Paper PD07

The Beginner's Guide to Data Analysis

Shuva Dey, Roche Products Ltd, Welwyn Garden City, UK

Abstract

What is it like to leave your comfort zone and begin a new journey in data analysis/science? Programming ADaMs, tables, listings and graphs and working with clinical data is the bread and butter of statistical programmers as well as working with different stakeholders.

This paper will share how to deal with the change from a beginner's perspective in a rapidly changing environment but first and foremost, to enjoy the experience. A growth mindset is incredibly important in an ever changing and evolving industry where pharmaceutical companies are preparing for the next 10 years and beyond and employees will be key to this change. This is my experience, my thoughts (including 'Imposter Syndrome') and I hope it will resonate with many others who are, or have been through this journey and appeal to others who may be thinking about such a change.

Introduction

The title of this paper refers to Data Analysis. This is meant in the context of clinical data where the analyses helps to determine whether the primary end point(s) of the study or clinical trial are met as defined in the protocol. Essentially analysis is performed in order to help answer a particular clinical question(s) or hypothesis. The analysis involves statistical concepts and methodologies to summarise and visualise the data collected throughout the trial. Increasingly as pharma disrupts the traditional operating model and ways of working the statistical programmer, data analyst etc. will be the data scientist of the future. Hence I will refer to the role as Data Scientist from here on.

I will share my journey into data analysis, my experience and learnings. My perspective is from a data scientist in *Early Development trials, namely Phase 1* but I trust it will be equally helpful to all. The pre-requisite to analysis is tabulating collected data in SDTM format which I will not be focusing on.

Background

I have spent most of my career working with clinical trial data within the realms of Data Management focusing on eCRF design, database build and data tabulation in SDTM format. I am a programmer and have knowledge of SQL, SAS, C# and Spotfire and have worked with various Clinical Data Management Systems including Oracle-Clinical and Rave. I especially enjoyed tabulating data to SDTM and SAS programming and wanted to have more involvement in how the data I tabulated in SDTM was analysed and how it informs decisions. At Roche, Data Management performed SDTM programming and when it transitioned to the Statistical Programming function, I felt that this was the perfect opportunity to follow this task and make this internal move!

Will I make a good data scientist?
Performing a brief SWOT analysis always helps you with where you are currently and where you need to be. It is a worthwhile brief exercise to evaluate yourself. In my case the opportunity had already presented itself and I sized up where I stood in the grand scheme of things.

Strengths:
Knowledge of SAS®
Knowledge of clinical trial data
Attention to detail
Existing user of the statistical computing environment
Good knowledge of SDTM
A-level in Pure Mathematics and Statistics
A passion to learn and a growth mindset

Weaknesses:
Statistical knowledge
Knowledge of ADaM
Knowledge of tables, listings and graphs
Therapeutic area knowledge
R knowledge
Processes

Threats:
Time
Availability of mentor/buddy
Confidence, 'Imposter Syndrome'
Change (organizational, functional, process, systems, software, role etc.

Many of the perceived weaknesses areas are addressed by training. My training was very good and was a mixture of e-learning and classroom-based. At the time I joined the Statistical Programming function a classroom-based accelerated learning academy was created which lead by experienced managers and programmers covered the breadth and depth of the statistical programming and analysis curriculum over a period of 3 weeks or so with hand-on exercises. The rest of my knowledge was gained on-the-job, my mentor/buddy, external courses e.g. R programming and eLearning.

It is also important to be aware of potential threats and manage these effectively. They may not be so apparent when you are thinking about a change in your career direction and probably the last thing on your mind but these are worth bearing in mind now with the hindsight I have!

Ultimately, your strengths and potential opportunities available to you, hopefully help you make the decision when considering data analysis and do not let the weakness and threats put you off.

As with any new role or job after the training it may be weeks or months until you need to recall how to do something or apply your learning in a particular task so training will be re-visited, reference materials consulted and questions asked of your colleagues!

Disruptors, change and focus

As big pharma like Roche, disrupts, revolutionises and evolves it's operating model (KPMG International, Jan 2017) there is a huge impact to all but even more so on those who have decided to undergo a personal transformation and change roles, departments and functions to further their career. In a new role the change may be a little unsettling as the status quo evolves. That feeling of unease and anxiety can be multiplied many-fold if the environment around you is also changing, the function or department is evolving, your line manager is changing and actually the entire organisation before your very eyes!

How to do you cope? Be present! To survive you always have to focus on the here and now, the task in-hand. The focus has to be me. The priority must be on learning the job and upskilling yourself in the basics and fundamentals, otherwise you will be overwhelmed. Try to be aware of what is happening and what changes may affect you e.g. software, statistical computing platform, programming languages etc. but other than that ensure that you are in the best position for the pending and forthcoming changes by being a good data scientist and focus on becoming competent.

Learnings

This section will cover what I believe is important and the areas that I focused on to develop myself towards becoming a competent data scientist.

Know your Study Protocol

The following are just some of the sections of the protocol that focus on key areas:

1) Introduction, study rationale, why is the trial being conducted, what clinical questions(s) does it hope to answer? Is there a formal hypothesis?
2) What are the primary and secondary objectives? How will the corresponding end points be determined?
3) Study Design
4) Statistical Analyses e.g. populations for analyses i.e. how are the different populations defined? For example, safety, efficacy, immunogenicity etc. What will safety and efficacy analyses consist of?
5) Not forgetting the SAP (Statistical Analyses Plan) if the study will be submitted to a Regulatory Authority, typically not the case for Phase 1 studies.

An example of the two primary objectives from a Phase 1 oncology protocol and how the corresponding end points will support them. In fact, the anti-tumour activity was added through a protocol amendment once the safety of the drug was established. In Early Phase trials, determining safety and PK profiles is paramount rather than proving efficacy.

Primary Objectives	Primary Endpoints
• To determine the safety, tolerability, maximum tolerated dose (MTD), and/or a recommended phase 2 dose (RP2D) for intravenous (IV) administration of [redacted] as monotherapy *(Part A)*.	• Nature and frequency of dose-limiting toxicities (DLTs). • Incidence, nature and severity of AEs graded according to the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) v5.0.
• *To evaluate the anti-tumor activity of [redacted] (Part B).*	• *According to Response Evaluation Criteria in Solid Tumors (RECIST) Version 1.1:* ○ *Objective response rate (ORR).* ○ *Disease control rate (DCR); defined as ORR +stable disease rate (SDR).* ○ *Duration of response (DoR).* ○ *Progression-free survival (PFS) defined as the time from the first study treatment (Day 1) to the first occurrence of progression per Investigator assessment or death from any cause, whichever occurs first.*

The protocol also defines further details on the exactly how those end points will be defined/displayed as below. There are references to summaries (tables) and listings.

Table 13 Safety Statistical Analysis Methods

Endpoint	Statistical Analysis Methods
AEs	The original terms recorded on the eCRF by the Investigator for AEs will be coded by the Sponsor. For classification purposes, preferred terms will be assigned by the Sponsor to the original terms entered on the eCRF, using the most up-to-date version of the MedDRA terminology for AEs and diseases. AEs will be graded according to NCI CTCAE grades v5.0. Adverse events will be summarized by mapped term and appropriate thesaurus level. Toxicity grade, seriousness, and relationship to study treatment will be presented, as well as summaries of deaths, AEs leading to death and premature withdrawal from study treatment. Glossary of AEs, medication, and procedures will be provided.
Nature and frequency of DLTs	DLT events will be presented by individual listings. The MTD will be estimated with a mCRM–EWOC using DLT-evaluable participants. The MTD estimate will be presented along with 90% Credible Intervals.

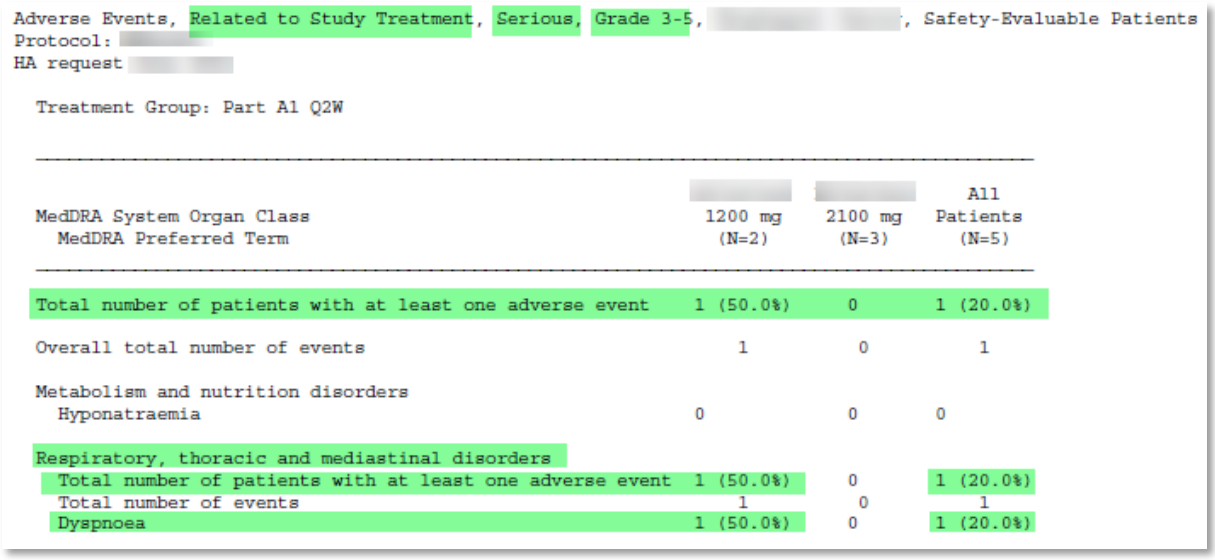
Below is an excerpt from the protocol defining the safety population i.e. all the subjects who are to be included in the tables and listings defined above in determining the safety of the compound.

Safety	All participants who received at least one dose of the study treatment, whether prematurely withdrawn from the study or not, will be included in the safety analysis. Unless otherwise specified, the safety population will be the default analysis set used for all analyses.

The above criteria will be used to flag all subjects who should be in the safety population and this translates in to the ADaM specification for ADSL as below and will be the basis of the programming.

Variable	Label	Derivation / Comment
SAFFL	Safety Population Flag	Set to "Y" if patient received a valid dose. Else to "N" NOTE: A valid dose is defined as (Dose per Administration [EX.EXDOSE] greater than 0) or (Dose per Administration [EX.EXDOSE] equal to 0 and Name of Actual Treatment [EX.EXTRT] contains 'PLACEBO').

Below is an example of a table summarising *serious*, *related adverse events* including *toxicity grade*



Adverse Events, Related to Study Treatment, Serious, Grade 3-5, [redacted], Safety-Evaluable Patients
Protocol: [redacted]
HA request [redacted]

Treatment Group: Part A1 Q2W

MedDRA System Organ Class MedDRA Preferred Term	[redacted] 1200 mg (N=2)	[redacted] 2100 mg (N=3)	All Patients (N=5)
Total number of patients with at least one adverse event	1 (50.0%)	0	1 (20.0%)
Overall total number of events	1	0	1
Metabolism and nutrition disorders			
Hyponatraemia	0	0	0
Respiratory, thoracic and mediastinal disorders			
Total number of patients with at least one adverse event	1 (50.0%)	0	1 (20.0%)
Total number of events	1	0	1
Dyspnoea	1 (50.0%)	0	1 (20.0%)

Statistical concepts and methodologies

These are mentioned in the statistical sections of the protocol and should inform your reading. Understand the terminology. Common text found; descriptive statistics, summaries/summarised by, listings, censoring, confidence intervals, p-values, sample size and power, survival/time-to-event analysis, Kapan-Meier curves etc. to name a few! More on this in a later section.

There are a wealth of resources on the internet, books and short courses or one-day courses that can be attended in person or online. A buddy or mentor or any internal training available will help. Speaking with a statistician will always be beneficial!

I attended some short basic introductory courses external to Roche. Some of the courses that I attended that I would recommend are:

Introduction to Research Methods and Statistics (UCL, Great Ormond Street Institute of Child Health run by the Centre for Applied Statistics Courses, CASC)

Critical Appraisal (also run by CASC as above)

Critical Appraisal and Research Methods (Royal Society of Medicine)

Other prestigious societies running courses include the Royal Statistical Society! Statistics is vast subject and these courses touch on the basics and are great for those without a background in statistics.

Basic statistics will typically cover the following (from the

Introduction to quantitative research
Research question development
Study design, sampling and confounding
Types of data Graphical displays of data and results
Summarising numeric and categorical data
Numeric and categorical differences between groups
Hypothesis testing
Confidence intervals and p-values
Parametric statistical tests
Non-parametric tests
Bootstrapping
Regression analysis

One may be a great programmer but never lose sight of why you are producing the analysis and what are the results telling you. How is the reviewer going to interpret the results and what are they looking for? This is where understanding why you are doing something and knowledge of the terminology is essential to be a complete data scientist.

Therapeutic area knowledge
Familiarise yourself with the therapeutic area within which you are working. For example terminology and medical concepts. In well-established areas of research e.g. oncology there are lot of concepts to understand i.e. RECIST1.1, Overall Survival (OS), Progression Free Survival (PFS), Response Rate, Confirmed Best Overall Response, Disease Control Rate (DCR), Duration of Response (DoR) etc. In oncology these form the basis of primary/secondary end points and will be discussed within the study team. It can be quite overwhelming at first especially if you are new to the therapeutic area or it has been a while since you last worked in that area. Team members will roll these terms off the tips of their tongues, especially Statisticians and Scientists! Therefore, it is worthwhile taking the time to acquainting yourself with therapeutic area medical terminology and understanding their definitions.

Most if not all of these terms will be defined in the protocol under the statistical sections usually.

Specification interpretation and writing
Primarily this will be the analysis dataset and output specifications.

The first time a dataset specification is viewed it can be quite a big surprise! The complexity and length of some of the derivations can be quite a shock at first. The dataset specifications will list all the variables required for analysis. There will be many derived variables. Some will require imputations for missing data e.g. dates/times, population flags, analysis flags etc. A key skill is to be able to interpret and understand a specification especially for a QC programmer or if using template standard specifications with a view to programmatic implementation as a first line programmer. It is also equally and vitally important to be able to write clear and concise specifications too in plain English!

Good practice is to write the specification as the code is developed. Alternatively to complete programming and QC and then translate the program code and enter in the specification. The specification should contain plain logic or pseudo-code. It should be programming language agnostic and a non-programmer should be able to read and understand it e.g. a reviewer from the FDA.

ADaM programming
Know your data
Understand your data; what is been collected and analysed from the eCRF? Not forgetting non-CRF data and how it has been tabulated in SDTM. Before commencing with programming there a few considerations; an understanding of the analysis dataset to be created, what derivations are required, which input datasets are needed, are the correct observations being used? This is where it is fundamental to understand the SDTM data.

CDISC ADaM Principles
In analysis dataset programming there are many derivations required as mentioned in the specification section and they can be quite complex. ADaM stands for Analysis Data Model and is now the expected format for submissions of datasets to Regulatory Authorities for clinical trials data. Assuming that ADaMs are used then learning and applying the ADaM principles for derivation programming is very important for a standardised approach and promotes traceability to the original data.

The ADaMIG v1.1 (CDISC, 2009) is a very useful document that I found extremely helpful especially when programming a study-specific custom BDS (Basic Data Structure) dataset. The structure is vertical or normalised and very flexible as both variables and observations can be derived and there are six rules to consider when deciding whether to create a new variable versus a new observation. It is a popular data structure, generally for assessments taking repeatedly throughout a trial e.g. Vitals (ADVS), ECG (ADEG), ADLB (Labs) etc.

In fact, all of the ADaM documents are helpful and will complement the in-house training most companies invariably have which are ultimately based upon these guiding principles. It is not until you program an analysis dataset from scratch that one fully understands the guidance in the practical application of creating a variable versus a new row in BDS!

In general, most companies will have standard safety and efficacy analysis dataset programming templates that require some ‘tweaking’. However, the newcomer will really come along leaps and bounds when programming non-standard/study-specific safety and efficacy analysis datasets.

Approach to non-standard/study-specific analysis dataset programming
The reason for creating an analysis dataset will be for the creation of some outputs, therefore the analysis dataset should contain all the information required to generate the output i.e. flags, derivations etc. Typically an output will be requested i.e. a table, listing or graph. You can be provided with a mock up especially if it may be quite a complex table or otherwise it will be in the output specifications.

Example of a table mock-up

Baseline Prevalence and Incidence of Treament Emergent Anti-Drug Antibodies (ADAs) to *[Drug]* : <Specify Population>
Protocol: xxnnnnn

	[Drug] Group 1 (N=nnn)	*[Drug]* Group 2 (N=nnn)	All *[Drug]* Patients (N=nnn)	Control Group (N=nnn)	All Patients (N=nnn)
Baseline Prevalence of ADAs					
Baseline evaluable patients	nnn[1]	nnn	nnn	nnn	nnn
Patients with a positive sample at baseline	nnn[2] (xx.x%[3])	nnn (xx.x%)	nnn (xx.x%)	nnn (xx.x%)	nnn (xx.x%)
Patients with no positive samples at baseline	nnn[4]	nnn	nnn	nnn	nnn
Incidence of Treatment Emergent ADAs					
Post-baseline evaluable patients	nnn[5]	nnn	nnn	nnn	
Patients positive for Treattment Emergent ADA	nnn[6] (xx.x%[7])	nnn (xx.x%)	nnn (xx.x%)	nnn (xx.x%)	
Treatment-induced ADA	nnn[8]	nnn	nnn	nnn	
Treatment-enhanced ADA	nnn[9]	nnn	nnn	nnn	
Patients negative for Treatment Emergent ADA	nnn[10]	nnn	nnn	nnn	
Treatment unaffected	nnn[11]	nnn	nnn	nnn	

A programmer needs to work backwards from the output/output specifications in order to create the analysis dataset.

Some of the questions to ask yourself:
What is the output displaying?
What information will be required in the dataset?
Which dataset structure will be most appropriate e.g. subject-level, BDS or occurrence?
What derivations will be required?

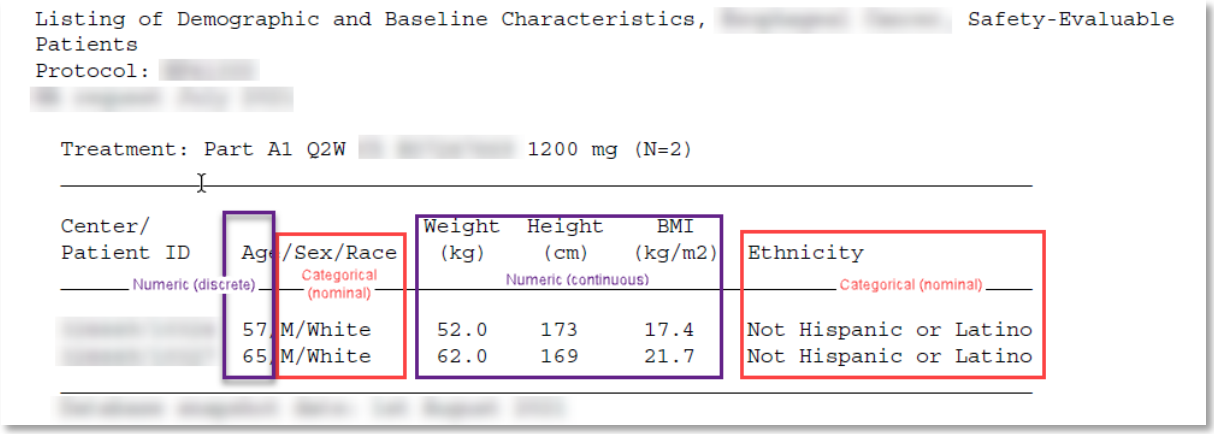
I would strongly recommend reading this paper (J. McGrogan and A.Wittle, 2020) covering ADaM programming for those coming from a SDTM programming background if like myself may be a little apprehensive about creating your first ADaM datasets especially from scratch.

One of the major challenges for me coming from SDTM to ADaM was the derivation programming. Derivations in SDTM are minimal and mainly standard as it is a data tabulation model for collected data but analysis derivations can be far from standard, very involved, requiring numerous calculations/operations and processing.

OUTPUTS

LISTINGS

What is a listing? A listing is a data listing for each individual subject with the variables of interest output. It can be created from an analysis dataset or even SDTMs directly. Below is a simple demographics listing by treatment at doses 1200 and 2100 mg.

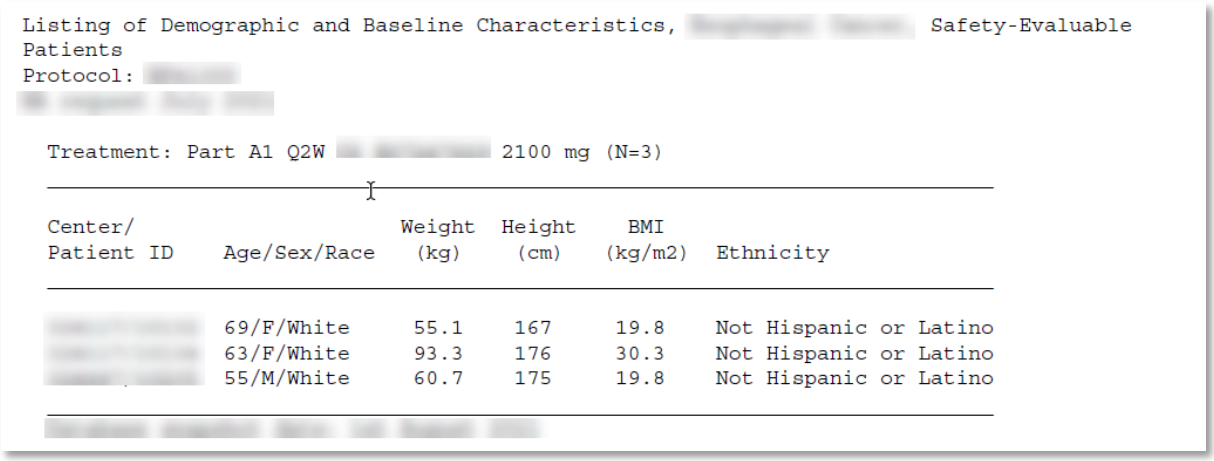


Listing of Demographic and Baseline Characteristics, [redacted] Safety-Evaluable Patients
Protocol: [redacted]

Treatment: Part A1 Q2W [redacted] 1200 mg (N=2)

Center/ Patient ID	Age/Sex/Race	Weight (kg)	Height (cm)	BMI (kg/m2)	Ethnicity
	Numeric (discrete)	Categorical (nominal)	Numeric (continuous)		Categorical (nominal)
[redacted]	57/M/White	52.0	173	17.4	Not Hispanic or Latino
[redacted]	65/M/White	62.0	169	21.7	Not Hispanic or Latino

[redacted]



Listing of Demographic and Baseline Characteristics, [redacted] Safety-Evaluable Patients
Protocol: [redacted]

Treatment: Part A1 Q2W [redacted] 2100 mg (N=3)

Center/ Patient ID	Age/Sex/Race	Weight (kg)	Height (cm)	BMI (kg/m2)	Ethnicity
[redacted]	69/F/White	55.1	167	19.8	Not Hispanic or Latino
[redacted]	63/F/White	93.3	176	30.3	Not Hispanic or Latino
[redacted]	55/M/White	60.7	175	19.8	Not Hispanic or Latino

[redacted]

Looking at the data in the listing, we see four numeric variables; Age, Weight, Height and BMI of which Age is discrete and the rest are continuous variables. The rest are categorical variables; Sex, Race and Ethnicity, to be more specific nominal categorical variables. These variables contain categorical results.

TABLES

What is a table? A table summarises data using descriptive statistics, also known as a summary table. Statistical methods are applied to the particular variables or observations of interest and displayed. To summarise data, subject level detail is lost and not required. An overall picture or view of the data of interest is seen. Typically, there will be multiple

analyses performed and the results of those individual analyses will be appended to create the final table as shown below. Different statistical analyses can be applied to categorical and numerical data.

Demographic and Baseline Characteristics, [redacted] Safety-Evaluable Patients
Protocol: [redacted]

Treatment Group: Part A1 Q2W

	[redacted] 1200 mg (N=2)	[redacted] 2100 mg (N=3)	All Patients (N=5)
Age (yr)			
n	2	3	5
Mean (SD)	61.0 (5.7)	62.3 (7.0)	61.8 (5.8)
Median	61.0	63.0	63.0
Min - Max	57 - 65	55 - 69	55 - 69
Age group (yr)			
n	2	3	5
<65	1 (50.0%)	2 (66.7%)	3 (60.0%)
>=65	1 (50.0%)	1 (33.3%)	2 (40.0%)
Sex			
n	2	3	5
Male	2 (100%)	1 (33.3%)	3 (60.0%)
Female	0	2 (66.7%)	2 (40.0%)
Ethnicity			
n	2	3	5
Not Hispanic or Latino	2 (100%)	3 (100%)	5 (100%)
Race			
n	2	3	5
White	2 (100%)	3 (100%)	5 (100%)

Percentages are based on N in the column headings.

The basic demographics table above contains descriptive statistics. The data is essentially been described using basic statistics. Example of descriptive statistics include; mean and median (measures of central tendency of the data), range, standard deviation and interquartile range (measures of the spread of the data) frequency distributions, proportions/percentages to name a few.

Four analyses are performed on the age (years) for each treatment and a pooling for all treatments i.e. 1200mg and 2100mg. Pooling refers to the pooling of columns and providing analysis statistics for all treatments. The number of patients who were administered a particular dose is the big N in the table heading.The analyses results displayed are the number/frequency of subjects for whom age is available (small *n*), mean, median and the range of ages.

A second analysis is performed on the Age by first creating two categories; age less than 65 and age greater than or equal to 65. Therefore, we now have a categorical binary variable! Then a frequency count is performed on each category. This is a good example of

performing categorical data analysis on a numeric data by first converting the numerical data to categorical.

The subsequent analyses are performed on categorical variables i.e. Sex, Ethnicity and Race.

By summarising the data in this table we can get a good idea of about the characteristics of the subjects recruited in to the trial for each treatment. For example the balance of demographic profile across the treatment groups.

Below we have an efficacy table of responses of cancer patients to the investigational drug in question.

Best Confirmed Overall Response based on Investigator, Efficacy Population
Protocol: [redacted]
IB 2021

Treatment Group: Part A1 Q2W Best Confirmed Overall Response by Investigator

	[redacted] 50 mg (N=4)	[redacted] 150 mg (N=4)	[redacted] 300 mg (N=4)	[redacted] 600 mg (N=4)	[redacted] 1200 mg (N=6)	[redacted] 2100 mg (N=13)	All Patients (N=35)
Responders 90% CI	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 52.71)	2 (50.0%) (9.76, 90.24)	0 (0.00, 39.30)	4 (30.8%) (11.27, 57.26)	6 (17.1%) (7.74, 31.06)
Objective Response Rate (ORR - CR\PR) 90% CI	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 52.71)	2 (50.0%) (9.76, 90.24)	0 (0.00, 39.30)	4 (30.8%) (11.27, 57.26)	6 (17.1%) (7.74, 31.06)
Disease Control Rate (DCR - CR\PR\SD) 90% CI	2 (50.0%) (9.76, 90.24)	3 (75.0%) (24.86, 98.73)	2 (50.0%) (9.76, 90.24)	3 (75.0%) (24.86, 98.73)	1 (16.7%) (0.85, 58.18)	7 (53.8%) (28.70, 77.60)	18 (51.4%) (36.46, 66.21)
Complete Response (CR) 90% CI	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 39.30)	0 (0.00, 20.58)	0 (0.00, 8.20)
Partial Response (PR) 90% CI	0 (0.00, 52.71)	0 (0.00, 52.71)	0 (0.00, 52.71)	2 (50.0%) (9.76, 90.24)	0 (0.00, 39.30)	4 (30.8%) (11.27, 57.26)	6 (17.1%) (7.74, 31.06)
Stable Disease (SD) 90% CI	2 (50.0%) (9.76, 90.24)	3 (75.0%) (24.86, 98.73)	2 (50.0%) (9.76, 90.24)	1 (25.0%) (1.27, 75.14)	1 (16.7%) (0.85, 58.18)	3 (23.1%) (6.60, 49.46)	12 (34.3%) (21.12, 49.55)
Progressive Disease (PD) 90% CI	2 (50.0%) (9.76, 90.24)	1 (25.0%) (1.27, 75.14)	2 (50.0%) (9.76, 90.24)	1 (25.0%) (1.27, 75.14)	5 (83.3%) (41.82, 99.15)	6 (46.2%) (22.40, 71.30)	17 (48.6%) (33.79, 63.54)
Not Evaluable (NE) Missing	0 0	0 0	0 0	0 0	0 0	0 0	0 0

90% CI for rates and responses were constructed using Exact method. Patients were classified as missing or unevaluable if no post-baseline response assessments were available or all post-baseline response baseline assessments were unevaluable.

There are a total of 16 rows in this table which actually correspond to 16 separate analyses. Let's focus on the first two rows of the table; Responders and the row below, 90% CI (confidence interval) and also the 'All Patients' column. For the first row of the table we see 'big' N, which equates to all patients for whom efficacy results are available and thus make up the efficacy population. A total of 6 responders out of 35 i.e. 6 patients (17.1%) had a response to the drug.

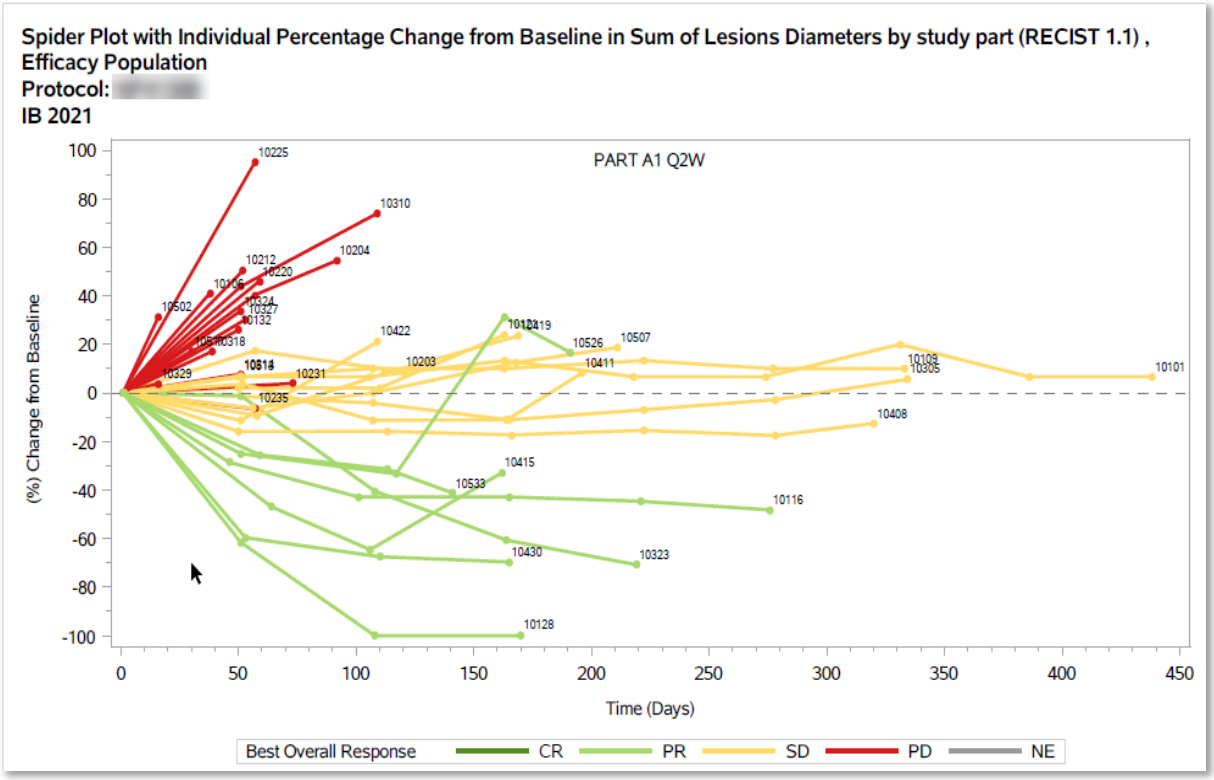
The second row shows the 90% CI which has been calculated using the Clopper-Pearson Exact method. A different type of statistic to what we have seen so far in tables. The intent is not to explain what a CI is but for the purposes of this paper I will give a very brief description so as to be able to explain why it is displayed in this efficacy table. In the trial of 35 patients we have a 17.1% response rate. From this we make an inference, a prediction or likely hood of what the response rate may be in the wider population of patients from which

the 35 patients have been recruited. We cannot say with certainty what the response rate will be but can provide an upper-limit of confidence, 31.06% and a lower-limit of confidence, 7.74% reading from the table. In summary, we are stating that we are 90% confident that in the actual population there would be a response rate of between 7.74 to 31.06%.

Looking further down the table there were no patients that had a ‘Complete Response’ i.e. the cancer completely disappeared but of the 6 that had a response there were all a ‘Partial Response’ i.e. there was some degree of shrinkage of the tumours.

VISUALATIONS

Graphs can display both subject level and summarised data as required. The spider plot below displays the change in the sum of the target lesions by patient. This plot portrays two levels of information; the line plot shows the change in lesion size throughout the trial and the colour of the line represents whether the patient had a complete or partial response or the disease was stable, progressive or not evaluable.



Hopefully this section provides a useful introductory insight in to outputs and how to interpret what you produce.

R, SAS, or <other statistical software>

The mention of programming at this point may seem to undermine the importance of this topic. This is not the case and is certainly not the aim but in my journey as someone having prior experience of programming it has been more important to understand the areas that have been covered so far, essentially the fundamental aspects in the analysis of clinical data before writing code. Ultimately analysis datasets need programming and outputs with statistical analyses need creation. However as a beginner first understand the analysis dataset and the principles of ADaMIG 1.1, understand how outputs are created from simple listings to tables and graphs; what are the components and how are these reports built? Once we understand the basic concepts and principles of reporting objects then it does not matter what tool, software or programming language is used as the fundamental principles are the same no matter what language you are acquainted with.

As mentioned previously the programming methodology is quite different for analysis programming due to the nature of the work. Even if you are comfortable with an existing language or software there will be the need to learn new functions, procedures and programming syntax to increase your skill level. Often you may need to reference other programs written by experienced analysts and understand them. This all helps in challenging you to grow and expand. Concurrently I personally experienced feelings of low confidence in my programming ability. Am I good enough? Will I be able to program competently? The code that I have been provided with is very complex I do not fully understand it. These feelings have been recognised as belonging to ‘Imposter Syndrome’ where the individual thinks they are not good enough for a particular role, task etc. You may feel that you need to prove yourself. It is easy to become very self-critical but do not let such feelings be a barrier. Perseverance and allowing yourself time to learn will help overcome such feelings in my experience.

Resilience

The world of data analysis can be quite dynamic with requests coming from either health authorities or internal requests from the clinical team or statisticians which materialize due to questions been asked of the data. These requests often require a fast turnaround as health authorities are not know for providing much slack in their timelines! As a data scientist this means being able to chop and change from tasks ‘at the drop of a hat’ when situations require. One could be in the depths of programming a complex analysis dataset that then needs to be dropped to pursue other higher priority requests. In fact, there could be another higher priority request coming after that or multiple requests. This environment is quite different from data management. Having a flexible and dynamic attitude or at least developing one will help deal with these situations and build resilience. Initially this was quite a shock but been mentally prepared for any eventuality that arises helps. You never know what request may end up in your inbox on a Monday morning or Wednesday afternoon!

Project Management
The basic tenets of project management are important and play a key role as a Data Scientist:

Risk Management
Leadership
Time Management
Communication
Stakeholder Management
Resource management

This is key in order to be able to plan and deliver for reporting events. Stating the obvious but keeping on top of tasks as much as possible is always preferable and knowing key milestones. It enables preparation and readiness so that you in the best position to handle reporting events when they occur. Delegation of work and knowing your team's strengths will enable for efficient delegation if required. On a phase 3 study you will be working within a large team but on earlier phase studies there may only be yourself and one other so the responsibility of project management may well fall on your shoulders very quickly. Keeping stakeholders and managing expectations is equally important and can be tough at the best of times but even more so when you just starting out and building your experience and confidence, as you also need to build the trust and confidence of your stakeholders in your capabilities.

Even if you are not managing a project these principles apply as much to ourselves as well as to the project. It is necessary for everyone to be able to communicate effectively and to manage our stakeholders, manage our own time and resource and any risks to delivering on a task!

Growth mindset
In order to thrive and develop a growth mindset is critical. The fact you may be considering learning data analysis is a great start and along the way, there will be challenges. See these challenges as opportunities to learn, absorb knowledge and gain experience…be curious. Each challenge, each reporting event will make you a better and stronger data scientist. Challenging situations, timelines and reporting events will occur but even during quieter times I advise giving yourself challenges e.g. developing your programming skills further, learning a new programming language or seeing how processes can be improved and gaps which you have noticed during particular reporting events. Volunteering for a particular role or an initiative even if you are not an expert. It will be a great way to learn. A personal example was that I volunteered to perform QC on some analysis dataset programming and outputs. I was not aware of some of the complexities and the background to the programming but it was an opportunity to learn and I was curious to know more about the analysis been performed that was non-standard. In hindsight as a new Data Scientist, it may not have sounded like a good idea with little experience. Nevertheless, curiosity and questioning why something is been done is key to QC and not accepting the specification at face value. I learned a great deal from this task and eventually had to implement the analyses on my study.

Conclusion

I hope it has given you an introduction to data analysis and given you an idea of what data scientists do and encourages you to explore more and perhaps even take the leap as I did. It is a steep learning curve going from a data tabulation to a data analysis programmer than vice-versa, as analysis programming is the next step up from data tabulation. With training, support and a little hard graft it's amazing how quickly you will learn. There will be bumps along the way, ups and downs as with any career move but resilience, passion and a growth mindset will help you progress on the way.

References:

1) KPMG International, Pharma outlook 2030: From evolution to revolution
2) CDISC ADaM Team, 2009, Analysis Data Model Implementation Guide Version 1.1
3) Jennifer McGrogan and Alyssa Wittle, Covance, Inc.,SDTM to ADaM Programming: Take the Leap! PharmaSUG 2020 - Paper DS-306

Contact information

Your comments and questions are valued and encouraged. Please contact me at

Shuva Dey
Roche Products Ltd
Hexagon Place
Shire Park
Welwyn Garden City
AL7 1TW

shuva.dey@roche.com