How to Guide

**This document contains all the information you need to complete the GSK Next step live online career challenge.**

**You will also need the following files:**

* DigData – Dataset [Excel file: subject\_data.xlsx]
* DigData – Excel Cheat Sheet [PowerPoint file]
* GSK Next Step [PowerPoint file]

**What’s in this document?**

This document contains the following:

1. Overview of the challenge
2. A series of questions to help you approach the challenge
3. DigData - dataset overview
4. How to present your answers/findings
5. **Overview of the challenge**

GSK are designing a study to test a new Flu vaccine on subjects. We need to recruit at least 200 subjects for our study. This subject data will help us to decide whether our vaccine progresses to the next stage of development. The GSK team need your help to select at least 2 hospitals to take our subjects from.

It is important to ensure that we only recruit subjects who meet our inclusion/exclusion criteria. This is so the subjects will be representative of the target population, and to ensure we can produce reliable and reproducible results. Therefore, we want to choose hospitals with the greatest number of subjects meeting these requirements.

Our Inclusion/exclusion criteria are as follows:

* Aged 50 or above
* Has a peak flow reading above 400 litres per minute (litres/min)
* Has not had a flu case or flu shot in the last 6 months
* Does not have an egg allergy
* Does not have a history of respiratory conditions
* Has a BMI (Body Mass Index) between 18-32 kg/m2
* Non-smokers only

**Extra details on our inclusion/exclusion criteria:**

* Peak flow is a simple measurement of how much and how quickly you can blow air out of your lungs. It is often used to help diagnose and monitor asthma. 400-600 l/min is considered normal, an individual with asthma will have a lower reading between 200-400 l/min. [Peak flow Measurements (wales.gov.uk)](http://resources.hwb.wales.gov.uk/VTC/2012-13/22032013/hsc/cym/unit-4/u5-ioph/unit-5-peak-flow.htm#:~:text=The%20highest%20peak%20flow%20reading,200%2D400%20l%2Fmin)
* Having flu in the last 6 months increases the number of antibodies a subject has, to fight flu and therefore might affect the results of the clinical trial, so we want to exclude subjects that have had flu in the past 6 months. Similarly, those who have recently been vaccinated against the flu are also likely to have developed immunity, so we want to also exclude these subjects as well.
* Some Flu vaccines are made using egg, so we are excluding subjects with egg allergies: [Flu vaccine - NHS (www.nhs.uk)](https://www.nhs.uk/conditions/vaccinations/flu-influenza-vaccine/)

We need your help to answer the following questions using our data:

* **What are the subject characteristics of each site? (See section 2.1)**
* **Is the data accurate/reliable for each site? Can you spot any problems in the data? (See section 2.4)**
* **We are looking for at least 2 hospitals to use for our clinical trial, based on your findings can you suggest which hospitals are best? (See section 2.3)**

1. Questions to help you approach the challenge
2. What are some summary statistics for each site? Examples: Average age per site? Proportion of Male/Female? Minimum and maximum values? (See Dig Data – Excel cheat sheet PowerPoint)
3. Can you create a new variable to better summarise the Peak Flow Rate of a subject? Calculate the max of the 3 values.
4. How many subjects meet the inclusion exclusion criteria per site?
5. Are there any missing values? Any unrealistic values? (e.g., age >150 or age <0) Are there repeated values?

These are just example questions to get you thinking. Feel free to explore and represent the data however you wish!

1. DigData– dataset overview (Excel worksheets)

Subject\_data.xlsx contains information on 2284 rows.  Don’t worry we don’t expect you to manually inspect every row! We'll show you how to explore a large dataset with ease in the excel cheat sheet PowerPoint.

The following table shows what each column in the dataset represents and the type of the column if it is a text column or a numeric column.

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| SITEID | Text | Unique ID value for each hospital |
| SUBJID | Text | Unique ID value for each subject |
| SEX | Text | Subject sex |
| AGE | Numeric | Subject age |
| RACE | Text | Subject race |
| ALLERGY | Text | Subject allergy information |
| MEDICAL | Text | History of medical conditions |
| SMOKER | Text (binary) | Is subject a smoker? |
| LAST\_DOSE\_FLU | Text (binary) | Has subject had the flu/a flu shot in last 6 months? |
| BMI | Numeric | Body mass index value |
| PEAK\_FLOW\_1 | Numeric | Peak flow reading 1 |
| PEAK\_FLOW\_2 | Numeric | Peak flow reading 2 |
| PEAK\_FLOW\_3 | Numeric | Peak flow reading 3 |

1. How to present your answers/findings

Once you have explored the data you can create charts to summarise your findings or screenshot tables. Add these onto a few slides along with a bullet point summary of which hospitals you think we should choose. You should not need more than 4-5 slides, with a few findings and a summary.

We’re also happy for you to present back in a word document or a spreadsheet, as long as the findings are clear.