HOW-TO-GUIDE

GSK - DIGDATA ONLINE STEP-UP CAREER CHALLENGE

This document contains all the information you will need to complete the GSK live online career challenge. You should have received the following files:

- How-to-Guide (this document)
- Three data sets (**clinical-study.csv, protein-levels.csv, clinical-study-ae-taskB.csv**)
- Presentation (**DigData–Step-Up-Presentation.pptx)**
- Excel cheat sheet (**excel_cheatsheet.pptx**)
- R Task sheet (**step_up_with_r.html**)
- Document about analytical roles in the pharma industry (**analytical_roles.pdf**)

THE SCENARIO

Researchers at GSK have been working on an innovative family of drugs to help people with solid tumours that continue to grow despite other treatments. A promising candidate, Miraculon-B (a fake name, obviously!), has been making its way through our development pipeline.

A late phase clinical trial for Miraculon-B has recently finished. This trial was focused on understanding if Miraculon-B was more effective than the standard of care in shrinking solid tumours in patients that don't respond to other treatments. To determine this, the clinical trial data needed to be analysed to understand whether patients saw enough of a benefit from the treatment. Work is underway to present a package to the regulatory agencies that describes who is most likely to benefit from this new medicine and to seek agreement on our proposed strategy for aiding prescribers considering the value of Miraculon-B for their patients.

WANT TO LEARN MORE ABOUT CLINICAL TRIALS?

In the presentation we introduced what an oncology clinical trial is, but if you want to learn more about clinical trials and how they work, check out the following links from the Cancer Research UK website:

- Phases of clinical trials | Cancer Research UK
- Randomised trials | Cancer Research UK

HOW-TO-GUIDE

INSTRUCTIONS

You can choose to work on **Path A or Path B:**

Path A: Wrangle and analyse the data to find out which patients benefit from the medicine. Can we conclude whether Miraculon-B is more effective than the standard of care?

Path B: Design a web application prototype that will help doctors understand how to effectively prescribe Miraculon-B to their patients according to the risk-to-benefit ratio obtained from the clinical trial data.

We suggest you choose path A if you enjoy problem-solving and want to strengthen your skills in data wrangling, data analysis and data visualisation using spreadsheet software or a programming language such as R.

We suggest you choose path B if you are interested in basic exploratory data analysis using spreadsheet software, the concept of 'design thinking' and its applications in creating data science products, practical experience in creating user interface mock-ups, and honing your creativity and problem-solving.

Both paths show a data role's approach to a problem and have the same level of difficulty, although they cover different skills.

Each section will have the relevant information to take you from zero to hero!

HOW-TO-GUIDE

PATH A: HELPING THE TEAM WITH DATA ANALYSIS

THE PROBLEM:

A new cancer treatment for solid tumours (Miraculon-B) is in development at GSK, data has been collected in our recent clinical trial comparing Miraculon-B to the current standard of care (i.e., referred to as the 'control'). The trial has completed, and we need your help to analyse the data and assess the effectiveness of Miraculon-B. To do this we need you to compare Miraculon-B to the standard of care, and explore how different patient sub-groups may benefit differently from treatment.

REQUIRED SOFTWARE SET-UP:

Posit Cloud – This is the online tool we recommend for tackling this path using the *R programming language*. It is an online instance of the popular RStudio software. You can create a free account at https://posit.cloud/ – simple!

Excel – If you feel more comfortable, you can choose to complete this task using Microsoft Excel spreadsheet software. If you don't have an instance of Excel on your computer you can use a free, online version of Excel by creating an account here: Free Microsoft 365 Online | Word, Excel, PowerPoint

THE DATA:

You should have two .csv files which contain the data from the clinical study.

DATA DESCRIPTION

GSK has shared 2 datasets, **clinical-study.csv** and **protein-levels.csv**

The dataset **clinical-study.csv** contains 772 rows (patients) and 7 columns (variables). The "**response**" column is our target showing whether the patient has responded to treatment ("Yes") or not ("No"). In solid tumours, response is based on whether tumours shrink, stay the same, or get bigger. If a patient has a tumour that is shrinking, they are classified as responder, if they have a tumour that stays the same or gets bigger, they are classified as non-responder. Again, please note that this is not real GSK data, but it is similar to the type of data and the type of questions we explore every day!

The dataset **protein-levels.csv** contains 768 rows and 2 columns, The **protein_concentration** column shows the concentration of a protein that has been identified as a potential predictive *biomarker* for solid tumours. Predictive cancer biomarkers can be used to identify the patients who are or who are not likely to derive benefit from specific therapeutic approaches. In this analysis we want to investigate whether the concentration of the protein could predict whether the patient will respond or not to treatment.

DATA DICTIONARY

clinical-study.csv:

COLUMNS	DESCRIPTION
subject_id	Patient ID
sex	Whether the patient is male or female
age	The age of the patient
weight	The weight of the patient
height	The height of the patient
trt_grp	Whether the patient is receiving the new drug or the standard of care (control)
Response	Whether the patient responded or not (Yes [Y]/No [N])

protein-levels.csv

COLUMNS	DESCRIPTION
participant_id	Patient ID
protein_concentration	Concentration of a blood protein (ug/L) that might be a potential predictive biomarker of response

YOUR JOB:

Your job is to examine and analyse the data provided to understand which subgroups of patients are benefiting more from the treatment.

We want to address questions like:

- Do patients that take the GSK drug respond more to treatment compared to those in the control group?
- Can the age, weight or protein concentration of patients predict whether they will respond better to treatment or not?

Below is a list of tasks you would typically go through to analyse the clinical data.

- If you are completing this challenge in R, please refer to the provided **.html** file for a rundown of the tasks required to analyse the data!
- If you are completing this challenge in Excel, please refer to the **excel_cheatsheet.ppt** for hints and tips on how to complete the tasks in Excel!

STEP-BY-STEP GUIDE

1. **Read in the data (Or open the file in Excel)**
2. **Examine the data**
3. **Clean-up the data**:
    a. You may have noticed that the first and second rows are identical, including the subject_id. Looks like someone accidentally added the same data twice. How can you remove such duplicates?
    b. You may have noticed that some rows are paediatric data (the age of the patient is below 18). These should be excluded from the analysis. How can you remove these rows?
    c. You may have noticed a small number of NA (missing) values in some of the columns. How can you deal with these?
4. **Create new variables**:
    a. You need to add a new column to the data for the BMI of the subject. Recall that BMI is calculated by dividing weight by the square of height.
    b. Finally, you should merge the separate data we have for each patient's protein concentration. This can be a tricky step but with the correct function you do it. (*Hint: You can merge the two datasets by the patient ID; it will be a lot easier if you renamed the patient ID columns, so they have the same name!)
5. **Aggregate the data**:
    a. Compare mean age in two treatment groups
    b. Compare mean age in responders vs non-responders
    c. Compare responders and non-responders in the two treatment arms
    d. Compare mean weight in responders/non-responders etc
    e. Compare protein concentration in responders vs non responders
6. **Visualize the data:**
    a. Creating plots/visualisations allows us to identify patterns in the data quickly. Try to create the following plots and see if you can identify any trends on which sub-group of patients is responding better to treatment:
        i. Boxplot of age(y-axis) by response (x-axis)
            1. Also separated by treatment group?
        ii. Boxplot of weight/BMI(y-axis) by response (x-axis)
            1. Also separated by treatment group?
        iii. Boxplot of protein_concentration (y-axis) by response (x-axis)
            1. Also separated by treatment group?
7. **Modelling (advanced task[1]):**

[1] Step 7 is an optional, advanced task for those with an interest or with previous knowledge of modelling, looking to apply that knowledge to a real problem.

a. Formal analysis of data can be done in many ways according to many different models - we use the insight from these models to learn the complex patterns in the data that are less evident from summary statistics or simple plots alone. The statistician George Box famously said, "all models are wrong but some are useful", meaning that a model will never capture everything, but even if it gets close enough then it can still be valuable. The task here is for you to try and fit a model (any model! E.g a logistic regression, a random forest etc) to the data and see what you can learn about the patterns in the data that might help us understand who is benefiting the most from our treatment.

8. **Interpret and report your findings:**
   a. An important requirement for a data role is to be able to present your work. Create a PowerPoint presentation to present your analysis and results to the key stakeholders.
   b. Record yourself giving a 3-minute presentation which demoes your work. There are several resources online on how to create effective presentations. Note: You could present to your friends and family. You are not required to send us the video.

YOU REACHED THE END OF PATH A! CONGRATULATIONS!

We hope you learned something, and we have inspired you to pursue an analytical role in the pharmaceutical industry.

TASK B: DESIGNING A WEB TOOL

THE PROBLEM:

GSK has run a clinical trial for a cancer treatment called Miraculon-B. Our statisticians analysed the results and found that patients were more likely to respond to this treatment than to the standard of care.
Unfortunately, patients treated with Miraculon-B were more likely to experience a tolerable but undesirable side effect. Our analysts were able to identify a protein in the blood which is linked to both response and this adverse event. Our analysts were able to identify a protein in the blood which is linked to both response and this adverse event.

To help doctors prescribing Miraculon-B evaluate the risk-to-benefit ratio, our data scientists at GSK developed an algorithm that shows which patients are more likely to benefit from treatment based on their protein concentration and other parameters. GSK intend to develop this algorithm and present a package for approval by the regulator for its use in helping prescribers of Miraculon-B.

YOU JOB:

Design a web application that is an interface for the algorithm. The web application should provide an overview of the clinical trial findings in an easy-to-understand format. It should also allow doctors to enter their patients' values to help them understand the benefit-to-risk ratio of prescribing Miraculon-B compared to the standard of care.

You should come up with your own ideas to recap the clinical trial findings and design the user interface of this web tool. Follow the step-by-step guide and you'll create an excellent prototype in no time!

THE DATA:

The **clinical-study-ae-taskB.csv** contains a clean data set (one row per patient).

COLUMNS	DESCRIPTION
participant_id	Patient ID
sex	Whether the patient is male or female
age	The age of the patient
weight	The weight of the patient
height	The height of the patient
treatment_group	Whether the patient is receiving the new drug or the standard of care (control)
Response	Whether the patient responded or not (Yes [Y]/No [N])
protein_concentration	Concentration of the protein which was linked to increase chance of dangerous side effects (ng/mL)
ae	Whether the patient experience a severe adverse event (Yes [Y]/No [N])

AN EMAIL FROM OUR DATA SCIENTISTS (KEY INFORMATION)

Below is the email you have received from the Data Science team. It contains the key information you need to know to help you complete this task:

Dear DigData participant,

We hope this email finds you well. We are writing to share with you some exciting results from our recent analysis of the clinical trial data of Miraculon-B.

Our team of data scientists conducted an analysis on the ***clinical-study-ae-taskB.csv*** *data and found the following:*
The clinical trial lasted 48 months and involved 768 patients with advanced-stage cancer. Half of the patients were given the new treatment, while the other half were given the standard of care. The study measured the efficacy and safety of the new treatment compared to the standard of care. The study found that patients were more likely to respond to Miraculon-B than to the standard of care. However, patients were more likely to experience an undesirable side effect with Miraculon-B . The occurrence of the undesirable side effects was directly linked to a protein in the blood: the higher the concentration of the protein, the more likely the patient was to experience the adverse event. However, the same protein was linked to the probability of response: the lower the concentration of the protein, the more likely the patient was to respond to Miraculon-B.

Using these data and published literature in this space, we have developed an algorithm that can be used to predict the risk-to-benefit ratio of prescribing Miraculon-B to individual patients with specific characteristics.
We believe that our model can help prescribers make more informed decisions about whether Miraculon-B is the right treatment to prescribe their patients.

Can you help us design the user experience of web application that summarises the study for doctors in an easily digestible manner and allows them to interact with our model?
You can start by looking at the data to familiarise yourself with the trends we identified.

Best regards,

The Miraculon-B Data Science Team

STEP-BY-STEP GUIDE

STEP 1: WRANGLING AND ANALYSING DATA

WHAT IS DATA ANALYSIS?

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

YOUR TASK

Open the file with a spreadsheet editor, such as Excel. Exploring this data will help you become familiar with the clinical study for Miraculon-B. Note that this is not real GSK data, but it is similar to the type of data and the type of questions we explore every day

If you want, you can use the data to perform data analysis tasks such as sorting, counting, averaging or plotting (similar to tasks in Path A, see page 5 above). These tasks can be done manually or via pivot tables (recommended). You **don't need to complete this step**; it is only here to help you familiarise yourself with the data — once you feel comfortable, you can move on to the next step!

STEP 2: DESIGN THINKING EXERCISE

WHAT IS DESIGN THINKING?

Design thinking is a problem-solving approach that focuses on understanding the needs of the people who will use a product or service and using that understanding to create solutions that meet their needs in an innovative and effective way. When it comes to creating a web application for doctors, design thinking is especially important because doctors deal with complex information and need user-friendly tools that minimise the chance of costly mistakes. Design thinking helps us to understand the needs of doctors and create applications that are user-friendly and meet their needs. This ultimately leads to better patient care.

In the design thinking process, we start by gathering information about the needs of doctors, such as what kind of information they need to make decisions, what their workflow looks like, and what their pain points are. We then use that information to brainstorm ideas for a web application that could meet their needs. Once we have some ideas, we create a prototype of the web application, which is a basic version of the application that we can test with a sample of doctors. We ask the sample to use the prototype and give us feedback on how well it meets their needs. Based on their feedback, we make changes and improvements to the prototype until we have a version of the application that is ready for release.

FOR THE PURPOSE OF THIS TASK, WE ASKED DOCTORS WHAT THEY NEED

One of the most complex tasks for doctors is determining the risk-to-benefit ratio of a treatment. While the primary goal of drug is to cure or alleviate symptoms, it is important to weigh the potential benefits against the potential risks. Often, this task requires a thorough

understanding of the patient's medical history, current symptoms, and potential side effects of the medication. Additionally, doctors must consider other factors that may contribute to the patient's health, such as underlying medical conditions, lifestyle habits, and other medications that the patient may be taking. In general, balancing the potential benefits and risks of a medication is a critical aspect of providing effective medical care.

In this case, the condition being treated is life-threatening and Miraculon-B is a product that improves patients responses. However, there are undesirable side effects associated with its use. Doctors need a simple way to quantify the risks and benefits of using it for specific patients; the information on a drug label for a marketed treatment is frequently based on the average benefit seen in eligible patients consistent with those patient types that were included in the original clinical trial. Were unique patient characteristics (e.g. a higher concentration of a specific protein) to be important then this detail would need to be additionally reflected in the label, possibly alongside an approved mechanism for helping prescribers use this information.

By providing doctors with an interactive web app, they can quickly and easily access the information they need to make informed decisions about patient care. This can improve patient outcomes.

YOUR TASK

Come up with ideas for the interface and behaviour of a web application that lets doctors interact with the statistical model for the risk-to-benefit calculations. This includes thinking of how its results are displayed and how its inputs are obtained.
Your app could also help aggregate relevant clinical information, such as information related to the clinical trial for this drug and the disease, patient populations, and more.

STEP 3: WIREFRAMING EXERCISE

WHAT IS WIREFRAMING?

A wireframe is a simple and basic visual representation of a website or an app. It is like an empty shell with placeholders. Wireframes help to plan the layout and user interface of the website or app and can be used to check with the final users that the application meets their requirements. Think of it like a rough sketch of what the website or app will look like.

YOUR TASK

Create a prototype of the web application. There are several free tools to do this, including pen and paper! Another popular free option is Figma (free, with tutorials on Youtube). You could also do this in PowerPoint.

STEP 4: PRESENTATION

WHY?

An important requirement for a data role is to be able to present your work. Without a good presentation, your work could fail to inspire change.

YOUR TASK:

Record yourself giving a 3-minute presentation which demoes your app and its features. There are several resources online on how to create effective presentations.

Note: You can present to your friends and family. You are not required to send us the video.

STEP 5: RETROSPECTIVE

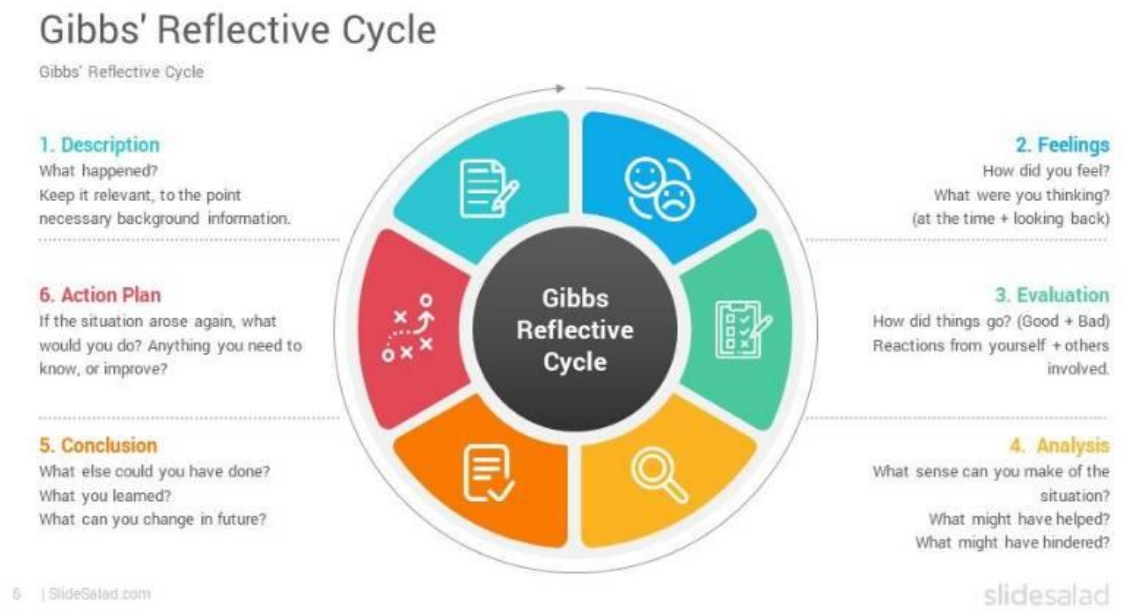
WHY?

A data person is always improving by formal learning and trial and error. This is what makes someone good at their job. As well as asking regularly for feedback, they should reflect.

YOUR TASK

You are nearing the end of this project. Write a project retrospective about what went well, what could have been better, and what you would do differently in a similar future project. This should be a text file between 500 and 1000 words. Your retrospective should cover the steps of Gibbs' reflective cycle.

[Gibbs' Reflective Cycle | The University of Edinburgh]



YOU REACHED THE END OF PATH B! CONGRATULATIONS!

We hope you learned something, and we have inspired you to pursue an analytical role in the pharmaceutical industry.