Self-screening tool for MPhil in Population Health Sciences candidates

Please attempt all questions (except in the "Coding" section, where you should do one of the tasks). Write down your answers, and note how you reached them, before looking at the answer key and the suggested resources.

Units and Orders of Magnitude

Population health sciences can involve very large numbers (and very small probabilities). It can involve comparing data across multiple datasets, which might not be displaying numbers in the same way, or even using the same units.

1. A respiratory mask filters out particles with diameter larger than 0.3×10^{-6}m. Which of the following hazardous substances would the mask filter out?
    a) Asbestos fibres with a diameter of 2.5×10^{-4}m
    b) Diesel exhaust carbon particles with a diameter of 1.2×10^{-5} cm
    c) Benzene molecules with a diameter of 6.2×10^{-10} m

Correct Answer: (a).
The mask filters out particles with diameter larger than 0.3 x 10-6m. 0.3 x 10-6m = 3 x 10-7m. Hence, a respiratory mask will filter out the asbestos fibres, which have diameters greater than 3 x 10-7m. Diesel exhaust carbon particles and benzene molecules have diameters less than 3 x 10-7m and so will not be filtered out.
Answering this question requires understanding standard form and orders of magnitude, and ensuring that comparison is made using the same units for each item.
Common errors: You may have done the conversion from 0.3x to 3x incorrectly; or you might not have noticed that the diesel particles are given in cm, and thus concluded they are large enough to be filtered out. For postgraduate study, attention to detail is important.
Suggested Resources:
Many resources are available, including:
https://www.bbc.co.uk/bitesize/guides/zxsv97h/revision/1
https://isaacnewtonacademy.org/sites/default/files/Year%2010%20Combined%20Science%20-%20Lesson%201.pdf

2. On average a 70kg man has 1.35×10^{5}kcal of potential energy stored in fat, 1.2×10^{3}kcal of potential energy stored in glycogen and 2.4×10^{4}kcal of potential energy available from metabolising protein. What is the total potential energy stored in these sources? Express your answer in units of kcal using standard form to 4 significant figures.

Correct Answer: 1.602 x 105 kcal
Key Points: This is a straightforward summation, which requires attention to orders of magnitude. You need to make sure the amounts of energy are in the same format, and then add them up.

Common errors: Some students assume something more complicated than adding three numbers is required, or fail to ensure all three numbers are expressed in the same units (e.g. 105 kcal) before adding.
Suggested Resources: as for question 1

Programming – Coding

As part of the MPhil you will be doing statistical programming/ data science in the R programming language. We don't assume any programming knowledge, however, it is useful to be familiar with some of the basics of how code is constructed.

3. If you don't have experience with programming, please complete Part 2 (Also called "Normal" on the app) of the RodoCodo Hour of Code game:

https://game.rodocodo.com/hour-of-code/

You will see that this programming game is aimed at children. Please don't be put off by this. If you didn't learn basic programming as a young child, it helps to go through these fundamental ways of thinking, in much the same way as you would do if you were trying to learn a foreign language as an adult.

If you have experience with programming, please do this question instead:

3 (alternative)
In the programming language of your choice:
   a. Write code which will print the statement "Hello World!"
   b. Write some code which will load a dataset saved in csv format
   c. Write some code which uses a for loop (or otherwise performs a task which would typically be done with a for loop).
   d. Write some code which uses an if-else statement (or otherwise performs a task which would typically be done with an if-else statement).
   e. Write some code which defines a function that takes as input two numbers and returns their sum

For instance, in R, I would write:

```
print("Hello World!")

read.csv($File_Path)

for (ii in 1:10){
        print(ii)
}

if (p<0.05){
        print("Significant")
} else{
```

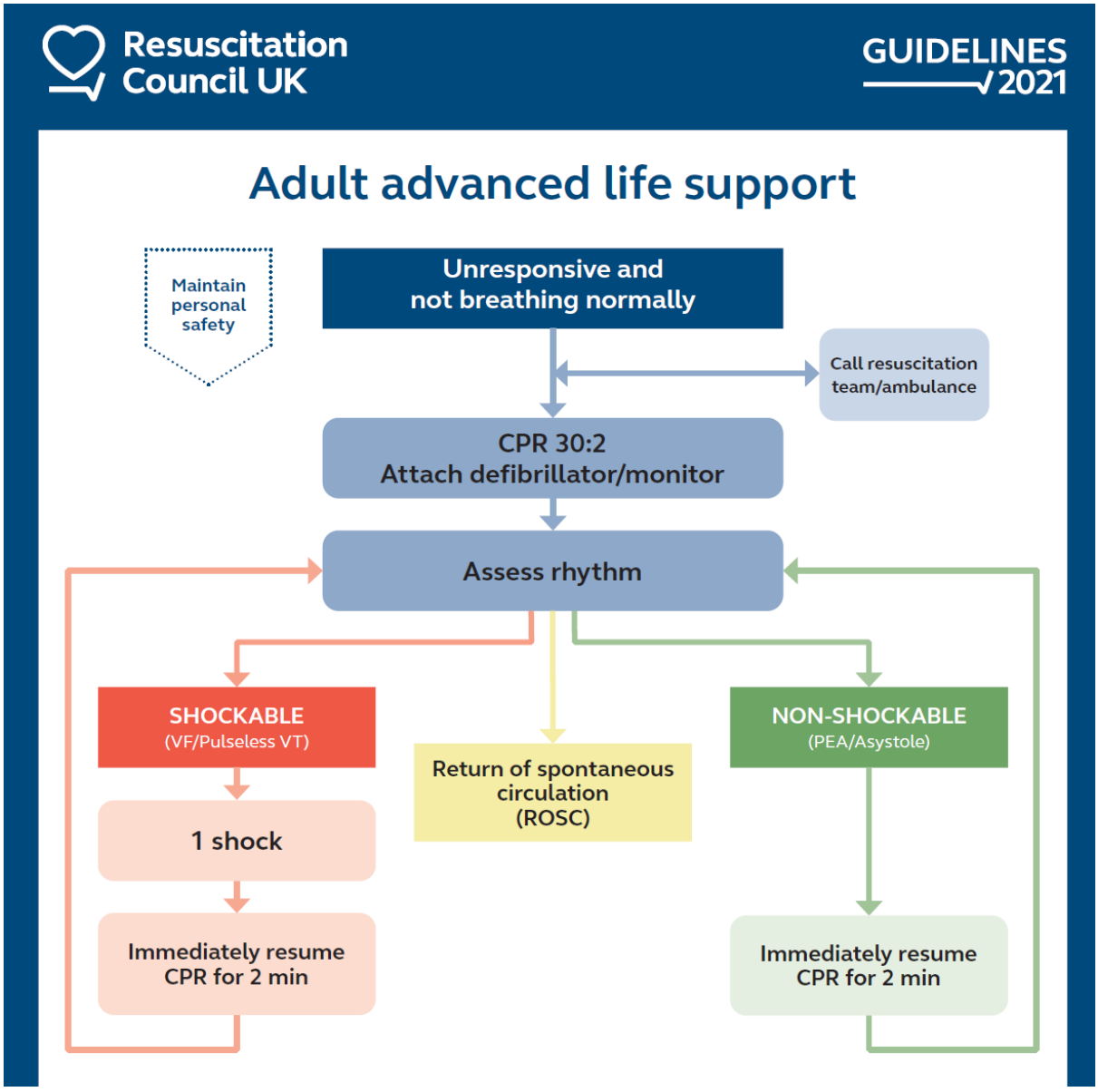
```
        Print(" Not Significant")
}

sumfun<-function(a,b){return(a+b)}
```

Programming – Algorithmic Thinking

As part of the MPhil you will be doing statistical programming/ data science in the R programming language. We don't assume any programming knowledge, however, it is useful to be familiar with the kind of logical thinking behind writing algorithms.

However, algorithmic thinking is not just used in programming. You will have encountered many algorithms in your day to day work and life. For instance, below is the Resuscitation Council UK's 2021 algorithm for adult advanced life support:



4. Think of an algorithm you have used in your work/ study. Write it down as a flow chart.

5. You and two of your friends are at a coffee shop. You have decided you will pick who pays for all the coffees at random. However, all you have is a fair (two sided!) coin.
    a. Come up with an algorithm which picks one of the three of you at random, with equal probability, and write it down as a flow chart.
    b. Can you extend your algorithm to when there are more than three of you?

The algorithm I would suggest is:

1. *Flip the coin twice*
2. *If the coin lands:*
    - *HH: You pay*
    - *HT: Friend 1 pays*
    - *TH: Friend 2 pays*
    - *TT: Flip the coin twice again*

A variation of this can be done for any number of people – flip the coin enough times that a unique sequence of heads and tails can be assigned to each person. If the coin flips make one of these sequences, the corresponding person pays. Otherwise, flip again.

6. The life support algorithm above has a loop – the command "Immediately resume CPR for 2 min" is repeated endlessly (or until ROSC is achieved). Come up with an algorithm you have used in work/study/life which has a loop, and write it down as a flow chart.

Algebra of Regression

One of the most common statistical tools used, particularly when we wish to control for the effect of confounding variables or make predictions, is regression. The Principles of Biostatistics module will cover linear regression and logistic regression. In order to interpret the output of these models, you will need to be able to comfortable with some algebra.

Note: for all of these questions, log means natural logarithm, base e; exp(x) means e^x

7. Write each of the following expressions as a single logarithm:
    a. log(a) + log(b)
    b. log(a) – log(b)

Correct Answers: (a) log(ab) (b) log(a/b)
Think about how to add and subtract logarithms, and fundamental properties of logarithms.

Suggested Resources:
Basic properties of logarithms:
http://www.mathcentre.ac.uk/resources/uploaded/mc-bus-loglaws-2009-1.pdf
or https://www.expii.com/t/intro-to-adding-and-subtracting-logs-same-base-4431#:~:text=Logs%20(Same%20Base)-,Logs%20of%20the%20same%20base%20can%20be%20added%20together%20by,)%20%2D%20log(y).

8. Write each of the following expressions as a single exponential term:
    a. exp(a) x exp(b)

b. exp(a) / exp(b)

Correct Answers: (a) exp (a + b) (b) exp (a - b)
Common errors: Multiplying and dividing rather than adding and subtracting to get exp(ab) and exp(a/b).

Suggested Resources:
https://courses.washington.edu/b513/handouts/LogExpfunctions.pdf
or http://www.pitt.edu/~mgahagan/Exponent.htm

9. Assuming x and y are real numbers, rearrange and simplify the following equations to express X as a function of Y.
    a. Y=log(X)
    b. Y =log(X^a)
    c. $Y=X^2-3$
    d. $Y= 10^{-4} \times X$
    e. Y =exp(2X+6)
    f. Y=(exp(X))/(1+exp(X))

Correct Answers:
a. X=exp(Y)
b. $X=(exp(Y))^{(1/a)} = exp(Y/a)$
c. X=+/-sqrt(Y+3)
d. $X=Yx10^4$
e. (log(Y)-6)/2
f. X=log(Y/(1-Y))

For these questions, you need an understanding of logarithms, exponents, and how to rearrange equations, particularly from logarithmic to exponential form and vice versa.
Common errors:
Not realising you can bring the power of 1/a inside the exponential in (b).
Not getting both square roots in (c).
Not noticing the negative exponent in (d).
Bringing terms into the logarithm inappropriately in (e).
Suggested Resources:
Some fundamentals of using logarithms and exponents:
https://revisionmaths.com/advanced-level-maths-revision/pure-maths/calculus/exponentials-and-logarithms

10. The graphs above show the relationships between various biomarkers (blood chemicals) labelled Y1-Y5 and age, amongst male and female children aged 0-10 years. Let *female* represent a binary variable which equals 0 for males, and 1 for females and *age* represent a continuous variable for age in years. Which of the following equations

represent the lines drawn for the biomarkers labelled Y1, Y2, Y3, Y4 and Y5 in the figures above?

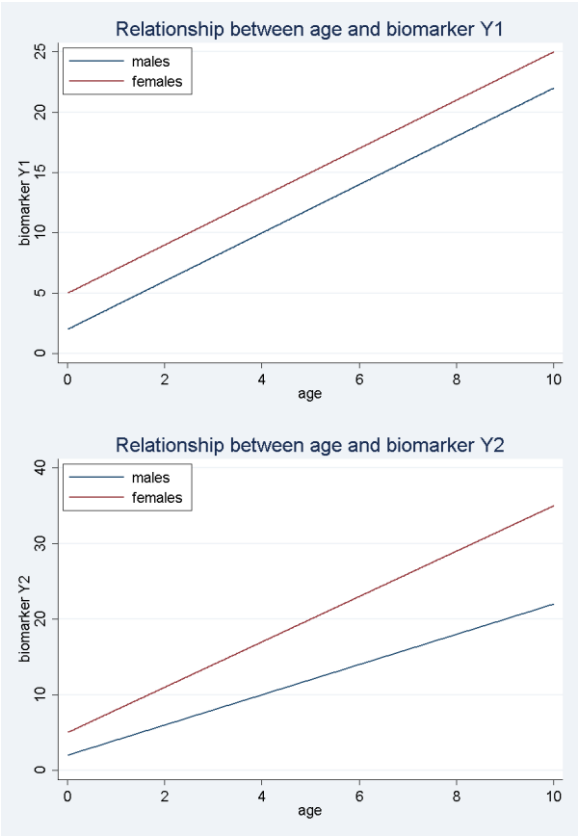
a. Equation 1: $Y = 2 + 2 \times age + age^2 + 3 \times female + age \times female + age^2 \times female$
b. Equation 2: $Y = 2 + \log(age) + 3 \times female$
c. Equation 3: $Y = 2 + 2 \times age + 3 \times female + age \times female$
d. Equation 4: $Y = 2 + 2 \times age + age^2 + 3 \times female$
e. Equation 5: $Y = 2 + 2 \times age + 3 \times female$

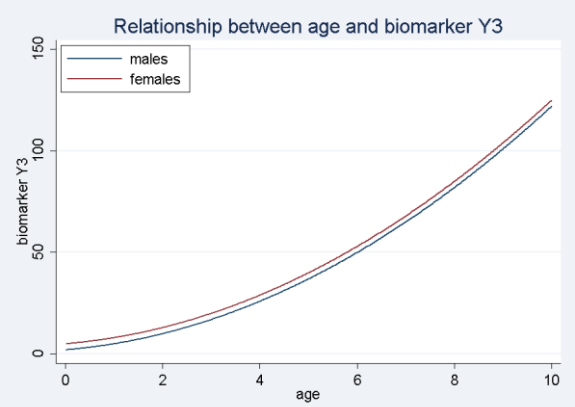
Y1 is equation 5, $Y = 2 + 2\times age + 3\times female$
Y2 is equation 3, $Y = 2 + 2\times age + 3\times female + age\times female$
Y3 is equation 4, $Y = 2 + 2\times age + (age)^2 + 3\times female$
Y4 is equation 1, $Y = 2 + 2\times age + (age)^2 + 3\times female + age\times female + (age)^2\times female$
Y5 is equation 2, $Y = 2 + \log(age) + 3\times female$

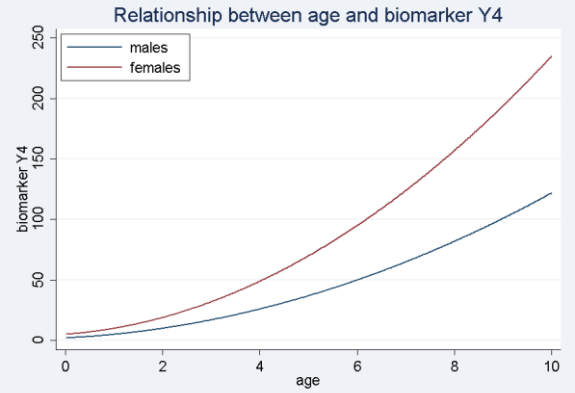
One way to approach this is to substitute the variable 'female' with its values, 0 and 1, in turn, each time collecting together terms of the same variables and seeing what the equation then looks like. Pay attention to whether the equation appears to be linear, quadratic, or logarithmic. If you are unfamiliar with the shape of linear (e.g. y=x, y=mx+c), quadratic (e.g. y=x2 or y = ax2 +bx + c), or logarithmic graphs (e.g. y=log(x)), it will be helpful to review these.

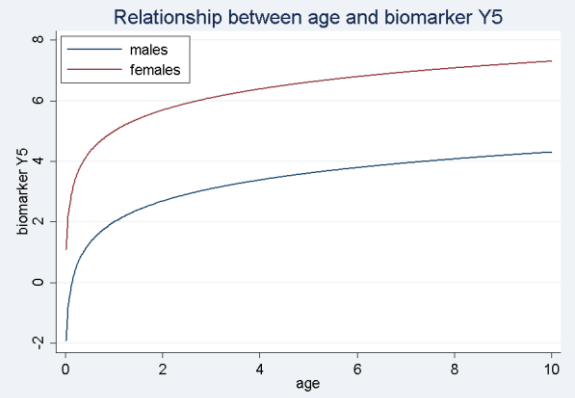
Suggested Resources:

Graphs of different functions: http://mathonweb.com/help_ebook/html/functions_4.htm









11. Select the single most correct answer for each of the following:

a. $\sum_{r=1}^{4} r^2$ =

(i) 1 + 2 + 3 + 4

(ii) $1^2 + 2^2 + 3^2 + 4^2$

(iii) $(1 + 2 + 3 + 4)^2$

(iv) $1^2 + 4^2$

b. $\sum_{r=1}^{4} \log(r)$ =

(i) log (1 x 2 x 3 x 4)

(ii) log 1 x log 2 x log 3 x log 4

(iii) log (1 + 2 + 3 + 4)

(iv) $e^{(1 + 2 + 3 + 4)}$

Correct answers: (a) $1^2 + 2^2 + 3^2 + 4^2$ (b) log (1 x 2 x 3 x 4)

Students need to understand what sigma (∑) means (summation). For part (b), it is important to also understand that log a + log b = log (ab).
Suggested Resources:
Very basics of sigma notation: https://www.khanacademy.org/math/ap-calculus-ab/ab-integration-new/ab-6-3/v/sigma-notation-sum
Summation and series expansion: https://www.studocu.com/en-ca/document/sheridan-college/computer-math-fundamentals/lecture-notes/lesson-2-3-summation-series-expansion/3574678/view

12. Solve the following simultaneous equations for x and y:
    a. 2y = 6 and 10y + x = 20
    b. 14x + 5y = 31 and 2x – 3y = -29

Correct Answers:
(a) y=3, x= -10
(b) x=-1, y=9

This question is about solving straightforward simultaneous linear equations.
Common errors include failing to notice the negative numbers, or adding when you should subtract.
Suggested Resources:
http://www.mathcentre.ac.uk/resources/workbooks/mathcentre/web-simultaneous1.pdf
http://mathsfirst.massey.ac.nz/Algebra/SystemsofLinEq/EMeth.htm

Statistics and Probability

When doing Population health science, we must deal with uncertainty constantly. Often we have only a small sample, and wish to make an inference about a wider population. Often we are interested in understanding risk factors (that is, the probability of an outcome occurring) rather than dealing with a deterministic process. This is why statistics is such a key tool.

The following tables give data from a non-randomised study medical study comparing the success rates of two treatments for kidney stones – open surgery vs percutaneous nephrolithotomy

Patients with small kidney stones	Treatment Success	Treatment Failure
Open Surgery	81	6
Percutaneous Nephrolithotomy	234	36

Patients with large kidney stones	Treatment Success	Treatment Failure
Open Surgery	192	71
Percutaneous Nephrolithotomy	55	25

13. Based upon this data:
    a. Which of the two treatments results in the greatest chance of success overall?
    b. Looking only at patients with small kidney stones, which of the two treatments results in the greatest chance of success?
    c. Looking only at patients with large kidney stones, which of the two treatments results in the greatest chance of success?
    d. What would be your clinical recommendation based on this data?

Open Surgery overall has a success rate of (81+192)/(81+6+192+71)= 78%
Percutaneous nephrolithomy overall has a success rate of (234+55)/(234+36+55+25)=83%
So, overall, percutaneous nephrolithotomy has the greatest chance of success

In small kidney stones:
Open surgery has a success rate of (81)/(81+6)=93%
Percutaneous nephrolithotomy has a success rate of (234)/(234+36)=87%
So, in small kidney stones, open surgery has the greatest chance of success

In large kidney stones:
Open surgery has a success rate of (192)/(192+71)=73%
Percutaneous nephrolithotomy has a success rate of (55)/(55+25)=69%
So, in small kidney stones, open surgery has the greatest chance of success

Here we see the impact of a confounding factor, size of kidney stone. The success rates are higher for small kidney stones than large ones, regardless of the treatment used. However, open surgery is more used when treating large kidney stones, while percutaneous nephrolithotomy is more used when treating small kidney stones. This means that, while percutaneous nephrolithotomy performs worse on each subgroup (and hence should not be clinically recommended, it appears to do better overall).

Note, this is an example of Simpsons Paradox:

https://towardsdatascience.com/simpsons-paradox-how-to-prove-two-opposite-arguments-using-one-dataset-1c9c917f5ff9

Medical data was collected from a sample of adults in the UK. The following table shows the percentage of adults in each age group who had a diagnosis of cardiovascular disease

Age	18-34 years	35-44 years	45-54 years	55-64 years	65 years and over
Percentage Diagnosed	6%	13%	27%	47%	65%

14. For each of the following questions, either answer the question or explain which piece(s) of additional information you would need to answer the question:
    a. In which age group is a diagnosis of cardiovascular disease most common?

b. In which age group are the majority of people in the sample with cardiovascular disease?
c. What is the overall prevalence of cardiovascular disease in the sample?
d. What is the overall prevalence of cardiovascular disease in the UK?
e. Is there a statistically significant difference between risk of heart disease in those who are 45-54 years old and those who are 55-64 years old?
f. Five people are chosen at random from the sample – one from each age group. What is the probability that none have been diagnosed with cardiovascular disease?

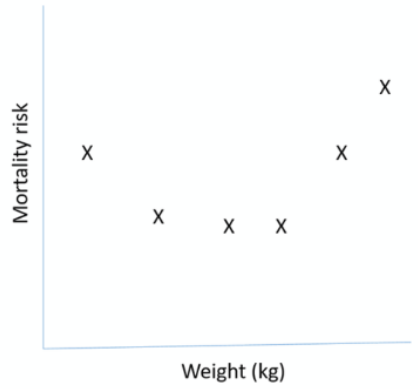
a. It is most common in the "65 years and over" group
b. We would need to know the total number of people in each age group to determine this
c. We would need to know the total number of people in each age group to determine this
d. Due to sampling error, we will never be able to know this from a sample – at most, we will be able to calculate a good estimate, and give a measure of our uncertainty (unless the sample actually contained data from all adults in the UK).
e. We would need to know the number of people in the two age groups in order to perform a statistical test and determine significance. The difference between 30% and 50% will not be significant when our sample size is 10 in each group, but will be highly significant when the sample size is 10,000 in each group.
*f. Prob(none diagnosed)=Prob(person from group 1 not diagnosed)*Prob(person from group 2 not diagnosed)*Prob(person from group 3 not diagnosed)*Prob(person from group 4 not diagnosed)*Prob(person from group 5 not diagnosed)= 0.94*0.87*0.73*0.53*0.35=0.11 i.e. 11%*

Suggested Resources:
Basic probability: https://www.bbc.co.uk/bitesize/guides/zk9dmp3/revision/2
Dependent and independent events: https://www.mathsisfun.com/data/probability-events-independent.html
https://www.khanacademy.org/math/statistics-probability/probability-library#multiplication-rule-independent

Drawing Conclusions from Data

The key task when performing any kind of quantitative research or data science is to be able to draw meaningful conclusions from the data collected. In order to do this, we need to be able to design good experiments, make sure the correct data is gathered, and that our statistical methods are appropriate to the task. However, sometimes even once we have the results of the study, the correct interpretation is not straightforward…

Note: these questions are not about statistical significance – there is no need to include p-values, etc, in your answer.



15. A doctor has been monitoring the relationship between patients' body weight and mortality (dying). In the graph, each point represents 100 patients, with their mortality risk (proportion of the 100 who died) shown in relation to their body weight. The correlation between body weight and mortality risk is 0.03.
    a. The doctor concludes that weight is of no use in predicting mortality risk. Do you agree?
    b. A different doctor concludes that weight is of use in predicting mortality risk. They suggest an intervention aimed at helping patients reach a healthy weight would reduce mortality. Do you agree?

Correct answer: No, you should not agree – the doctor's conclusion is not correct.

Correlation is a measure of the goodness of fit of a linear model for the relationship between the two variables. In this case, the relationship is not linear (you can see this from the graph, but you could also bring up that your prior belief about the effect of weight on health would predict that people with lower weight and higher weight would have poorer outcomes), so a linear model does not fit well. However, a non-linear model may provide a good fit and accurate prediction; students could suggest a quadratic relationship based on the curve. Students could suggest analysing low weight and high weight separately. Students could suggest using absolute difference from mean weight / some clinical idea of "ideal" weight.

Hopefully even students who don't understand what is going on will look at the graph and see that there is some kind of a relationship there. The key thing to understand is that the correlation calculated is about a straight line, whereas here we appear to have a non-linear relationship.

Suggested resources:
This document offers ideas on how to interpret correlations, including when there is a clear, but non-linear relationship:
http://users.sussex.ac.uk/~grahamh/RM1web/Eight%20things%20you%20need%20to%20know%20about%20interpreting%20correlations.pdf

This resource can help you get a feel for what different correlations actually look like:
https://gallery.shinyapps.io/correlation_game/

However, the presence of a correlation does not imply causation. In this case, it is possible that the weight is directly responsible for the change in mortality risk, however it is possible that there is a third variable affecting both (a confounding factor), or that the causation is the other way around (for instance, patients in late stages of illness losing weight). Does this matter? It depends on the purpose of our study. If we were interested in predicting mortality, correlation might be sufficient. However, if we are suggesting an intervention, we require causality – hence RCTs.