Self-screening tool for MPhil in Population Health Sciences candidates

Attempt all questions. 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

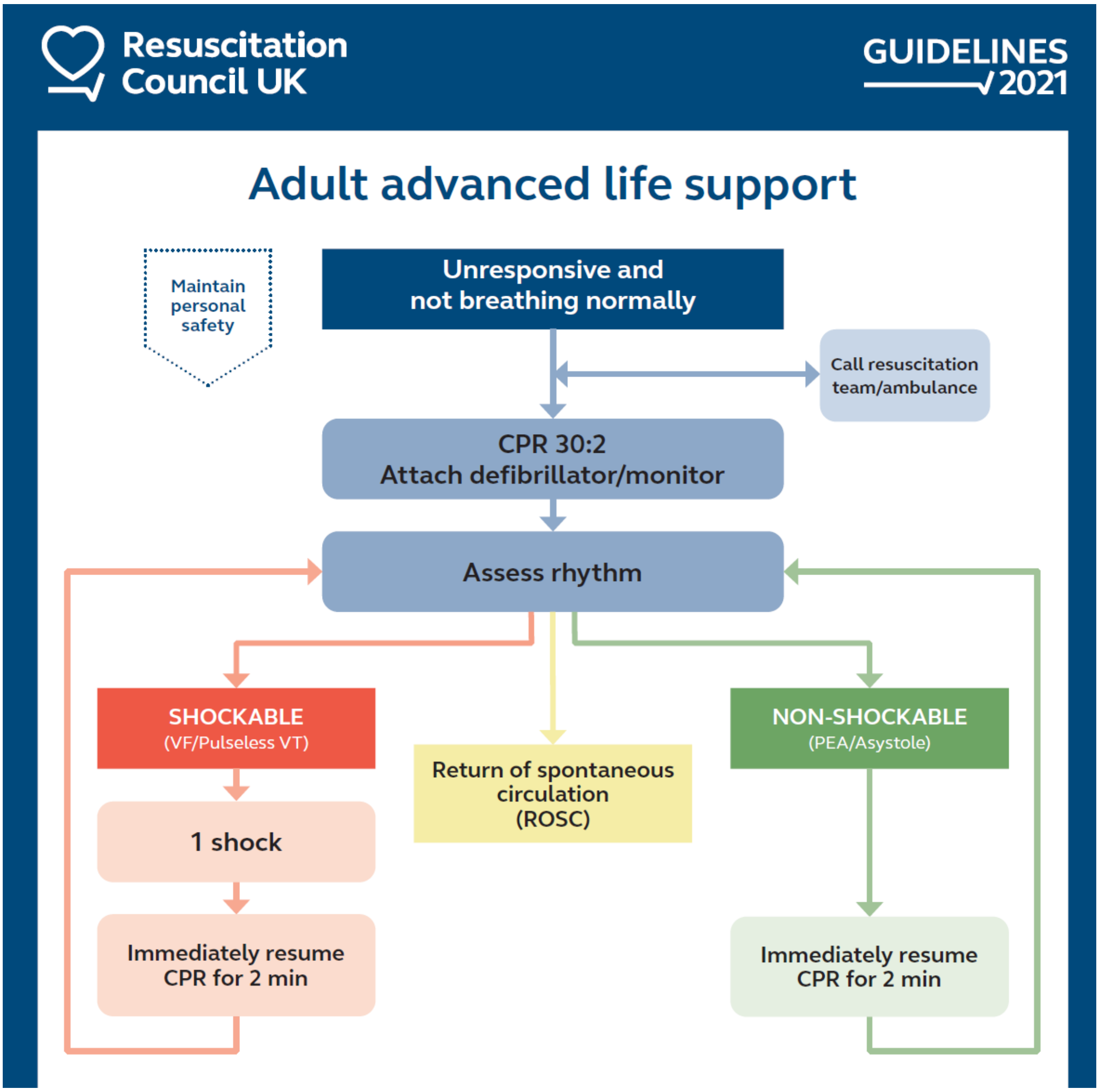
Population health sciences can involve very large numbers (and very small probabilities). It can involve comparing data across multiple datasets, which might not be displaying numbers in the same way, or even using the same units.

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

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 Resusitation Council UK's 2021 algorithm for adult advanced life support:

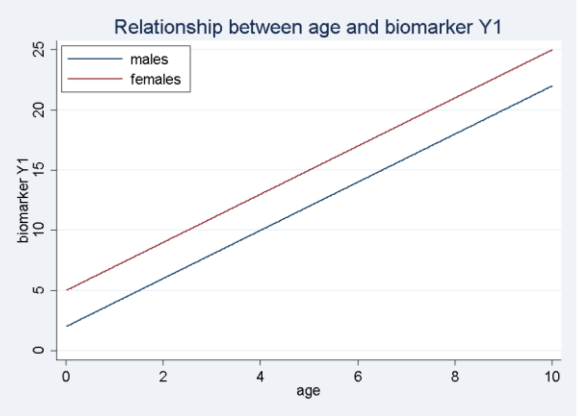


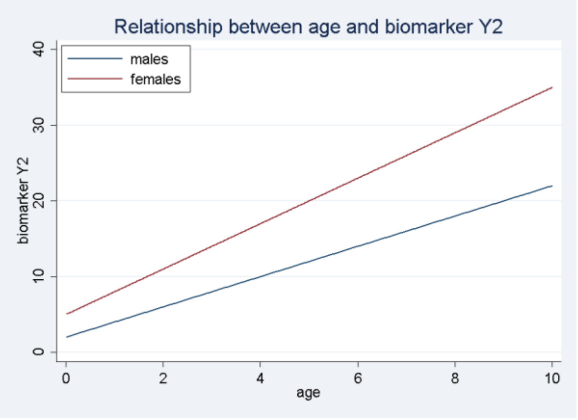
2. The life support algorithm above has a loop – the command "Immediately resume CPR for 2 min" is repeated endlessly (or until ROSC is achieved). A loop may repeat a fixed number of times, be applied to all individuals in a sample or repeat until a certain condition is met. Come up with an algorithm you have used in work/study/life which has a loop, and write it down as a flow chart.

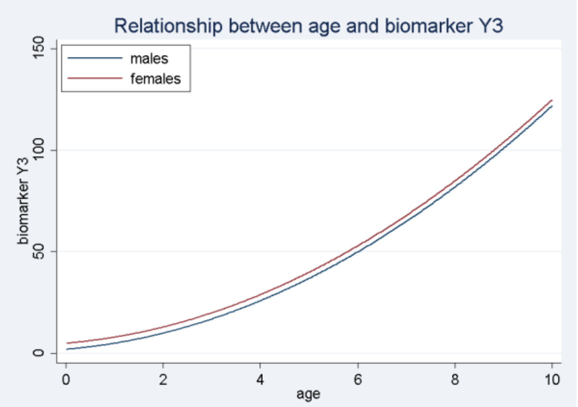
Algebra of Regression

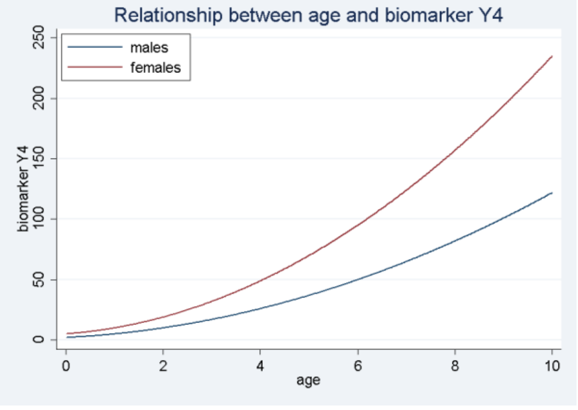
One of the most common statistical tools used, particularly when we wish to control for the effect of confounding variables or make predictions, is regression. The Principles of Biostatistics module will cover linear regression and logistic regression. In order to interpret the output of these models,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

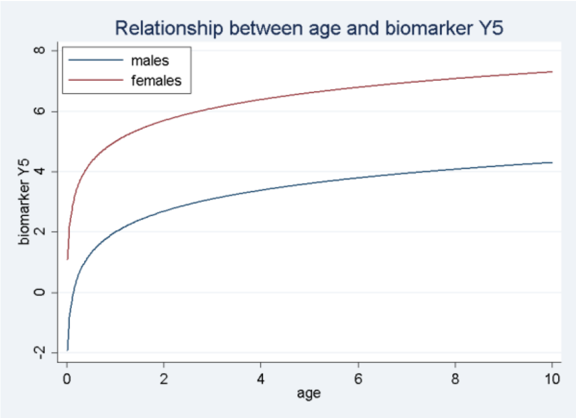
Note: for all of these questions, log means natural logarithm, base e; exp(x) means e^x


Relationship between age and biomarker Y1
males
females
biomarker Y1
0
5
10
15
20
25
0
2
4
6
8
10
age


Relationship between age and biomarker Y2
males
females
biomarker Y2
0
10
20
30
40
0
2
4
6
8
10
age


Relationship between age and biomarker Y3
males
females
biomarker Y3
0
50
100
150
0
2
4
6
8
10
age


Relationship between age and biomarker Y4
males
females
biomarker Y4
0
50
100
150
200
250
0
2
4
6
8
10
age



3. 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 x age + age^2 + 3 x female + age x female + age^2 x female
    b. Equation 2: Y = 2 + log(age) + 3 x female
    c. Equation 3: Y = 2 + 2 x age + 3 x female + age X female
    d. Equation 4: Y = 2 + 2 x age + age^2 + 3 x female
    e. Equation 5: Y = 2 + 2 x age + 3 x female

4. 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} x X
    e. Y =exp(2X+6)
    f. Y=(exp(X))/(1+exp(X))

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

6. Write each of the following expressions as a single exponential term:
    c. exp(a) x exp(b)
    d. exp(a) / exp(b)

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

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

8. 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?

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

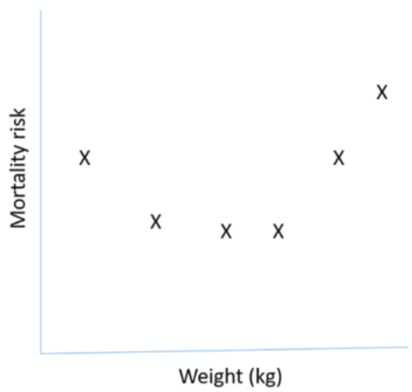
Percentage Diagnosed	6%	13%	27%	47%	65%

9. For each of the following questions, explain which piece(s) of additional information you would need to answer the question:
    a. In which age group are the majority of people in the sample with cardiovascular disease?
    b. What is the overall prevalence of cardiovascular disease in the sample?
    c. What is the overall prevalence of cardiovascular disease in the UK?
    d. 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 in the sample?

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.



10. 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?