Statistics small glossary	
Term	**Definition**
(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.
(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.
Hypothesis	A statement that can be testes using quantitative evidence (data).
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.
Null Hypothesis	The statement to disprove in order to find associations between two or more variables. E.g. There is no statistically significant difference between two groups of patients.
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.

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 variable/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". Factor which 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.
Interaction	Interaction that occurs when a predictor variable has different effect on the outcome depending on another predictor.
Mean	Measure of central tendency, used to summarise data values
Outcome	This is the event or main quantity of interest in a particular study, e.g. death, contracting a disease, blood pressure.
Predictor	Variable that is included in 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.
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).
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.
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.