(95%) Confidence interval	This is an estimated range of values calculated from a given set of sample data which are likely to contain the ‘true’ population value, e.g. mean BMI. By “contain the true value”, we mean that the true value lies above the lower value of the confidence interval but below the upper values of the confidence interval. For example, suppose that the sample mean BMI is 25, with a 95% confidence interval for the mean BMI of 23.5 to 26.5. If you take 100 samples of patients, measure their mean BMI, and calculate the 95% CI for each sample, the population mean would lie within 95 of those 100 95% CIs. See the Reading on confidence intervals in the first course.
(point) Estimate	A single estimate of a measure that is calculated from the sample, e.g. mean BMI. It serves as a estimate of the population parameter (true value)
(population) Parameter	A single statistic or measure of interest in the population. We are unlikely to study the population as this is often unfeasible, so parameters are usually unobservable and instead we estimate them from the sample.
Alternative hypothesis	The alternative hypothesis is the converse of the null hypothesis. The alternative hypothesis is often that a difference between groups does exist. If the null is rejected due to a small p-value, then we can accept the alternative. If the null hypothesis is not rejected using statistical inference, we cannot assume that the alternative hypothesis holds. Instead, we can only conclude there was not enough evidence to reject the null hypothesis.
c statistic (area under the ROC curve)	Also known as the model's discrimination. For logistic regression, it measures how more likely the model is to give a higher probability to a patient who has the outcome of interest than to one who does not in fact have it. High values (nearer to 1) are best. 0.5 is useless.
Case	A case is an individual with the outcome under study. Epidemiological research is based on the ability to quantify the occurrence of disease in populations. This requires a clear definition of what is meant by a case. This could be a person who has the disease, health disorder, or suffers the event of interest (by “event” we mean a change in health status, e.g. death in studies of mortality or becoming pregnant in fertility studies).
Censoring	In survival analysis, censoring refers to our lack of knowledge about a patient, particularly whether they had the outcome of interest, e.g. because the study ended or they were lost to follow-up

Chi-squared test	This is a statistical procedure for testing whether two proportions are similar (e.g. whether the proportion of people eating their five portions of fruit and veg a day in Ghana is significantly different from the proportion of people eating their five a day in India).
Collinearity	Collinearity is when predictor variable/s in a multiple regression model can be linearly predicted from the other predictor varaible/s with a substantial degree of accuracy. This is a problem.
Control **(as opposed to a case)**	A control is a person without the outcome under study (in a type of epidemiological study called a case-control study) or a person not receiving the intervention (in a clinical trial, as in the Parkinson’s disease example). The choice of an appropriate group of controls requires care, as we need to be able to draw useful comparisons between these controls and the cases/intervention group.
Correlation coefficient	A measure of how two variables depend on each other. The value of either the Pearson or the Spearman rank correlation coefficient can lie between -1 and +1, where zero means no correlation at all.
Count	The most basic measure of disease frequency is a simple count of affected individuals. The number (count) of cases that occurred in a particular population is of little use in comparing populations and groups. For instance, knowing that there were 100 cases of lung cancer in city A and 50 in city B does not tell us that people are more likely to get lung cancer in city A than B. There may simply be more people in city A. However, the number of cases may be useful in planning services. For instance, if you wanted to set up an incontinence clinic, you would want to know the number of people with incontinence in your population.
Covariate	See "Predictor". Literally, one thing that varies (is associated statistically) with another thing.
Exposure	When people have been ‘exposed’, they have been in contact with something that is hypothesised to have an effect on health, which can be either positive or negative e.g. tobacco, nuclear radiation, pesticides in food (all negative effects), physical exercise and eating fruit and vegetables (all positive effects). This is the most obvious meaning of 'exposed', but it can also refer to any patient characteristic or risk factor for the outcome of interest. This concept will be covered in the epidemiology specialisation.
Hazard	In survival analysis, the hazard is the risk of having the outcome of interest, e.g. death, given that the patient has not already had it. One hazard is divided by another to give the hazard ratio for a particular predictor.
Heteroscedasticity	When the variability of a variable is **unequal** across the range of values of a second predictive variable

Homoscedasticity	When the variability of a variable is **equal** across the range of values of a second predictive variable
Hypothesis	A statement that can be tested using quantitative evidence (data) in a hypothesis test, the foundation of modern science.
Interaction	An interaction occurs when a predictor variable has a different effect on the outcome depending on the value of another predictor variable. This is also called effect modification in epidemiology.
Least squares regression	The statistical method used to determine a line of best fit in a linear regression model by minimizing the sum of squared distances of the observations from the line.
Linear regression	A statistical method to fit a straight line to data to estimate the relationship between a dependent/outcome variable and independent/predictor variable. In Linear regression we obtain estimates for the intercept and slope (regression coefficients). **Multiple linear regression** is when two or more independent/predictor variables are used to explain a dependent/outcome variable.
Mean	A measure of central tendency. It is computed by summing all data values and dividing by the number of data values summed. If the observations include all the values in a population the average is referred to as a **population mean**. If the values used in the computation only include those from a sample, the result is referred to as a **sample mean**.
Non-linear	Not a straight line or not in a straight line.
Normal distribution	This symmetrical distribution describes how common the values are of many things in nature, at least approximately, e.g. height, weight, blood pressure. It’s also the basis of many statistical tests because, if you know the average value (usually called the mean) and the standard deviation, then you can draw every point of a normal distribution and you know what proportion of values are greater than (or less than) any given point, e.g. the % of men more than two metres tall. Some things are not normally distributed (e.g. proportions of anything, serum concentrations of electrolytes) but can be made to fit quite well after some simple mathematical trickery.

Null hypothesis	The null hypothesis is what the investigator sets out to disprove in order to find evidence of an association between two or more things. The null hypothesis is often that there is no difference between patient groups regarding the outcome of interest. The null hypothesis can then be rejected using statistical tests and their associated p-values
Odds	The odds is a way to express probability, e.g. the odds of exposure is the number of people who have been exposed divided by the number of people who have not been exposed. The mathematical relationship between odds and probability is: Odds = probability / (1 – probability)
Odds ratio	The odds ratio for an exposure measure is the ratio between two odds, e.g. the odds of exposure in the cases divided by the odds of exposure in the controls in a type of study called a case-control study: Odds ratio = Odds of exposure in the diseased group (cases) divided by Odds of exposure in the disease. In the example in the logistic regression course in the statistics for public health specialisation, however, the outcome of interest is diabetes, so we're interested in e.g. the odds of diabetes if you're female divided by the odds of diabetes if you're male
Outcome	This is the event or main quantity of interest in a particular study, e.g. death, contracting a disease, blood pressure.
Overfitting	Overfitting is a phenomenon that occurs when too many variables (with respect to number of observations) are included in a model and the model ends up explaining random error rather than real relationships. This is a problem.
p-value	This is the probability of obtaining the study result (relative risk, odds ratio etc.) or one that's more extreme - if the null hypothesis is true. The smaller the p-value, the easier it is for us to reject the null hypothesis and accept that the result was not just due to chance. A p-value of <0.05 means that there is only a very small chance of obtaining the study result if the null hypothesis is true, and so we would usually reject the null. Such as result is commonly called “statistically significant”. A p-value of >0.05 is usually seen as providing insufficient evidence against the null hypothesis, so we accept the null.
Pearson’s correlation coefficient	A statistic that can be calculated as a measure of the linear association between two continuous variables. It has a value between +1 and −1.
Population	The set of all people of interest to a study. We can't study them directly and so must instead draw a sample of people from the population.
Predictor	Something that goes into a regression model that is potentially associated with the outcome variable. Predictors of death include age and disease. Predictors of disease include age and genes. In this specialisation, we'll often use the word "covariate" to mean the same thing.

R-squared statistic	In linear regression, this is the proportion of the variation in the outcome variable that is explained by the model i.e. by the model's predictors. It can be between 0 and 1. For non-linear regression, versions of the R-squared have been proposed, some more useful than others.
Rate	A rate expresses how quickly the outcome of interest occurs, so is subtly different from a risk (even if many non-epidemiologists use the two words interchangeably). The denominator is some measure of person-time (see the epidemiology specialisation)
Residual	The difference between the observed value of the dependent/outcome variable (y) and the predicted value from the model (ŷ). Each type of regression has its own types of residuals.
Risk	The number of people with the outcome of interest divided by the total number of people at risk of the outcome.
Risk set	In survival analysis, this is the set of patients who are at risk of the outcome of interest.
Sample	A sample is a relatively small number of observations (or patients) from which we try to describe the whole population from which the sample has been taken. Typically, we calculate the mean for the sample and use the confidence interval to describe the range within which we think the population mean lies. This is one of the absolutely key concepts behind all medical research (and much non-medical research too).
Sample population	A subset of the population that can be used in a statistical analysis and for which to draw inference about the 'population'. Choosing the sample is a crucial step.
Sample size	(Usually) The number of people in your sample.
Sampling	The process by which people are selected from the population. To produce unbiased sample statistics the sample needs to be drawn at random from the population i.e. each member of the population should have an equal chance of being selected.
Scatter plot	A graph that plots the coordinates from two sets of data points (two variables). Scatter plots can reveal patterns between the two variables.
Spearman's rank correlation	A statistic that can be calculated to measure the degree of agreement between two rankings for continuous and ordinal variables. It has a value between +1 and −1.
Standard deviation	The average squared difference from the mean, i.e. a measure of "spread"

Standard error	The standard error of a statistic, e.g. the sample mean, is the standard deviation of its sampling distribution. In other words, it's a measure of the accuracy with which the sample represents the population.
Statistic	A numerical measure that describes some property of the population. A statistic is obtained from a sample. We hope the statistic estimated from the sample is statistically equal to the same statistic if we could collect it from the population. If so, the estimate is said to be an unbiased estimate (of the population value)
Statistical test	This is the only way to decide whether the results of your analysis, e.g. your measure for group A compared with your measure for group B, are likely to be due to chance or could be real.
Stepwise regression	A unsatisfactory way to automate the approach to select varaibles for inclusion into a regression model. The weaknesses of the stepwise approach include regression coeffcients being biased high, standard errors being biased low, and are ineffective in the presence of collinearity. These approaches can include 'forward' and 'backward' selection procedures.
t-test	A statistical test for comparing two means of a normally distributed variable.
Variable	A variable is a characteristic or item that can take different values. They can be categorical or numerical variables: for example, disease stage or age.
Variance	The average of the squared differences of the data values from the mean value of observations divided by N observations (or N-1 for sample variance). It's just the square of the standard deviation.