statistics of observational studies: Basic concepts for medical writers

ies with time [2]. Use of this measure implicitly assumes that the incidence rate is constant over different periods of time, such that for an incidence rate of 16 per 1,000 persons-years, 16 cases would be expected for 1,000 persons observed for 1 year or 50 persons observed for 20 years.

The relationship between prevalence and incidence is shown in figure 2.

Figure 2: Relationship between prevalence and incidence



Attributable risk, absolute risk, relative risk, odds ratio

Attributable risk is the difference in the rate of a condition between an exposed population and an unexposed population [3].

The absolute risk reduction is the decrease in risk of a given activity or treatment in relation to a control activity or treatment. It is the inverse of the number needed to treat (NNT) [4]. NNT describes the number of subjects to be treated with a certain treatment to have, for instance, one positive outcome when compared with reference therapy. For example, consider a hypothetical drug which reduces the relative risk of colon cancer by 50% over five years. Even without the drug, colon cancer is fairly rare, maybe 1 in 3,000 in every five-year period. The rate of colon cancer for a five-year treatment with the drug is therefore 1 in 6,000, so by treating 6,000 people with the drug, one can expect to reduce the number of colon cancer cases from 2 to 1, meaning that the NNT for the new treatment is 6,000.

NNT is often used to compare the effects of different treatments even between different illnesses [5].

Relative risk (RR) is the risk of an event (or of developing a disease) relative to exposure. Relative risk is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group [6].

The hazard ratio in survival analysis is the effect of an explanatory variable on the hazard or risk of an event. Hazard ratio can be considered an estimate of relative risk.

The odds ratio [7] is a measure of effect size, describing the strength of association or non-independence between two binary data values. It is used as a descriptive statistic, and plays an important role in logistic regression. Unlike other measures of association for paired binary data such as the relative risk, the odds ratio treats the two variables being compared symmetrically, and can be estimated using some types of non-random samples, such as case-control studies.

These apparently ‘difficult to understand’ measures are explained much better by examples than by descriptions, so a brief summary and example of these estimation measure are shown in tables 1 and 2 (with thanks to Wikipedia).

Table 1: Working table example of contingency table [5]

	Example 1: risk reduction		Total	Example 2: risk increase	
	Experimental group (E)	Control group (C)		(E)	(C)
Events (E)	EE = 16	CE = 100	116	EE = 75	CE = 100
Non-events (E)	EI = 136	CI = 150	286	EI = 75	CI = 130
Total subjects (E)	ES = EE + EI = 150	CS = CE + CI = 250	400	ES = 150	CS = 250
Event rate (ER)	EER = EE / ES = 0.1 or 10%	CER = CE / CS = 0.4 or 40%	N/A	EER = 0.5 (50%)	CER = 0.4 (40%)

Table 2: Calculation of Absolute Risk (AR), Relative Risk(s) (RR), Number Needed to Treat (NNT), Odds Ratio (OR), Attributable Risk (AR) and Attributable Risk Percent (ARP) from contingency table [5]

Equation	Variable	Abbr.	Example 1	Example 2
EER - CER	< 0: absolute risk reduction	ARR	-0.3, or -30%	N/A
	> 0: absolute risk increase	ARI	N/A	0.1, or 10%
(EER - CER) / CER	< 0: relative risk reduction	RRR	-0.75, or -75%	N/A
	> 0: relative risk increase	RRI	N/A	0.25, or 25%
1 / (EER - CER)	< 0: number needed to treat	NNT	(-1) 3.3	N/A
	> 0: number needed to harm	NNH	N/A	10
EER / CER	relative risk	RR	0.25	1.25
(EE / EI) / (CE / CI)	odds ratio	OR	0.167	1.5
EE / (EE + CE) - EI / (EI + CI)	attributable risk	AR	(-1) 0.34, or (-)34%	0.095, or 9.5%
(RR - 1) / RR	attributable risk percent	ARP	-300%	20%

Covariates, confounders and effect modifiers

Covariates are simply variables (a characteristic of a person, object or phenomenon) which may take on different values. Covariates can be confounders, effect modifiers, or other important prognostic factors.

Factors that could mask the association between a cause and effect (for example a risk factor and development of a disease) are called confounding variables. A confounder is associated with the exposure (risk) we are studying and is also an independent risk factor or predictor for the outcome. Assume we want to know if there is an association between drinking coffee and pancreatic cancer. In this case, smoking is a confounder. It is associated with coffee-drinking (i.e., many smokers also drink coffee), and it is an independent predictor of pancreatic cancer. Unless we control for this confounder in the organisation of our study population (e.g., through separating ‘coffee-only’ and ‘coffee + smoking’ groups of people), we can’t determine the association between coffee drinking and pancreatic cancer.

Identifying confounding factors or variables is also critical in assessing the effectiveness or safety of a drug treatment. For example, early research into hormone replacement therapy (HRT) showed that it was effective in reducing symptoms of menopause and also helped protect women against cardiovascular disease. Later research refuted those claims, and eventually it was discovered that women who took HRT also engaged in other steps to protect their health and youth, such as diet and exercise. The association between HRT and reduced risk of cardiovascular disease might have been due to the confounding factor of healthy lifestyle among HRT users.

Statistics of observational studies: Basic concepts for medical writers

- Confounding is a nuisance effect. It distorts the association between an exposure of interest and an outcome, because the confounder affects the outcome and is unequally distributed between the subjects who are exposed or unexposed to the exposure of interest. So we aim to control confounding.

A covariate can also be an ‘effect modifier.’ An effect modifier is an independent variable that could modify the effect of the outcome we are studying but does not cause or predict that outcome. This type of covariate is sometimes called an interaction modifier because it systematically interacts with or modifies the association between the predictor (independent variable) and the outcome (dependent variable).

Effect modification is a real effect, independent of the study design. It modifies the association between an exposure and an outcome according to the level of a third factor. Unlike a confounder, an effect modifier is not linked to the exposure of interest. In observational studies we want to detect and report effect modification, not control it. Exploring the nature of effect modification can be very helpful in understanding the biological processes underlying an association between an exposure and an outcome.

Suppose we want to investigate the effect of vitamin X supplements on childhood growth. Among children who are vitamin X deficient, it is likely that vitamin X supplements will be associated with increased growth. However, among children who are not deficient in vitamin X, supplements may have no effect. This is an example of effect modification, also known as interaction. In this situation, the association between an exposure (here, vitamin X supplements) and an outcome (childhood growth) varies by levels of a third factor (level of vitamin X before supplementation).

Regression analysis

Regression analysis refers to techniques for the modelling and analysis of numerical data consisting of values of a dependent variable (also called a response variable) and of one or more independent variables (also known as explanatory variables or predictors or covariates). The dependent variable in the regression equation is modelled as a function of the independent variables, corresponding parameters (‘constants’), and an error term. The error term is treated as a random variable and represents unexplained variation in the dependent variable.

Parameters are estimated to give a best fit to the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used. In observational studies we are often interested in the way one variable is influenced by several variables. The statistical method used to analyse that type of data, when the outcome variable is continuous, is multiple linear regression, known also as multiple regression. Multiple regression analysis yields a regression model in which the dependent variable (the dependent variable meaning the study outcome) is expressed as a combination of the explanatory variables.

Regression models predict a value of the y variable given known values of the x variables. Prediction within the

range of values is known as interpolation. Prediction outside the range of the data is known as extrapolation, and this is more risky.

Even more difficult and risky is to identify, from a large number of explanatory variables, those which are genuinely related to the dependent variable and, finally, assess how well the model obtained fits the data. Selecting a small subset of a large number of possible explanatory variables to include in a final regression model is a highly complex task, which is informed by both statistical and clinical considerations. In fact, while the statistician tries to identify from a large number of variables those which are related to the dependent variable, he also assesses how well the model obtained fits analysed data. Because of this apparent paradox, results from such analyses have any clinical relevance only in the context of the evidence belonging to other studies and clinical practice; ideally, results not reflecting acquired evidence should be confirmed in ad-hoc designed studies.

Logistic regression

Logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. It is a generalised linear model used for binomial regression, meaning that for those studies where the variable of interest is the presence or absence of a certain condition, instead of using multiple linear models, logistic regression models are used. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. For example, the probability that a person has a heart attack (the binary variable) within a specified time period might be predicted from knowledge of the person’s age, sex and body mass index [8]. Logistic regression is used extensively in the medical and social sciences as well as marketing applications such as prediction of a customer’s propensity to purchase a product or cease a subscription or to evaluate survival [9].

The basic principle of logistic regression is much the same as for ordinary multiple regression. The main difference is that instead of developing a model that uses a combination of values of a group of explanatory variables to predict the value of a dependent variable, logistic regression models predict a transformation (from yes to no or vice-versa) of the dependent variable [1].

Post-marketing (non-study) exposure: EMEA requests

Where marketing of the medicine has occurred, the marketing authorisation holder should provide data on patients exposed after marketing. Exposure data based on the number of kilograms of medicinal product sold divided by the average dose is only valid if the medicinal product is always taken at one dose level for a fixed length of time—which is not the situation with most medicinal products.

Signal events belonging to observational post-authorisation studies should be analysed according to guidelines on the use of statistical signal detection methods in the Eudra-

Statistics of observational studies: Basic concepts for medical writers

vigilance data analysis system [10]. These guidelines state which statistical analyses have to be used for the estimation of adverse events frequencies and confidence intervals after product commercialisation. Furthermore, these guidelines describe how these results have to be reported, validated and, finally, integrated with results belonging from other sources.

Disclosing observational studies

Because of the complexity of some analyses used for observational studies, it is important to highlight the importance of reporting all details of more complex methods, such as multiple regression. The strategy adopted, and all variables included in the model (not only those remaining in the final model) have to be widely reported making the reader aware of the assumptions and possible limitations. It is essential to explain the coding of categorical variables, especially those featuring in models that are described. For example, there are numerous ways of categorising the amount of alcohol assumed daily.

Descriptive statistics of each dependent and independent variable should be reported. For each model, regression coefficients should be described in detail and their standard error should be given together with any other result that might support complete understanding of the structure of the model by the reader.

In 2007, several prominent medical researchers issued the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, in which they called for observational studies to conform to 22 criteria that would make their conclusions easier to understand and generalise [11].

Andrea Rossi

Eli Lilly Italia S.p.A
Sesto fiorentino, Italy
rossi_andrea_a@lilly.com

Reference:

1. Altman, Douglas G. *Practical statistics for Medical Research*. London. Chapman & Hall, 1991.
2. Coggon D, Rose G, Barker DJP. Quantifying diseases in populations. *Epidemiology for the Uninitiated*. Fourth edition. *BMJ*, 1997.
3. Comparing disease rates. <http://www.bmj.com/epidem/epid.3.html>.
4. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:728-33.
5. Relative risk. http://en.wikipedia.org/wiki/Relative_risk#cite_note-mpmid14695382-0.
6. Siström CL, Garvan CW. Proportions, odds, and risk. *Radiology* 2004;230(1): 12-9.
7. Edwards, AWF. The measure of association in a 2x2 table. *Journal of the Royal Statistical Society* 1963; Series A 126 (1): 109-114. <http://www.jstor.org/stable/2982448>.
8. Hilbe, Joseph M. *Logistic Regression Models*. Chapman & Hall/CRC Press, 2009.
9. Agresti, Alan. *Categorical Data Analysis*. New York. Wiley-Interscience, 2002.
10. Guideline on the use of statistical signal detection methods in the eudravigilance data analysis system. EMEA. 2008. p. Doc. Ref. EMEA/106464/2006 rev. 1.
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Med* 2007;4(10):e296.

Statistics: What a joke!

Statisticians are funny people. They must be because which other topic could attract almost 200 jokes?

Gary C. Ramseyer, Emeritus Professor of Psychology at Illinois State University, has a website to help instructors and students experience the lighter side of statistics. He has gathered together what he refers to as the greatest collection of statistics jokes on the Internet—and there is no reason to disbelieve his claim. The galleries of jokes named after famous statisticians can be accessed through <http://my.ilstu.edu/~gcr Ramseyer/Gallery.html>.

Here are a few examples from the short jokes:

It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest.

Charlie Brown in a Peanuts cartoon was addressing his baseball team at the end of the season. He recited numerous dismal statistics such as: Runs scored by us 12, by opponents 125. At the end of the speech he yells out: “And what are we going to do about it?” to which the team answers in unison: “Get a new statistician!”

Three ladies were talking about catching husbands. Kate said she was flying to Chicago for the International Conference of Statisticians. Sue looked puzzled; Julie said, “Huh?” Kate responded by telling them that 86% of Statisticians were single males under the age of 37. Sue said, “Wow! Odds are good!” Julie said, “Yeah, but the goods are odd.”

Never use Excel for statistical analysis

A talk by Jonathan D. Cryer given at a statistical meeting in 2001 concludes that Microsoft Excel should not be used for statistical analysis. Excel meets almost none of the criteria of a good graph and the vast majority of chart types, default scatterplots and histograms etc. offered by Excel are faulty. Cryer’s talk in which he sets Excel’s defects in clear and concise bullet point and offers examples can be accessed at <http://www.cs.uiowa.edu/~jcryer/JSMTalk2001.pdf>

Defining the statistician

Stephen Senn gives some amusing definitions of statisticians on his website:

Theoretical statistician: A second class mathematician who imagines that he is a first class statistician.

Applied statistician: A second class statistician who imagines that he is a first class scientist.

Medical statistician: A second class scientist without any imagination.

<http://www.senns.demon.co.uk/wdict.html>