How to interpret scientific statistics

Leukaemia Care
YOUR Blood Cancer Charity

How to interpret scientific statistics

Health-related statistics are everywhere; within the news, in adverts, in your hospital appointments. Understanding what these statistics really mean for you, however, can be difficult. In this toolkit, we look at the use of statistics and give advice on interpreting them.

What are statistics?

Statistics are a way of numerically summarising information, based upon evidence from groups of people. They are often used to predict future events or indicate the likelihood of certain events happening.

Incidence

Incidence means the occurrence of disease in a population within a given time.

It is often written as: the number of people who are diagnosed out of a given number of people.

For example, in 2015 the incidence of leukaemia within the UK was 16 in 100,000 .

This means that for every 100,000 people in the UK, 16 were diagnosed with a leukaemia in that year.

What is this as a percentage?

Percentage is the same as saying 'in 100 people'.

For incidence of leukaemia, this would be worked out by:

(16 ÷ 100,000) x 100 = 0.016

So, 0.016% of the UK population were diagnosed with leukaemia in 2015.

Key to remember:

When sums are written in brackets these are worked out first.

e.g. (1+3) – 2 = 4 – 2

= 2

How many people are really affected in the population?

The population of the UK in 2015 was around 65.13 million. Therefore, to work out real numbers of people diagnosed you would do the following:

(16 ÷ 100, 000) x 65,130,000 = 10,421 people diagnosed with leukaemia in the UK in 2015.

Comparing incidence

If you are comparing incidence figures written as '_ in _' it is often worth converting them to percentages or the same population size.

For example, something that affects 3 in 10 people is more common than something that affects 1 in 5 people.

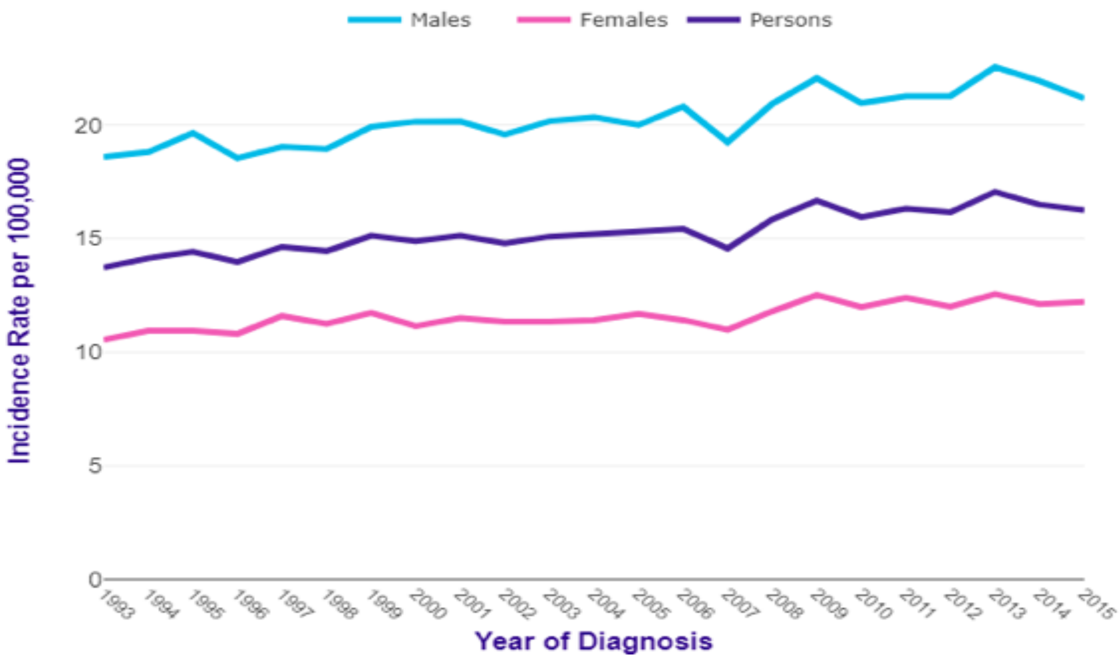
3 in 10: (3 ÷ 10) x 100 = 30%

1 in 5: (1 ÷ 5) x 100 = 20%

Or you could convert '1 in 5' to 'a number in 10' by doing the following:

(1 ÷ 5) x 10 = 2 in 10.

Incidence	Percentage of people affected
1 in 2	50%
1 in 5	20%
1 in 10	10%
1 in 25	4%
1 in 100	1%
1 in 1000	0.1%



Credit: Cancer Research UK

The horizontal axis is the year and the vertical axis is the incidence.

It is important to check the units used, which can be found in the title of axes. In this case, the incidence is per 100,000 people.

There should also be a key that tells you what each of the lines are. Here, it can be found at the top of the graph.

Risk

The incidence of disease within a population in any given time also indicates the chance of (or risk of) someone being diagnosed with a disease.

Based upon the 2015 incidence data, we can assume that the risk of someone being diagnosed with leukaemia in any given year is 16 in 100,000.

For example, if there were 100,000 people in a room, the chances are that 16 of them could be diagnosed with leukaemia in that year.

It is important here to acknowledge the time frame used to deduce risk. For example, the risk of getting leukaemia in your lifetime would not be the same as the risk of developing leukaemia within one year.

It is also important to remember that there are certain risk factors that may alter the chances of getting a disease. For example: age, sex, genetics, and environmental factors, such as smoking.

A leukaemia specific example would be age. The risk of being diagnosed during your 20's would be much lower than your risk of being diagnosed over the age of 65 years old

Percentage differences

Percentage differences are used to show the change between two different data sets.

One example is looking at how incidence of leukaemia has changed between 1993 and 2015. CRUK statistics reveal there has been an 18% increase in total diagnoses . This figure appears quite high, but it is important to distinguish between absolute and relative differences .

The absolute difference is the value of how much something has changed.

Relative difference then expresses the absolute difference as a percentage change from the first value.

For example:

In 1993 the incidence of leukaemia in the UK was: 0.0137%

In 2015 the incidence of leukaemia in the UK was: 0.0163%

The absolute increase in incidence is:

0.0163% - 0.0137% = 0.0026% absolute increase.

To calculate relative increase:

(0.0026 ÷ 0.0137) x 100 = 18% relative increase

Therefore, while an 18% increase in leukaemia incidence over 12 years is significant, the use of relative increases can be slightly misleading, as leukaemia diagnoses are still rare within the total population and the absolute increase is much smaller.

Averages

Averages are a way to identify the mid-point value in a set of data and there are three different types: mean, median or mode.

The mean is calculated as: the sum of the numbers   how many numbers there are.

e.g. If there were 5 people aged 10, 13, 14, 13 and 15 years old, the mean age would be: (10 + 13 + 14 + 13 + 15) ÷ 5 = 13 years old

The median is simply the middle number in a set of values.

e.g. Using the above ages, you can put the values in order and find the middle value: 10, 13, 13, 14, 15 (the median is 13)

If there are an even number of

values in a data set, you must find the mean of the two middle values.

E.g. If there are 6 people aged: 10, 13, 13, 14, 15, and 16 years old the middle numbers are 13 and 14 (10, 13, 13, 14, 15, 16). The median value would be: (13 + 14) ÷ 2 = 13.5

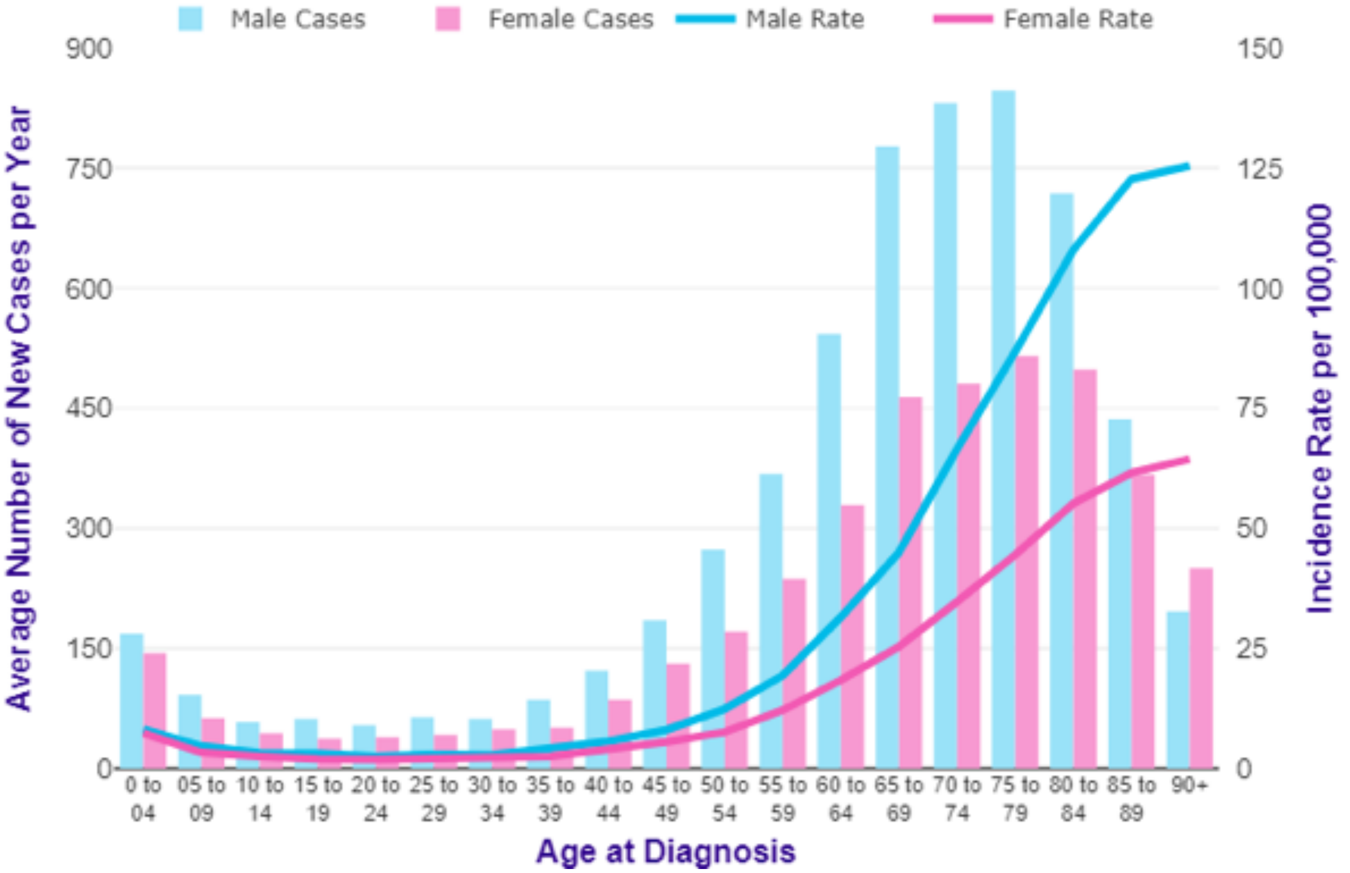
The mode is the value that appears the most in a set of numbers.

E.g. Using the above example of 6 people aged 10, 13, 13, 14, 15 and 16, the mode would be 13 years old.

An average is a good statistic if the range between the biggest value and the smallest value in the data set is relatively narrow, however, if there is a big range the average is often not reflective of the data.

For example, the average age of leukaemia diagnosis is approximately 65 years old . This does not, however, reflect the range of ages whereby people can be diagnosed with leukaemia and the fact that different leukaemia types are more common in different age groups.

More people are diagnosed with leukaemia between the age of 75 to 79 years old than any other age group, but the average age is lower, because there are a significant number of children and young adults diagnosed with leukaemia.



Credit: Cancer Research UK

Sometimes averages are reported with a standard deviation (or SD), which demonstrates how much variation there is from the average. These are formatted as average ± SD (±means 'plus or minus'). The smaller the standard deviation, the narrower the range of data.

Survival/Relapse

Survival and relapse are normally determined by researchers using Kaplan-Meier estimates.

Kaplan-Meier estimates involve working out the probability of an event (e.g. survival or relapse-free) happening at different time intervals . The equation is:

Probability at a time interval =

$$\frac{\text{Number of patients at the start of the time - number of patients that event occured (e.g.died or relapsed)}}{\text{Number of patients at the start of the time}}$$

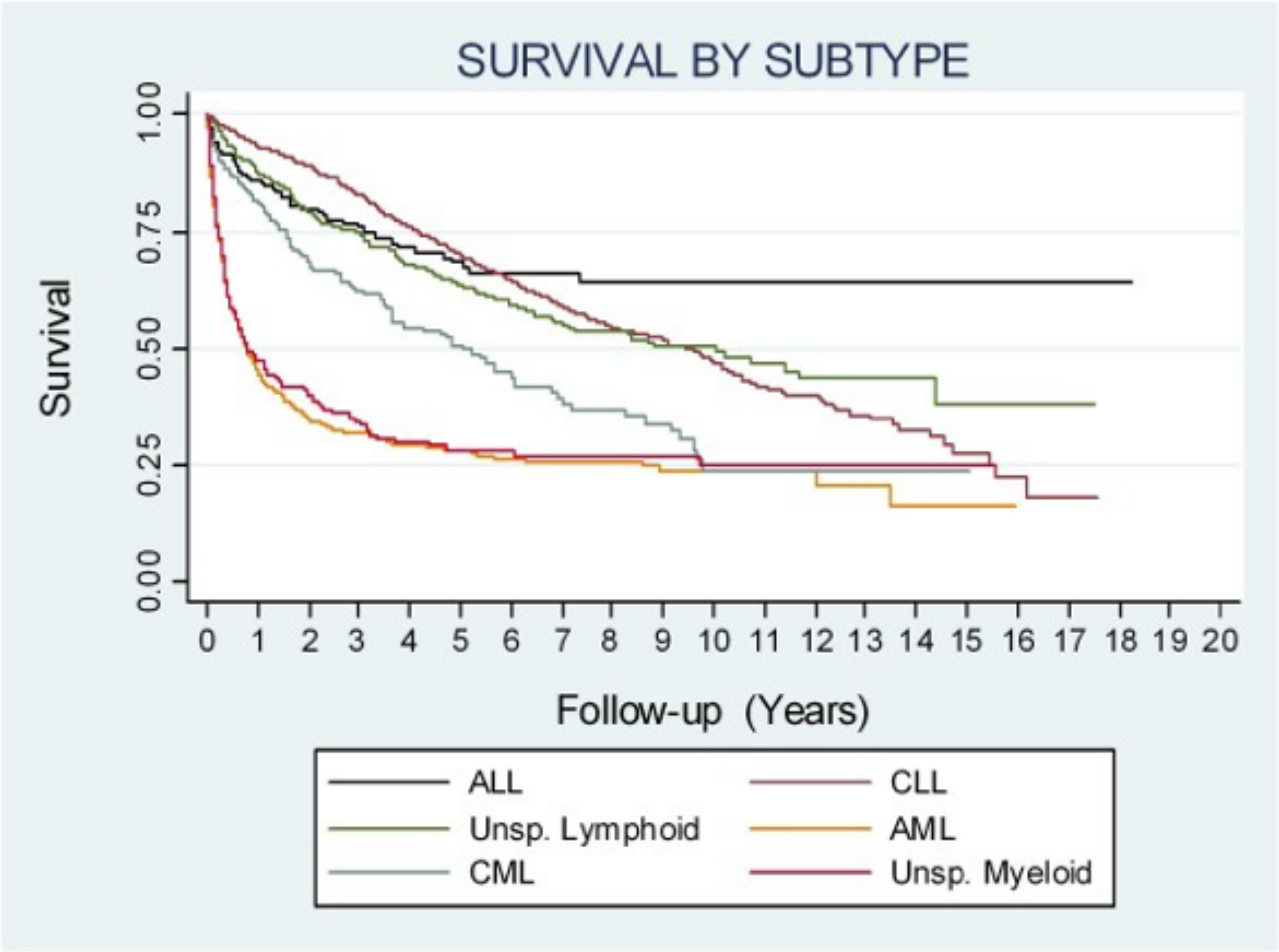
To calculate the probability over time of an event happening, the probabilities from each time interval need to be multiplied.

For example:

Time interval	Probability at interval	Probability at the end of time interval
Day 1	P1	P1
Day 2	P2	P1 X P2
Day 3	P3	P2 X P3

This is because the probability of someone surviving to the end of day 2, for example, would mean someone needs to survive both day 1 and day 2.

The probability over time can then be plotted onto a diagram, known as a Kaplan-meier plot. An example, from a 2009 paper on Leukaemia survival is shown below:



Credit: Bhayat, F. et al. (2009)

Probability of survival is shown on the vertical axis and the time intervals are along the horizontal axis.

Probability is a value between 1 (100% survival) and 0.

The median (or middle) survival is the timepoint where probability is equal to 0.5. This is sometimes highlighted on the plot.

Using the black line on the plot, you can see that at 4 years the survival for Acute Lymphoblastic Leukaemia (ALL) is around 0.75. This means that after 4 years, 75% or 3 in 4 ALL patients were still alive.

There are key things to consider when interpreting statistics:

Using the above research on UK leukaemia survival rates as a case study

1) Where is the data from?

In the above study of survival, the data was taken from a general population data set of patients across 330 GP practices across the UK. Other statistics may be based upon clinical trial data or surveys, for example.

2) How many patients are included?

The study used data from 4162 leukaemia patients. Statistics are more likely to reflect the real-world experience if it is based upon a bigger sample size of patients.

3) What patients are included?

In this paper, patients were included who were marked as having a 'leukaemia diagnosis' by their GP. In this instance, it is grouping together patients who may be very different in terms of staging and risk or type of treatment received, for example. Therefore, the survival statistics are very general.

In other studies, there may be very specific patients included and therefore, results and statistics may not be relevant for all.

4) When was the data taken from?

The paper was published in 2009 and based upon patients diagnosed between 1987 to 2006. While at the time this may have been up-to-date, there have been huge advancements in leukaemia therapies over the past decade, meaning it is unlikely to reflect survival now.

It is key that you always look at when statistics were published and when the data was collected.

5) Are there factors that will alter the statistics?

The Kaplan-Meier survival plot in this paper gives survival for different leukaemia types. The researchers also identified that age, gender and socio-economic background change the survival probabilities, however, these are not shown or referenced within this plot.

Therefore, it is important to acknowledge that statistics are summaries of certain data and do not necessarily represent a full picture.

6) What does other research or statistics say?

Even the best research can have

its weaknesses and it is always important to look at what other evidence is available. In this example, and in most research papers, the authors compare their findings to others that are available.

There are many cases of the media reporting statistics from single studies that do not necessarily reflect the findings of other studies. Therefore, looking at other available statistics and comparing is always important.

Further questions:

If you have any further questions about interpreting scientific statistics then you can contact our Campaigns and Advocacy team. They are available Monday to Friday from 9:00am – 5:30pm. If you would like to speak to them, you can:

- Call our office line on 01905 755977

- Send them an email at **advocacy@leukaemiacare.org.uk**

- You can also call the Helpline, free of charge on 08088 010 444. Available weekdays 9am – 10pm and weekends 9am – 12:30pm.

The team will pass your enquiry onto the Campaigns and Advocacy team.

Please note that our Campaigns and Advocacy team are unable to provide:

- Detailed medical advice or recommendations

- Legal advice

- Advocacy for a course of action which is contrary to the aims and objectives of Leukaemia Care