Statistics: an introduction to basic principles

Philip D Welsby [1], Mark Weatherall [2,3]

[1]Retired, Edinburgh, UK
[2]Department of Medicine, University of Otago, Wellington, New Zealand
[3]Older Adults and Rehabilitation, Capital and Coast District Health Board, Wellington, New Zealand

Correspondence to
Dr Philip D Welsby, Retired, Edinburgh, UK;
philipwelsby@aol.com

Received 25 February 2021
Revised 15 March 2021
Accepted 18 March 2021
Published Online First 26 May 2021

Check for updates

© Author(s) (or their employer(s)) 2022. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Welsby PD, Weatherall M. *Postgrad Med J* 2022;**98**:793–798.

ABSTRACT

The essential principles of statistics as applied to surveys, studies, sampling, epidemiology, screening, and trials are described and explained.

"Lies, damn lies, and statistics"
"When in doubt tell the Truth"
Mark Twain

STATISTICS: A SIMPLIFIED MATHEMATICAL BACKGROUND

Statistics uses systematic collection and analysis of numerical values to assist conclusions about whole populations when details of whole populations are vague or incomplete. Both deterministic and stochastic measurements underlie observed measurements. *Deterministic* refers to (1) specific 'hard' numerical measurements of whole population or from representative samples or (2) derived from inductive reasoning (inferring general results from particular instances) or (3) derived from deductive reasoning (inferring particular results from general instances). *Stochastic* refers to 'fuzzy' inputs in which the actual value of individual measurements is unnecessary but, importantly, from which meaningful patterns can be derived by statistical methods.

Definitions of commonly used statistical terms

The *mean*, 'the average', is the most commonly used statistic in everyday life. It is the sum of a set of continuous numerical measurements divided by the total number of measurements. It is intuitively reassuring, although possibly fallible when inappropriately used, to identify centre of the distribution for a population, or what constitutes a typical value of a population.

The *median* is the middle observation in a ranked series of measurements, although its exact derivation can differ regarding the processing of two 'middle' values when an even number of measurements are made.

The *sensitivity* represents the proportion of positives that are correctly identified (figure 1).

The *specificity* reflects the proportion of negatives that are correctly identified as being negative, leaving most of the remaining proportion being positive results (some may be false positives) (figure 1).

A *sample* is a selection of individual measurements from a population.

Bias is conscious or unconscious favouritism in collection, analysis or interpretation of results.

The *range* is the difference between the maximum and minimum observations.

Efficacy is the outcome of an intervention *in a controlled setting*.

Effectiveness is the outcome *in the setting for which an intervention was intended*.

A *null hypothesis* usually reflects a statement of a value or summary of values that would typically be present and not the result of any intervention. Two terms are often used in relation to null hypotheses. *Type 1 error* is rejection of a true null hypothesis—a 'false positive'. *Type II error* is non-rejection of a false null hypothesis—a 'false negative'.

Hypothesis testing tests the validity of a claim that is made, usually against a null hypothesis.

P values quantify the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. P values of ≤0.05 are taken to mean that the observed results may be due to chance in less than 1 in 20 instances.

The SD in relation to the *standard normal distribution curve* (figure 2) illustrates the amount of variation among numbers in a dataset and enables quantification and appreciation of distances of individual measurements from the mean. A symmetrical appearance—the bell-shaped curve—reveals a smooth distribution of whatever is being measured. In a normal distribution, the peak is both the mean and the median of measurements. Measurements constituting a normal distribution curve occurs in many but not all biological numerical assessments, and enables succinct quantification of scatter, dispersion or variability. If distributions are not bell shaped, the distribution(s) may be skewed or contain two or more peaks. If measurements are made from a normal distribution, then individual measurements can be put in context of the population from which the numerical results were acquired. For example, a student's biochemistry mark may be 1.25 SD above his class mean but his physiology mark may be 1.1 SD below his class mean. This compares his performance with the class no matter what the numerical scores were in the separate marking systems. In a normal distribution, 68% of measurements occur within plus or minus 1 SD either side of the mean (that is a total of 2 SD) and 95% occur within 2 SD from the mean, and almost all observations occur within 3 SD from the mean. Results without 2 SD are considered to be unusual and might require explanation or investigation. Numerical laboratory results often report their assessment of the reference range with respect to an observed result as being between 2 SD above or below the mean.

Unawareness of statistical methods may lead to false conclusions. "I had a patient like this who responded to X, therefore most patients should respond"—an example of inductive wishful

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.

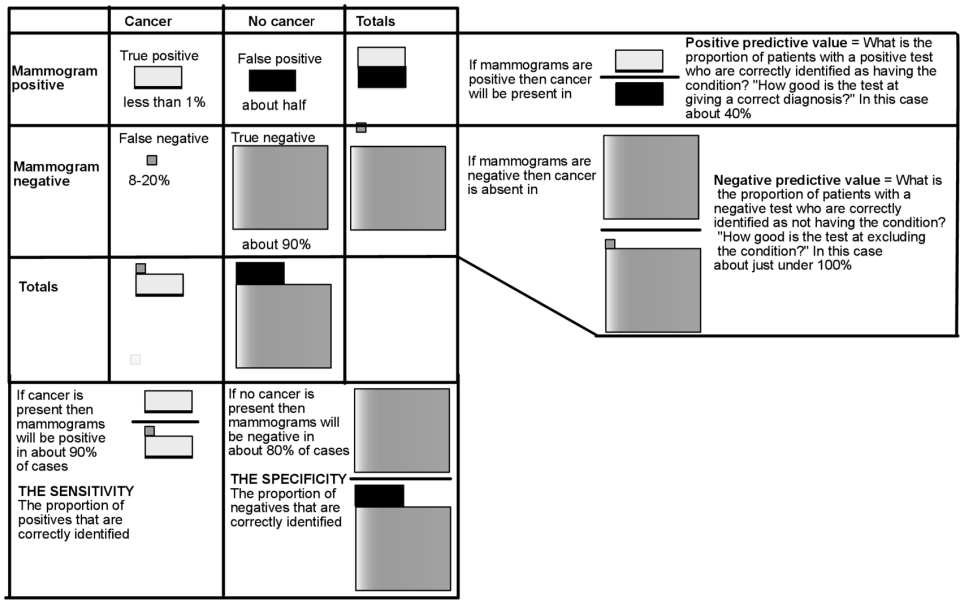

Figure 1

thinking (observation of one example can be applied in many situations) or "Most of my patients have responded to X therefore this patient will respond to X"—an example of deductive wishful thinking (observations of many situations can be applied to one situation).

Use of average results often imparts little information. For example, [2,2,2,2,2,8,8,8,8,8], [4,4,4,4,4,6,6,6,6,6] and [1,2,3,4,5,6,7,8,9,10] are hugely different patterns despite all having identical averages of 5.

Errors may occur if results from a distribution conceals one or more subgroups. For example, heights of all adults conform to a normal distribution but this conceals the merging of two separate peaks (male and female), each of a normal distribution with different means and possibly different SD. Measurements of means and SD alone will not reveal two such peaks. Also other distributions (such as the distances of the right and left ear from the left ear) are 'numerically non-standard' because they have two peaks.

Methods for statistical analysis for categorical variables are often based on the binomial distribution and include χ^2 tests

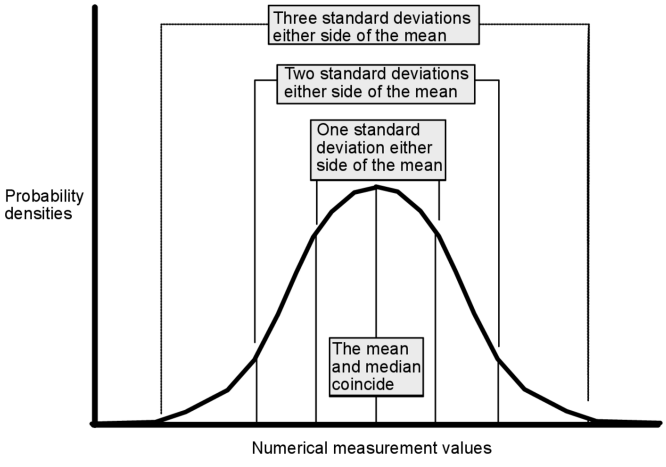

Figure 2 The standard 'bell-shaped' distribution curve. A standard normal distribution is a normal distribution with a mean of zero and an SD of one.

but the use of SD for interpretation of distance is not as applicable, and analysis and interpretation can be more complex and are too sophisticated for this account. The implications of this statement are clear—unless you have statistical experience you should consult a statistician *before* commencing any non-simple study to understand a reasonable approach to data collection and analysis.

Sampling and surveys

Appropriate sampling enables reasonable statistical generalisations about whole populations to be made. Generalisations obtained from samples can almost never be certain, but the degree of uncertainty can be quantified by using statistical methods. Remarkable conclusions can be drawn from surprisingly small samples provided the procedures of sampling were robust and randomisation was used. Random does not mean haphazard.

Sampling requires questions to be asked of the sample members. Sampling may not be robust as illustrated by partisan pre-election polls in the USA, some of which used landline phone calls (landline users do not reflect the general population) and representative samples should be likely to include all elements of the population.

Questions are asked in surveys but individual questions and questionnaires may be flawed. Those who complete them may be a self-selected group and especially with electronic surveys the number of those approached may be unknown, and thus findings derived from the sample may not be representative no matter what statistics are used on the sample. For generalisations to be made from such ill-defined populations, the authors have to demonstrate that, in retrospect, their sample did reflect the general population, by having similar distribution of important relevant factors. Poorly formulated questions or even fatigue caused by over long questionnaires may invalidate conclusions.

Some surveys are obviously intended to deceive. Statements that '90 percent of people prefer product X' are inappropriate when samples are less than 100.

Survey purposes should be explicit, the target population should be explicit, the sampling procedures should be explicit, the questions should be ruthlessly focused, the follow-up

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.

The UK Office for National Statistics produced a new socio-economic classification in 2001

1 Higher professional and managerial occupations

2 Lower managerial and professional occupations

3 Intermediate occupations

4 Small employers and own account workers

5 Lower supervisory and technical occupations

6 Semi-routine occupations

7 Routine occupations

8 Never worked and long-term unemployed

The previous classification (not uncommonly still used)
1. Professional
II. Semi-profesional
III . Skilled. Non-manuallly IIIN
Manually IIIM
IV. Semi-skilled
V. Unskilled

Figure 4 Social class classifications.

procedures should be explicit and appropriate, the results should be capable of analysis, and conclusions should be justifiable. 'Yes' or 'No' questions should not evoke an 'It all depends' responses. Graded answers 'on a scale of 1–10' may be required.

Beware percentages! Some (that is ill-defined) researchers report results as 50% reduction irrespective of the baseline proportions. A 50% reduction is more impressive when 10 in a 100 respond to treatment X compared with 20 in a 100 who do not than when this reduction was from two in a million to one in a million.

EPIDEMIOLOGY

Clinical medicine is mostly concerned with *individuals* who have a disease, whereas epidemiology is the study of the incidence, distribution and determinants of diseases in human *populations* with various conditions and uses questions 'Which groups?', 'What?', 'Why?' and'"What might be done?'.

Epidemiological studies may deal with individuals who have a specific disease, individuals who do *not* have the disease, changes in frequency of specific diseases, patterns of disease, incidence rates of specific disease or prevalence rates of specific diseases.

Epidemiological description involves reporting the variation of disease and the putative cause, the geographical or situational

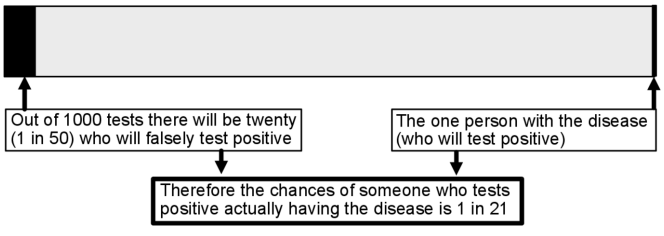

Figure 3 The limited usefulness of test with a low false-positive rate (1 in 50) in screening for an uncommon disease (1 in 1000).

EVENTS	RESPONSES
SYMPTOM A **AND** DISEASE B	"I have observed that this symptom and disease are often associated. They must be correlated"
SYMPTOM A **BUT NOT** DISEASE B	"I did not realise the some occurrences of symptom A occurred in the absence of disease B were not with event B until I looked"
NOT SYMPTOM A **BUT** DISEASE B	There are occurrences of the disease that did not have the symptom. Perhaps the causal relationship is not as strong as I thought"
NOT SYMPTOM A **AND NOT** DISEASE B	I cannot see the need to discover the occurrence of these two risk events"

It is necessary to know this so that the total number of absences of symptom A can be known. For example if the results were

	Disease B present	Disease B absent	TOTALS
SYMPTOM A	80	20	100
SYMPTOM A ABSENT	40	10	50

It is necessary to know the totals so that it can be realised that the same proportion (80/100 = 80 percent) who have the symptom have the disease as do those who do not have the symptom (40/50 = 80 percent). The symptom A is therefore **not** a marker of the disease, even though it is twice as likely to occur when disease B is present

Figure 5 The need to consider all combination of symptoms and diseases.

variation, variation of putative causes variation of the disease or variation in the disease in those affected.

Descriptive epidemiological studies are studies of the variation in incidence of a disease according to time, place, or person and may suggest causes of disease, enable quantification of the health problem or enable quantification of the financial implications. *Analytical* studies are designed to quantify the risk of disease associated with a putative risk factor.

Epidemiological data can be obtained by many means including surveys, morbidity data, mortality data, registers of disease prevalence and incidence, general practice or hospital attendance, or inspection of centrally held records.

EPIDEMIOLOGICAL SCREENING

Investigatory epidemiology usually involves screening. Ideally, screening tests should be highly selective and highly sensitive. Problems arise when screening large populations for symptoms of disease even when false-positive tests are uncommon (figure 3). Screening for common diseases (using screening for breast cancer with mammograms as an example) may cause many women to be subjected to further investigations when they are unlikely to have breast cancer (figure 1).

The major task of epidemiology is the accurate identification of the causes of what is observed. There may be several candidate causes and several potential underlying cofactors. For example, the effect of social class (figure 4) on many conditions may be obvious, or hidden or controversial. For example, are the higher mortality rates in social class 8 *associated* with unemployment, but are they *caused* by unemployment? Simple questions may require complex answers.

Identified symptoms should not be associated with diseases without appreciation that it is necessary to assess all possible combinations of symptoms, no symptoms, disease and no disease (figure 5).

Definitions relevant to epidemiological inquiry

The *incidence is* the number of new cases of a disease in a population occurring in a defined time, in other words: 'How many

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.

	Prevalence high	Prevalence low
Incidence high	Common persisting conditions eg. Osteo-arthritis in the elderly	Commonly occurring brief infections eg. the common cold
Incidence low	Persisting diseases that are uncommon eg rheumatoid arthritis	Rare brief diseases eg.malaria in the UK

Figure 6 The difference between incidence and prevalence.

new cases?' The *incidence rate* is the incidence divided by the total population.

The *prevalence* is the frequency of the disease at a specified point in time in a defined population, in other words: 'How many cases are there?' (figure 6). The *prevalence rate* is the prevalence divided by the total population.

Sensitivity and specificity have been defined previously (figure 1).

Life expectancy is the mean number of years that individuals drawn from a specified population can expect to live.

Relative risk is the amount of disease that a putative factor might cause is the incidence among those exposed to the putative factor **divided** by the incidence among those not exposed to the putative risk factor. *Absolute excess risk* is the incidence among those exposed to the putative factor **minus** the incidence among those not exposed to the putative risk factor (figure 7).

A *standardised mortality rate* is the expected number of individuals with the condition compared with the actual number. These rates should reflect whole populations and subsections of populations (age often affects disease) and it would be important to find the standardised mortality rates in those of various ages.

Prospective studies

Prospective studies (also known as longitudinal or cohort studies) entail following up individuals who are exposed to a putative risk factor and discovering how many develop the condition in question.

Prospective studies depend on knowing that exposure to the putative risk factor antedates the disease, knowing that there should be accurate observations 'looking for the condition *at the time* rather than relying on retrospective adequacy of notes made previously' and appreciating that unexpected associations may become apparent if all possible relevant observations were made.

Prospective studies may take many years to produce results—and thus are often not career-enhancing for those in training and may require a large number of observations (especially if the disease in question is rare).

	Disease	No disease	Rates of disease
Exposed to putative cause	a	b	$\frac{a}{a+b}$
Not exposed to putative cause	c	d	$\frac{c}{c+d}$
The *relative* risk is	$\frac{a}{a+b}$	**divided by**	$\frac{c}{c+d}$
The *absolute excess* risk is	$\frac{a}{a+b}$	**minus**	$\frac{c}{c+d}$

Figure 7 The risks of disease.

Retrospective studies

Retrospective studies have to be assessed with caution as the condition selection procedures at the time may not be applicable to current circumstances. Relative risks often cannot be identified because the proportions of the population from which each group have been drawn are usually not known.

The advantages of retrospective studies include short completion times, cheapness and that they are useful for rare diseases (waiting for rare diseases to occur can be very boring). Disadvantages might include uncertainty as to whether the putative risk factor preceded the condition (as it should), reliance in part on the individual memories of events or notes of events made at the time. Identification of relevant individuals are often unfocused because of failure to look for relevant risk factors of the disease *at the time* and no straightforward assessment of the excess risk is possible (although indirect assessments for rare diseases may be possible).

Intervention studies

Interventional epidemiological studies assess the effect of planned interventions on disease patterns and share many techniques with trials.

PREVENTION

There are four types of prevention using epidemiological principles. *Primary prevention* is the prevention of future occurrences in well individuals by removing causes that may include environmental, economic, social, educational, and dietary factors, or increasing resistance to disease by vaccination. *Secondary prevention* is the prevention of clinical disease before symptoms appear by screening, early detection and/or treatment of conditions such as hypertension. *Tertiary prevention* is the reduction of future harm by treating symptomatic diseases. *Quaternary preventions* are attempts to mitigate or avoid unnecessary or excessive therapeutic interventions that are in fact harmful.

For an association between a putative cause of a disease to be demonstrated, it is important that

- There is considerable overlap of the putative cause distribution and the disease distribution.
- Exposure to the putative cause preceded the disease.
- Altering the putative cause affects the disease.
- Removal or reduction of the putative cause results in the reduction of the disease.
- There is a dose–response relationship between the amount of the putative cause and the amount of the disease.
- The geographical or other distribution of the disease varies with the geographical distribution of the putative cause.
- A similar population could be identified who have similar patterns of the disease and observing that the same putative cause is present and the incidence of the disease is similar.
- The association is a constant finding in several studies.
- There are no obvious confounding factors (eg, alcohol is associated with lung cancer but is mostly explained by the fact that those who drink a lot also tend to smoke a lot).

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.

- Biases have been excluded or known to be insignificant.
- The association is unlikely to be caused by chance.
- Other possible risk factors appeared to be minimal.
- Laboratory evidence supports the hypothesis that the association is causal.
- There are no other explanations.

TRIALS

Trials usually are often initiated because observations suggest possible benefits of a particular intervention, after which a hypothesis is stated, namely, that an intervention has no effect, but which can then be investigated.

At their best, randomised trials allow rigorous valid conclusions to be drawn about the effects of treatment in a specified sample of participants that can then be generalised to the population from which the participants were gathered. Trials are costly, both financially and in the time required, such that they tend to be of brief duration even for potential long-term treatments.

Research participants may not be representative of the general population and intensive observations needed for trials may not occur in real-life clinical practice. A high degree of certainty that an association between treatment and outcome is causal requires meticulously designed and executed randomised controlled trials. With such trials, patients either receive the treatment or not, typically 50% are allocated to each arm of a two-armed trial, as this is usually the optimum allocation for a two-armed trial with respect to the ability to detect a difference (the power) between two samples should one actually exist.

Some trials may have potentially misleading outcome measures, including substitution of surrogate markers such as laboratory results for clinical outcomes. Does reduction in serum cholesterol translate into overall clinical benefits?

Trials are classified as *open-label* (after randomisation everyone knows who is receiving what). In *single blind* trials, individual participants do now know which arm of the trial treatment they may be receiving but the trial organisers are aware. In *double blind* trials, neither participants nor the organisers know what treatment (if any) the participants are receiving. In *crossover*

Current research questions

⇒ Identify important misuses of statistics in journals or conferences, in particular Type II error together with poor development of scientifically or clinically important differences to detect.
⇒ Identify failure of understanding of early post-graduate health practitioner's interpretation of the difference between association and causation and possible remedial education.
⇒ Identify failure of understanding of bias from sample surveys in health practitioners and possible remedial education

Key references

1. Armitage P, Berry G, Matthews *JNS. Statistical methods in medical research* 4th edition. J Wiley and Sons 2001.
2. Bland M. *An introduction to medical statistics* 4th edition. Oxford University Press 2015.
3. Campbell MJ, Machin J, Walters SJ. *Medical statistics – a textbook for the health sciences* 4th edition. J Wiley and Sons 2007.
4. Campbell MJ, Swinscow TDV. *Statistics at square one* 11th edition. Wiley-Blackwell BMJ Books 2011.
5. Woodward M. *Epidemiology study design and data analysis* 3rd edition. Chapman and Hall/CRC texts in statistical science 2013.

Multiple choice questions

1. The mean of a statistical distribution is best described as?
   A. The numeric average of that distribution
   B. The expected value of the distribution
   C. The central location of the distribution
   D. The middle value of the distribution
   E. The central percentile of the distribution
2. For a sample survey which one of the following is most correct?
   A. The width of the confidence interval for estimation of a proportion is determined by the size of the population
   B. Non-response bias rarely leads to problems with estimation
   C. Most sample surveys carried out using land-lines are unbiased
   D. Requests for personal information should generally be asked for first in sample surveys
   E. Bias refers to the difference between the population parameter and its estimated value in the sample
3. In frequentist-based analyses which statement best describes the P-value for a hypothesis test?
   A. The probability of a result at or more extreme than the observed results, assuming the Null hypothesis is true
   B. The probability the association is significant
   C. The post-hoc probability after accounting for the prior probability
   D. That this is the width of the appropriate confidence interval
   E. The probability that a Zed statistic is greater than 1.96
4. In a randomised controlled trial why is cause and effect inference more robust than in a non-experimental cohort study
   A. Because cause precedes effect
   B. Because the clinical sample is more likely to be representative of the target population
   C. Because randomisation ensures extraneous factors that may influence outcomes are randomised between treatment groups
   D. Because two-sample tests are more powerful than one-sample tests
   E. Because baseline measurements can be taken into account
5. For screening tests for disease in a population which one of the following statements is most correct
   A. A test with a high specificity means there will be few false positives
   B. The utility of the test only depends on the sensitivity and specificity
   C. The most important element of a screening test is always its sensitivity so that there are few false negatives
   D. If a disease is very rare then a low specificity for a screening test is not very important for the test to be useful
   E. If the prevalence of a disease in a population is 1 in a 100 (1%) then a test with a sensitivity of 100% and a specific of 90% will identify more false positives than true positives

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.

trials, patients are randomly started on one treatment or placebo and subsequently changed to the other arm treatment usually with a 'washout' period in between. *Placebo-controlled* trials are usually double blind (although drug effects may reveal that some participants are receiving an active drug). Placebo-controlled trials are often used when evidence of effect of new treatments is required to satisfy regulatory authorities. Not surprisingly clinicians favour comparative trials 'Which treatment is better', whereas drug companies favour placebo-controlled trials 'Is our drug effective?'.

There are several ways in which control groups may be found. *Historical studies* in which patients receiving new treatments are compared with those in whom this treatment was not given or was not available. *Geographical studies* in which comparisons are made with similar patients in different locations who had not received the treatment. It is important that studies and trials are independently and continually scrutinised such that trials are prematurely terminated if significant positive or negative results occur.

Contributors We are the sole contributors.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

ORCID iDs
Philip D Welsby http://orcid.org/0000-0002-0369-068X
Mark Weatherall http://orcid.org/0000-0002-0051-9107

Answers

1. B
2. E
3. A
4. C
5. A

Postgrad Med J: first published as 10.1136/postgradmedj-2020-139446 on 26 May 2021. Downloaded from http://pmj.bmj.com/ on October 1, 2022 by guest. Protected by copyright.