PharmaSUG 2019 - Paper SS-334

The Anatomy of Clinical Trials Data: A Beginner's Guide

Venky Chakravarthy, BioPharma Data Services, Ann Arbor, Michigan

ABSTRACT

We are so caught up in our own work that we sometimes lose sight of the big picture. It is worthwhile to look at the Drug Discovery and Development process at a high level. This will serve as a refresher of the complexities at various stages and highlight the role of SAS® programming in the wider context of this important human health field. An audience that wants to learn more about clinical trials will appreciate this comprehensive introduction to the role of a SAS® programmer in Clinical Trials Phases I-III. The presentation will cover the role of Clinical Data Interchange Standards Consortium (CDISC) and the implications for capturing and reporting clinical trials data. Those who have just begun their careers in the Pharmaceutical industry will also benefit from attending this presentation. If you have been working in the Pharma industry for some time this will be a refresher and could lead you to discover hidden gems. The presentation will begin with an introduction to Human Clinical Trials. A short history of the evolution of standards in clinical trials will be provided. The talk will be a microcosm of a clinical trial study with greater emphasis on the role of a SAS programmer. It will cover the study protocol, eCRFs (capture of data) and SAP (plan to analyze data). There will be a greater focus on how the eCRF data is standardized to form the Study Data Tabulation Model (SDTM) and a further refined dataset model for analysis which is called the Analysis Data Model (ADaM). This will eventually lead to an appreciation of how such standardization leads to a more streamlined production of study related outputs such as Tables, Listings and Figures (TLF).

INTRODUCTION

In this gentle introduction to the Pharmaceutical world, you will learn about the various stages of drug development. You will also get to know the complications in drug development and the chances of a drug receiving Food and Drug Administration (FDA) approval. This is a US specific regulatory agency, but the process is similar for other agencies like the European Medicines Agency (EMA) or the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan. You will then learn about the Human Trials stage where the bulk of SAS Programming occurs. At this stage, you will get to know the types of documents that a programmer needs to familiarize. Next comes SAS data followed by a review of the evolution of standards and the current data standards. You will get an appreciation of these data standards in analyzing data and generating SAS reports.

THE PHARMACEUTICAL DRUG DEVELOPMENT PROCESS

Let us start at the very beginning and understand how a drug goes through the Discovery and Development process. Let us look at the complications involved and the chances of successfully getting an FDA approval from the time the drug was conceived. The drug development process is shown below in Figure 1. At the far left of the figure is the Basic Research process which studies a disease of interest to identify a target to attack the disease. Then Scientists develop a molecule to hit the target in the Drug Discovery phase and advance one or more molecules for the Pre-Clinical testing phase. In this stage, the drug is extensively tested in Lab and animal models to determine whether it is safe for humans. Once the toxicity levels are safe enough to advance to humans, the company presents the evidence to the FDA and files an Investigational New Drug (IND) application. If the FDA gives the go ahead, the drug proceeds to human trials. Phases I through III have different objectives in a continuum. In Phase I, a few healthy volunteers are administered the drug to test safety and to get an idea of a safe dose range. If the results are good, the drug moves to Phase II where it is more extensively studied in patients having the targeted disease for which the drug is tested. Typically, the goal at this stage is to continue monitoring safety and seek answers to the drug's effectiveness against Placebo and/or a drug considered the current Standard of Care. In this Phase II stage, they further tune the dose strength to try to get the ideal mix of safety and efficacy. The final stage before seeking FDA approval is Phase III. This is a very expensive stage of development with thousands of patients, hundreds of clinics and hundreds of millions of dollars.

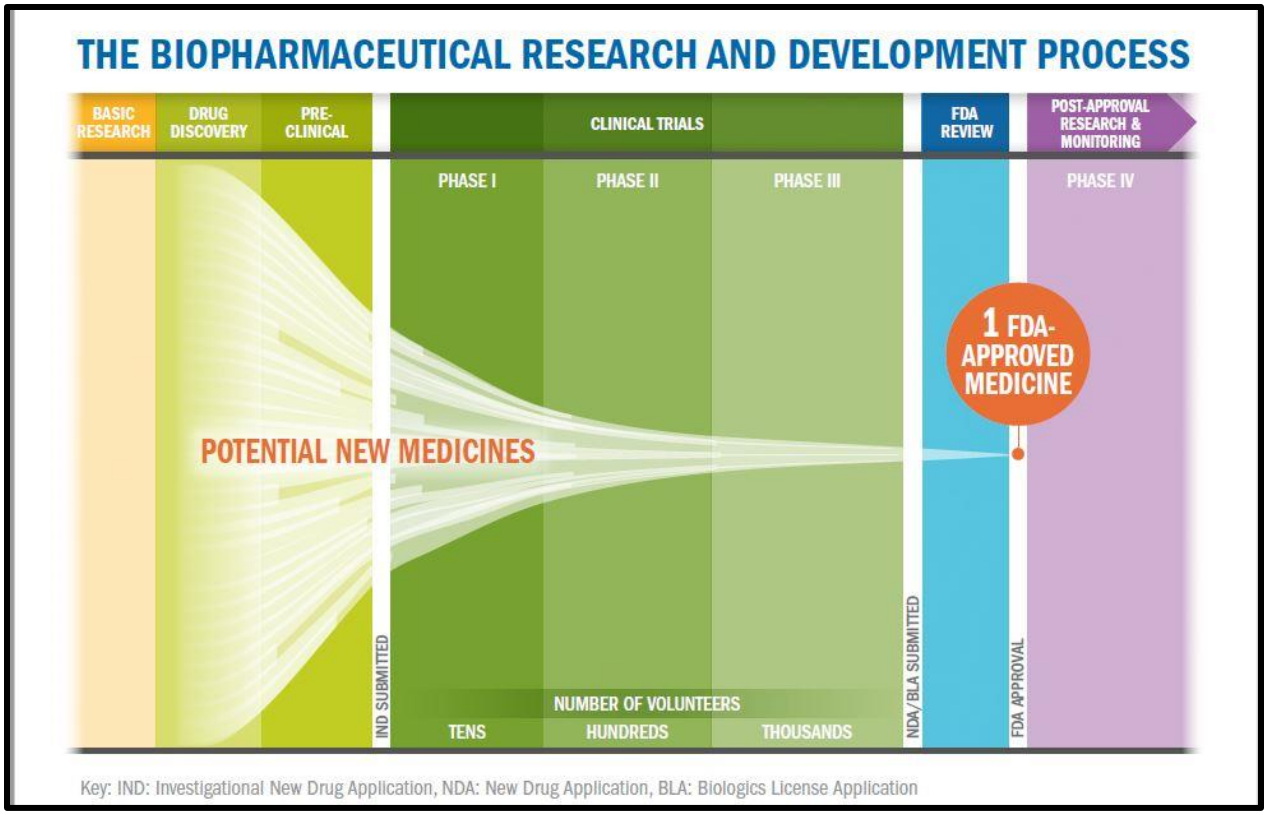

Figure 1 Research and Development Process, Source: PhRMA, 2015

Considering the high cost of Phase III, drug candidates are carefully evaluated after Phases II. Is it worth the risk to advance to Phase III considering the overall failure rates of drug development? Speaking of failures, you are probably aware that the drug industry has an extremely high rate of failure. It is worth talking about it a bit. See Figure 2 below and focus on the rectangle outlined:

	DISCOVERY	PRECLINICAL	PHASE I	PHASE II	PHASE III	OVERALL
Time per stage	3.5	0.46	1.4	2.3	5.1	12.9
Chance of failure	0.4	0.35	0.22	0.3	0.1	0.984
Progression probability	0.5	0.5	0.4	0.4	0.15	
Chance that one project entering this stage will eventually successfully leave it	0.58	0.52	0.42	0.52	0.44	
Number of projects needed to achieve one successful launch	30	13	6.2	3.6	1.7	

Figure 2 Development times by Phase and Risk of failure, Source: Drug Discovery World 2004

As seen above it takes roughly four years before the drug enters the human trials phase. Although this data is a bit dated from 2004, the overall development of a drug has stayed remarkably consistent around 10-12 years. Do take note that the risk of failure is 0.984. That is not a misprint. After accounting for the sunken costs in this high rate of failures, the cost of a single drug from Discovery to Launch is estimated to be a staggering $2.6 Billion (Grabowski and Hansen, 2014).

If you consider the approval rate after the drug reaches human trials, it is roughly 10 percent of all drugs that reach the market. The success rate by each Phase in human trials is presented in Figure 3:

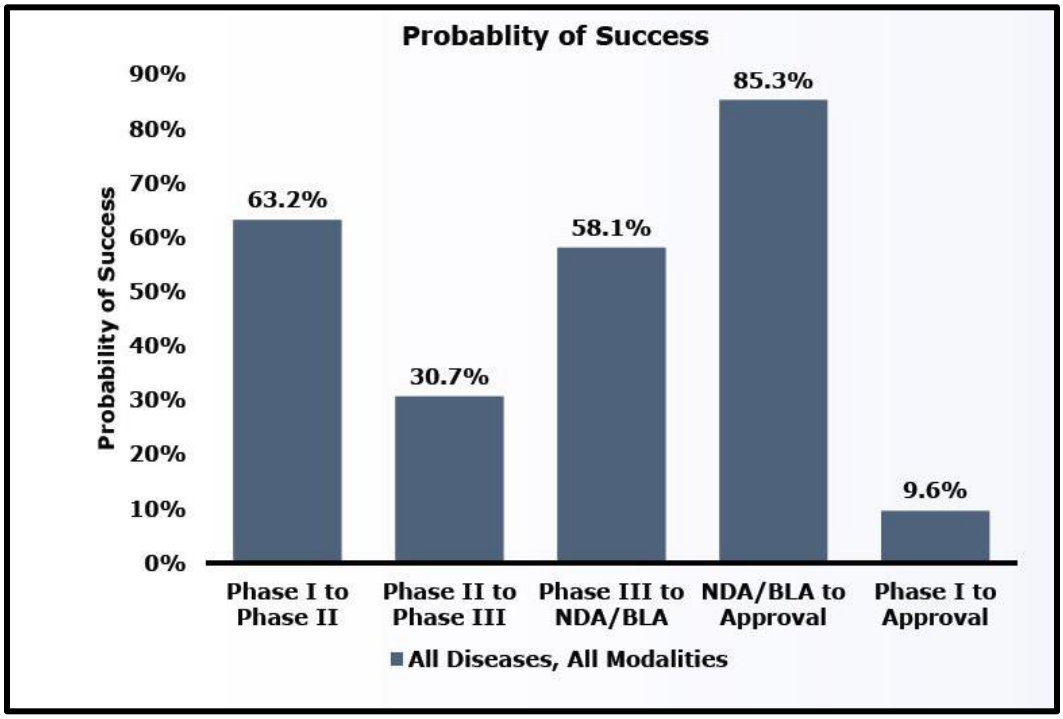

Figure 3 Probability of Success from the Start of Human Trials, Source: BIO Industry Analysis (2006-2015)

The above figure is largely self-explanatory. You may have noticed that the Phase II to Phase III transition probability is just 31 percent. You were made aware earlier that the cost of Phase III is prohibitive and unless a company considers the Phase II results as quite promising, they are unlikely to move it further along. Once the drug passes Phase III, the company will file a New Drug Application (NDA) for a synthesized drug formed from chemical entities. A common type of pill such as Ibuprofen is a chemical entity. Otherwise, if there are living organisms in the drug a Biological Licensing Application (BLA) is filed. You may think of vaccines as a BLA candidate. The FDA reviews all the evidence and decides whether to approve the drug to enter the market.

The above discussion should give you a basic understanding of the complications, rigor, risks and the length of drug development. It is time to move on to what happens during these human trials and what a SAS programmer does. Let us start with Phase I human trials. There are many documents that a programmer deals with during these Trials. We will not cover them all but there are three critical documents that a program needs to be very familiar – **(1) Study Protocol, (2) Case Report Form and (3) Statistical Analysis Plan**. Let us now look at these documents a little closer.

HUMAN CLINICAL TRIALS

When a company receives FDA's approval to conduct trials in humans, the drug begins another lengthy phase to reach the market. The first objective is to establish the safety of drug in normal and healthy volunteers. These are your Phase I studies that use a small number of healthy subjects. Phase II studies continue with safety assessment and fine tune the drug dose strength. Phase III studies are conducted on large sample sizes. Statisticians carefully evaluate the sample size sufficient to represent the incidence rate of the disease in the general population.

To carry out a study, there are many components and functions that need to come together. The sponsoring company puts together a detailed study plan called a **Study Protocol** (mentioned above) with the help of Clinical Scientists, Statisticians, Programmers, Data Managers and Medical Writers all of whom contribute to the development of the Protocol.

STUDY PROTOCOL

A typical protocol begins with the study title. Programmers are concerned primarily with the study design and methodology, the primary end points, the visit schedules of patients etc. A sample title of a Protocol is below:

DOUBLE-BLIND, PLACEBO-CONTROLLED, DOSE ESCALATING, SAFETY AND PHARMACOLOGY STUDY WITH THREE DOSAGES OF GBR 12909 IN COCAINE EXPERIENCED VOLUNTEERS

Protocol No: NIDA-CPU-0002
IND No. 57,007

Figure 4: Title in a Study Protocol. Source: ClinTrials.Gov, NIDA study, Version 1, Date: April 3, 2002

The Study Protocol is a thorough document that encompasses multiple function areas. The SAS programmer will focus mainly on the Study Design, the analysis methods and the Patient Visit Schedule. The Study Design in this Protocol is presented in Figure 5

5 STUDY DESIGN

This is a single dose with escalation, double-blind, placebo-controlled inpatient study, in which 24 cocaine experienced volunteers that meet the protocol eligibility criteria during a 30 day (maximum) screening period will be randomized into three dose groups (n=8 each). In each group, 6 subjects will be randomized to receive a single daily dosage of 50, 75 or 100 mg of study agent and 2 will receive matched placebo for 11 days. Each cohort of 8 subjects will

Figure 5: Study Design from the Protocol for GBR 12909. Source: ClinTrials.Gov, NIDA study Version 1, Date: 3 April, 2002

Although the SAS programmer needs to be familiar with the protocol, the key sections are the Study Design and the plan for analysis. However, before you analyze any data, let us see how that data are

captured and what type of a data structure can you expect. The data collection plan is traditionally known as a **Case Report Form (CRF)**.

CASE REPORT FORM

We now commonly use the electronic version which is called eCRF. A sample is provided below in Figure 6:

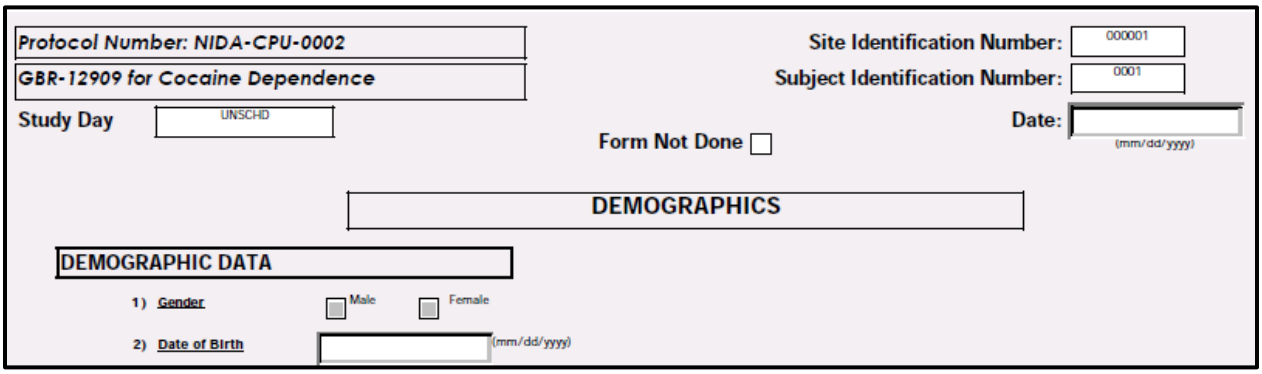

Figure 6 Collection of Demographics information

The above is one of the first forms used to capture the demographic information. What you see above is a partial form. Some fields are pre-populated such as the study number and the drug name on the top left. The date of capture is entered on the top right. The actual demographic information collection starts below the DEMOGRAPHICS banner. There is obviously much more information collected at the first visit called **Screening Visit**. This is also the visit that determines whether a patient qualifies for this study. Let us next look at the form that captures the side effects of the medication in Figure 7.

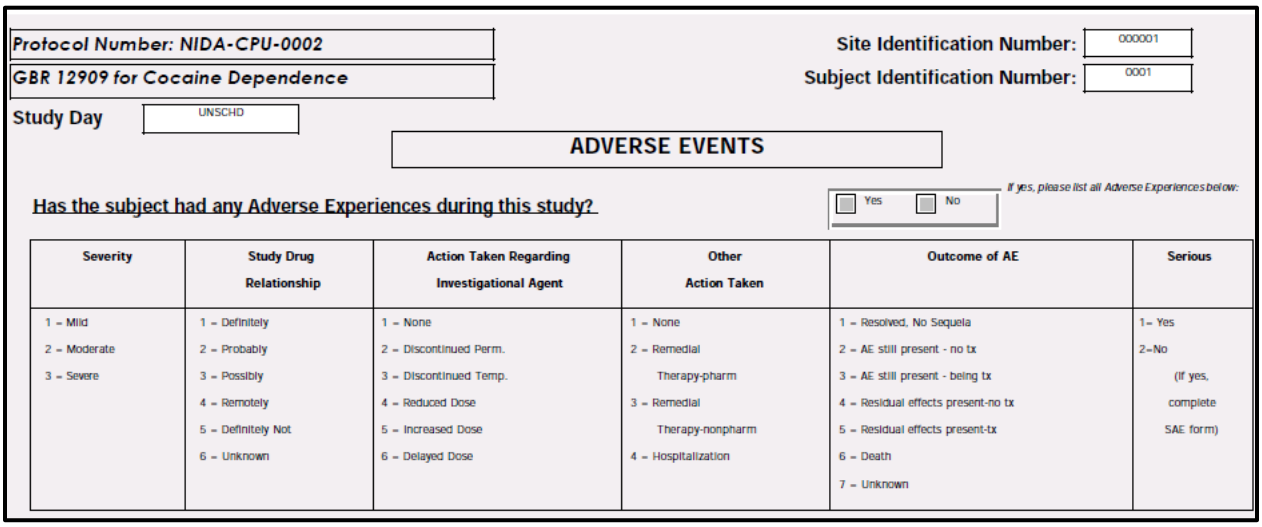

Figure 7 Collection of Adverse Events

What you see above is the data collection form for any **Side Effects** during the study period. In scientific terms these are called **Adverse Events** (AE). Programmers refer to them by the acronym AE. This form captures whether there were any AEs, how severe, was it related to the drug, how did they treat the AE, was there any hospitalization? Was the study drug stopped? What was the outcome – meaning was it resolved? These are some of the information collected. The specific side effect is captured and then coded to a standardized medical terminology. As an example, an investigator at the site may record an Adverse Event as "**Retinopathy due to diabetes**". The standardized Preferred Term in the Medical Dictionary for Regulatory Activities (MedDRA) dictionary is "**Diabetic retinopathy**".

Each data collection form or module is called a domain. There are many domains in any given study. At the time of the Screening visit, a patient needs to qualify to be in the trial. The protocol specifies many conditions that a patient must meet. This is called the **Inclusion Criteria**. There are also many conditions that a patient should fail. That means if a patient meets those **Exclusion Criteria**, then that patient is excluded from the trial. Besides all the above domains, there are a few common data domains. Let us first take Drug Accountability– this means dispensing a certain number of pills at a visit that will sufficiently cover the patient until his next visit. A few more are always provided to account for losses through spills etc. This form also includes collecting back remaining unused pills from the previous visit. Accounting the usage of the medication is critical to determine compliance to the dosing regimen. This information allows the sponsor to aim for the best dose that has the greatest compliance. You can have the greatest drug but if patients do not take as prescribed, it is not that great. Another module is Medical History. Sponsoring companies want to know what a patient's medical history is as far back as they can remember. This is critical even in your own experience with your doctor visits. Sponsors also want to know what medications the patients have taken in the past and continue taking now. This data domain is called Prior and Concomitant Medications. At any given scheduled visit to the clinic, blood and urine samples are normally collected to run a battery of Lab tests. This domain is appropriately called Labs. Likewise, as you see in your own care by your provider, Vital Signs such as your pulse and blood pressure are measured at each visit.

The blank CRFs that you have seen above do not offer you as a SAS Programmer an idea of the variable names involved. This is the reason CRFs get annotated. These annotations become variable names in the Oracle Clinical Database. Let us look at an annotated sample CRF in Figure 8. All the texts in red are the annotations and these become your variables in the electronically captured data in the Clinical Database.

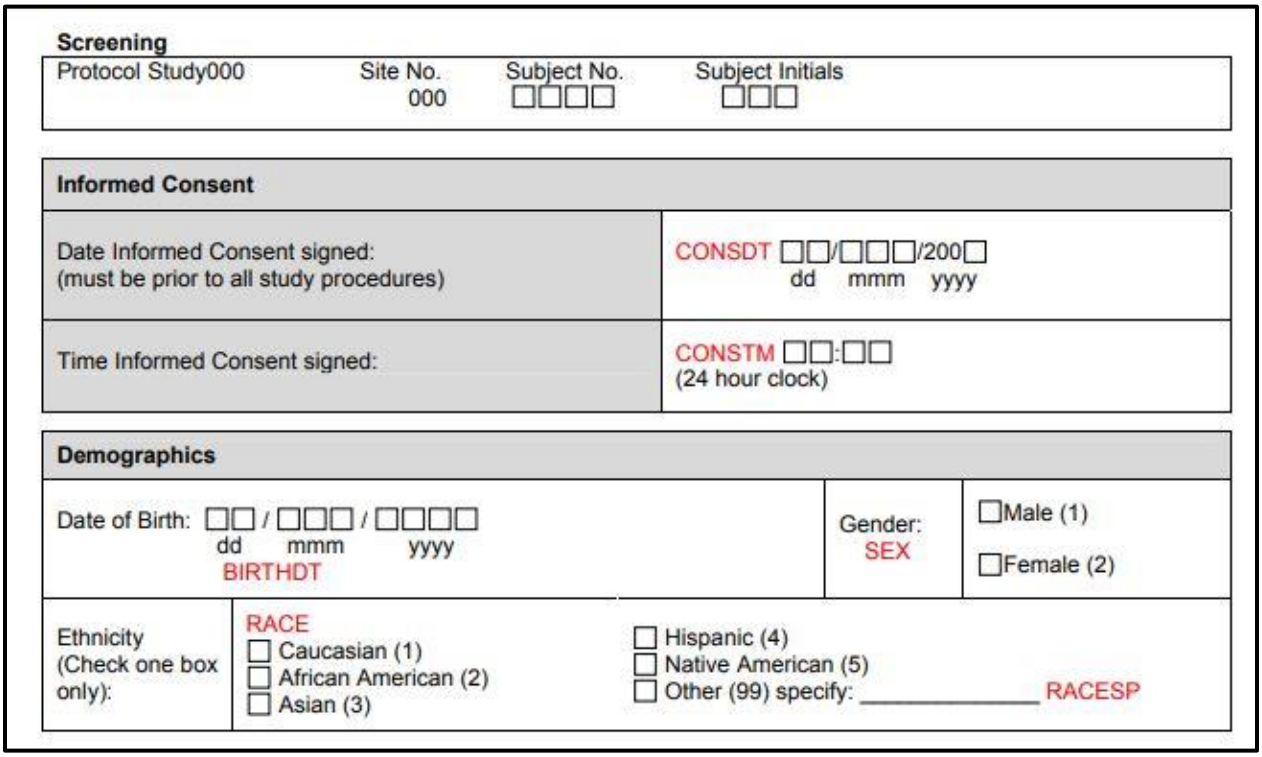

Figure 8 Sample Annotated CRF for Variable Names

You were introduced to two of the most important documents that a programmer deals with – the **Study Protocol** and the **Case Report Forms**. We will deal with the third document – **Statistical Analysis Plan** a little later when we discuss SAS data.

We have seen the gathering of data above. Before we look at where that data goes and how it eventually makes it into SAS data sets, we need to understand the practices that drug development companies follow for quality assurance.

LARGE SHADOW OF REGULATORY AGENCIES

The Regulatory Agencies scrutinize every part of the drug development process to assure patient safety and to ensure proper conduct of trials. A company needs to have quality systems and practices in place that are acceptable to the Regulatory Agencies. These practices ensure quality, integrity, reproducibility, traceability and accountability. The study conduct, data collection, data storage, drug manufacturing, laboratory sample collection and analysis etc. follow Good Clinical Practices (GCP). Specifically, the manufacturing of drugs needs to follow Good Manufacturing Practices (GMP). The Laboratory sample collection and analysis needs to follow Good Laboratory Practices (GLP). These Good Practices are commonly referred as GxP during the conduct of the trials.

The data from multiple domains above (Demographics, Adverse Events etc.) are electronically captured into a database. Oracle Clinical is one of the most common Relational Data Base Management Systems that manages clinical trials data. Any electronic form of data needs to follow regulatory guidelines commonly referred as 21 CFR Part 11 to maintain integrity and an audit trail (CDER, 2003).

From our perspective, we need to concern ourselves with what happens next to the data conforming to the regulatory standards. The variables in Oracle Clinical for each domain are dumped to SAS datasets bearing the same domain name. For example, Demographics may be captured as DEMO in Oracle and written out to a SAS dataset DEMO. This is repeated for all CRF forms. While there is some structure to this dataset there is still no defined Standard as each company has unique requirements with their data collection approach. If the end goal is seeking approval to market, the Regulatory Agencies can benefit from seeing and evaluating a standard approach to seeing patient data. This became a driver in the founding of the Clinical Data Interchange Standards Consortium (CDISC). We will now look in a bit more detail about those standards.

STUDY DATA TABLULATION MODEL (SDTM)

CDISC began as a volunteer effort by a collection of individuals from across the pharmaceutical industry. It has since become a non-profit organization. It continues to grow to cover a broad range of data standards for Drug and Medical Devices Development and Healthcare in general. We will next explore the Study Data Tabulation Model (SDTM) that represents the collected data in a standardized way. Let us focus on a single data domain to facilitate continuity from the Raw data through the production of outputs. We will work with Lab data. The lab data on any patient is huge so let us just restrict this further to a single test Triacylglycerol Lipase measured in U/L. “Lipase is a protein (enzyme) released by the pancreas into the small intestine. It helps the body absorb fat. This test is used to measure the amount of the lipase in the blood.” (MedlinePlus, 2018).

Below are the key Oracle Clinical fields data dump to SAS relevant to our discussion (see Figure 9). The SUBJID and the SUBJECT fields are deidentified to protect patient’s privacy. The VISIT value “Day 1” means the first time the patient was given study drug. Subsequently, the patient came to the clinic three more times - 8 weeks after first dose and then 16 and 24 weeks after first dose. Notice the highlighted

columns on the last row. Compared to the previous visits, the Lab test results returned a very high value for Week 24 in the standard results column (STDRESN). The meaning of this result is interpreted through Toxicity grades that are graded through Common Terminology Criteria for Adverse Events (CTCAE). This is another look up table. Toxicity grades range from 1 through 5. The least toxicity is at value 1 whereas 5 means death. Toxicity grade 4 returned here is quite severe that would require hospitalization.

	SUBJID	SUBJECT	VISIT	LBTEST	STDRESN	STDUNIT	TOXGR	TOXGRDSC
1	10101-10001	Pharma-2018-10101-10001	Screening	Lipase	31	U/L	0	
2	10101-10001	Pharma-2018-10101-10001	Day 1	Lipase	22	U/L	0	
3	10101-10001	Pharma-2018-10101-10001	Week 8	Lipase	35	U/L	0	
4	10101-10001	Pharma-2018-10101-10001	Week 16	Lipase	36	U/L	0	
5	10101-10001	Pharma-2018-10101-10001	Week 24	Lipase	587	U/L	4	Lipase increased

Figure 9 Raw LAB data, Sample Data for Lipase

Next, we look at how this captured data is standardized as per the requirements of SDTM. This is one of the first places where a SAS programmer is involved. There are dedicated teams to convert raw datasets to SDTM. At first glance, the SDTM data below (see Figure 10)appears just like the raw data with no additional fields other than the LBTEST field. However, there is more to it. Notice that the SUBJECT field is now called USUBJID. This is a requirement of the SDTM standards. It is a combination of the STUDYID (Pharma-2018), Investigator Site number (10101) and the Subject number (10001). One of the simplest SDTM rules is that the domain specific variable names begin with the domain name. Notice that the values for this Lab domain all begin with LBxxxxx to satisfy this rule. Variable names are restricted to 8 characters. The values of character variables are restricted to 200 characters. This is because the regulatory agencies like FDA need to maintain vendor neutrality. The only SAS transport data that is in the public domain is in version 5 format, an old SAS version, which has these limitations. More rules and details are in the SDTM Implementation Guide (SDTMIG, 2013). The below dataset is a very simple representation to keep the focus to a narrow set. In real trials there are usually 100 plus Lab tests, date fields (LBDTC), Results collected in Original units (LBORRES and LBORRESU), normal ranges LBNRHI and LBNRLO, and many more variables. There is also a Supplemental Qualifier dataset for most of the domains. For this LB dataset there will be an accompanying SUPPLB dataset that hosts the rest of the raw variables which may or may not be required for analysis.

	USUBJID	SUBJID	LBTEST	LBSTRESN	LBSTRESU	LBTOX	LBTOXGR	VISIT
1	Pharma-2018-10101-10001	10101-10001	Triacylglycerol ...	31	U/L		0	Screening
2	Pharma-2018-10101-10001	10101-10001	Triacylglycerol ...	22	U/L		0	Day 1
3	Pharma-2018-10101-10001	10101-10001	Triacylglycerol ...	35	U/L		0	Week 8
4	Pharma-2018-10101-10001	10101-10001	Triacylglycerol ...	36	U/L		0	Week 16
5	Pharma-2018-10101-10001	10101-10001	Triacylglycerol ...	587	U/L	Lipase increas...	4	Week 24

Figure 10 SDTM Lab dataset LB for Lipase

We have seen how the electronically captured data are converted to the SDTM standard. You would have noticed that we have, thus far, avoided the topic of how we are going to analyze all the collected data. This is a good time to introduce the third document **Statistical Analysis Plan (SAP)** that a SAS programmer needs to be familiar.

STATISTICAL ANALYSIS PLAN AND REPORTS

Recall that you were introduced to the **Study Protocol** and the **Case Report Form (CRF)** earlier. The **SAP** is the third and the single most important document for a programmer producing the standard Analytical Data Model (ADaM) datasets and reports in the form of Tables, Listings and Figures (TLF). The

SAP contains important elements from the protocol but focuses mainly on the goals of analysis to examine the primary and secondary Efficacy end points and assess the Safety of the study drug. The SAP clearly defines what those end points are and how to measure them. If the measurement involves arriving at a complicated score, the SAP provides guidance for the programmer to implement the algorithm. The SAP also provided rules for handling missing data. For example, if a patient provides a partial date (2018-03) for the start of an Adverse Event, the usual imputation rule would default to a conservative estimate and assign it to the start of the month. The imputed date is 2018-03-01. The SAP also sets the rules for arriving at baseline values. For example, a patient may have missed the study Baseline visit but an earlier but recent screening record could be available to use as baseline. This method of imputation is called the Last Observation Carried Forward (LOCF). Another important detail is the population of analysis. A hundred patients may be randomized but one patient may have never taken the study drug. In this case the Randomized population is 100 and the Safety population is 99. There are many other such minute details contained in the SAP. Next comes the crux of the SAP – the statistical models to be used for analyzing the end points. Finally, all supporting reports in the form of Tables, Listings and Figures (TLF). Figures are also referenced as Graphs in which case the acronym becomes TLG. Presented below are some key sections of a study SAP. We will start with a sample study design in Figure 11 of a study on Serbian Smoking Reduction/Cessation Trial using a smokeless tobacco. This is found in the FDA repositories and the SAP sections below are extracted from this study.

1.2 Study Design

This is a multicentre, double-blind, placebo-controlled, randomised clinical trial designed to evaluate snus versus placebo as an aid to reduce smoking among adult cigarette smokers in Serbia.

Figure 11 Study Design from a SAP. Source FDA, 2010

This below part of the SAP in Figure 12 partly captures what happens with the patient at the Baseline visit.

Baseline – Clinical Visit

- Review of Inclusion and Exclusion Criteria
- Informed consent must be obtained and a subject number assigned prior to any evaluations being performed
- Demography

Figure 12 Partial Details of the Baseline Visit

The next item of importance is the type of analysis population. This SAP discussed three types of population – (1) Intent to Treat (ITT), (2) Modified Intent to Treat (MITT) and (3) Safety population. Below is a screen capture of the definition of ITT in Figure 13:

Intent-to-Treat (ITT) Population

The ITT population is defined as all randomised subjects, regardless of when they withdrew from the study. All subjects at Site 04 will be excluded from the ITT population. The ITT population will be used to present all the efficacy data (including the primary efficacy endpoint) by randomised treatment group. Subjects will be summarised according to the treatment to which they were randomised, regardless of which treatment they actually received.

Figure 13 Population of Analysis Defined

The SAP also provides guidelines for the programmer to derive variables of importance (see Figure 14). Captured below are the definitions of Age and Body Mass Index (BMI). We may think that Age is simple to calculate but the degree of precision is extremely important. Is age the difference from Date of Birth to the day the patient was Screened, Randomized or the first dose. These are the things that a SAP must clarify.

3. DEFINITIONS AND DERIVED VARIABLES

3.1 Demography and Baseline Characteristics

Age. Age will be calculated using the Date of Birth (DOB) and the date of the Baseline visit, and presented as age at last birthday as an integer.

Age = Integer part of [(Date of Baseline visit – Date of Birth) / 365.25]

Body Mass Index (BMI). BMI is the subject's body weight in kilograms divided by the square of the subject's height in metres.

$$\text{BMI} = \text{Weight in kilograms} / (\text{Height in metres})^2$$

Figure 14 Derivations of Variables

The SAP also clearly defines primary and secondary endpoints. Captured below is the primary endpoint in Figure 15:

4.1 Primary Efficacy Endpoint

The primary efficacy endpoint is the achievement of "smoking reduction" at Week 24. Smoking reduction is defined as a self-reported reduction of $\geq$ 50% compared to Baseline in the average number of cigarettes smoked per day during weeks 21, 22, 23 and 24, verified by a reduced concentration of CO in exhaled air of at least 1 ppm at Week 24 compared to Baseline. The primary efficacy endpoint will be analysed using a binary goal attainment variable (smoking reduction achieved [yes/no]).

Figure 15 Primary Endpoint of the Study

We have seen a few things from the SAP to give us an idea of its importance to the analysis of the study. The final part of the SAP that is of great relevance to the programmer is the section on outputs collectively called TLFs. The table numbering satisfies regulatory requirements. Let us see a few of these outputs described. See **Figure 16** below :

7.2.1 Demographic and Background Data

Table 14.1.1	Summary of Subject Disposition, by Treatment and Overall (All Subjects)
Table 14.1.2	Summary of Final Status and Reason for Withdrawal, by Treatment and Overall (All Subjects)

7.2.2 Efficacy Data

Table 14.2.1.1	Summary and Analysis of Proportion of Subjects who Achieved Smoking Reduction at Week 24 (ITT Population)

Table 14.2.7.8.1	Summary of Laboratory Parameters by Treatment, Visit and Overall (ITT Population)

7.2.3 Safety Data

Table 14.3.1.1	Summary of Adverse Events by Treatment and Overall (Safety Population)

Figure 16 Sample of Outputs for the Study

In the previous few sections we have seen the derivation of variables and the request for outputs. This brings us back to the data. Recall that we standardized the electronically captured data to SDTM. Go back to that section and refresh if necessary. Recall also that the derivation of variables need to happen before the planned outputs are programmed. It is time to introduce the next data standard for the analysis variables.

ANALYSIS DATA MODEL (ADaM)

The data model which facilitates the analysis of the study data is called the Analysis Data Model. We mentioned this briefly in the beginning. It is commonly referenced by the acronym ADaM. We will also go back to the SDTM dataset LB and see how variables are added to facilitate the production of **Table 14.2.7.8.1 Summary of Laboratory Parameters** that we framed in **orange** above in the list of outputs. You may recall that we have been following a single patient with a single lab test. Please refer to the section on SDTM data if you need to refresh your memory. Next, see below Figure 17 for a sample of the ADaM dataset ADLB. Notice how the BASE value for the lab test Lipase is attached to all the visits. Also, the variables CHG (change from baseline) and PCHG (percent change from baseline) are attached to facilitate producing the output.

	SUBJID	VISIT	PARAM	AVAL	BASE	CHG	PCHG	ATOXGR	ATOX	SHIFT1
1	10101-10001	Screening	Triacylglycerol Lipase (U/L)	31	22	.	.	0		
2	10101-10001	Day 1	Triacylglycerol Lipase (U/L)	22	22	.	.	0		
3	10101-10001	Week 8	Triacylglycerol Lipase (U/L)	35	22	13	59.090909091	0		
4	10101-10001	Week 16	Triacylglycerol Lipase (U/L)	36	22	14	63.636363636	0		
5	10101-10001	Week 24	Triacylglycerol Lipase (U/L)	587	22	565	2568.1818182	4	Lipase increased	Grade 0 -> Grade 4

Figure 17 ADaM dataset ADLB for SDTM dataset LB

The basic tenet of the ADaM dataset is that the variable values are traceable to the source SDTM and maintain a specific Basic Data Structure. All ADaM datasets begin with ADxxxxxxx. The dataset and variable names are restricted to 8 characters just as with SDTM data. The same rules apply for the width of character variable values. There is an ADaM Implementation Guide (ADAMIG, 2009). This is similar to the SDTM Implementation Guide in laying out the rules of the ADaM dataset. The resultant dataset should allow a programmer to produce outputs with the least amount of further manipulations. Ideally the final output is a DATA or SQL step and a Statistical PROC away. As you can see, the above ADLB dataset is basically ready to produce the numbers in the Summary of Lab Parameters output seen in Figure 18 below.

	Treatment Group A (N = xx)	Treatment Group B (N = xx)
LIPASE		
Baseline		
N	xx	xx
Mean (SD)	xx (xx.x)	xx (xx.x)
Median	xx	xx
Q1, Q3	xx, xx	xx, xx
Min, Max	xx, xx	xx, xx
{Week 8}		
N	xx	xx
Mean (SD)	xx (xx.x)	xx (xx.x)
Median	xx	xx
Q1, Q3	xx, xx	xx, xx
Min, Max	xx, xx	xx, xx
Change from Baseline at *{Week 8}*		
N	xx	xx
Mean (SD)	xx (xx.x)	xx (xx.x)
Median	xx	xx
Q1, Q3	xx, xx	xx, xx
Min, Max	xx, xx	xx, xx

Figure 18 Sample of Summary of Lab Parameters

A simple PROC MEANS or SUMMARY can be output with all the statistics that we need for the above output. Another couple of steps to manipulate the formats and we have a publishable report after that. It goes without saying that this is a very simple output to aid in the understanding of how data are gathered, organized and eventually produced into a report. There are many graphical reports and patient level listings that are also produced. It is not uncommon to see two hundred plus TLFs for a single study.

CONCLUSION

In this paper you saw a mini and simplified version of how a drug is born and the role of a SAS programmer. Briefly let us recapture what we did.

1. We first walked through the stages of the drug development process at a very high level.
2. Next, we briefly reviewed the length of time from discovery to market, the prohibitive costs and extreme risks involved in drug development.

3. We then moved our focus to the human trials and introduced the key documents to familiarize as a SAS programmer.
4. Once we understood the job of the Study Protocol, we saw how the Case Report Form is used to collect data on the patients.
5. We observed how the electronically collected data makes it way to a Clinical database and then to a SAS dataset while noting the large presence of Regulatory agencies at every step of the way. We decided to focus on a single patient with a single lab test result and follow that patient's course through the rest of the way.
6. Next, we understood the importance of Standards initiatives and how they apply to the collected data. The Study Data Tabulation Model or SDTM was introduced to you to represent all collected patient level data.
7. Next, we examined the most important document for the programmer to refer to plan his analysis and reports. We noted this to be the Statistical Analysis Plan or SAP and looked at some important sections from it.
8. Once we learned the tasks for the programmer in the SAP, it became apparent that the SDTM data were not in a format for ready analysis and reporting. We needed a separate data standard – Analysis Data Model or ADaM to derive variables and attach baseline markers and other useful variables for creating reports.
9. We looked at what kind of reports are produced and how the ADaM data facilitated these analysis by being one DATA or SQL step and a Statistical PROC away from the requested numbers.
10. Finally, we visualized the output and acknowledged that this entire microcosm of a study was a simplified version to aid the understanding of a beginner. In the real world the data and analysis are much more complex.

Hopefully, you have come away with a basic understanding of what a SAS programmer does in the drug development industry.

REFERENCES

Grabowski, G. Henry and Hansen, W. Ronald. 2014. "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs". Accessed March 6, 2018. http://csdd.tufts.edu/files/uploads/Tufts_CSDD_briefing_on_RD_cost_study_-_Nov_18,_2014..pdf

Center for Drug Evaluation and Research (CDER), FDA. 2003. "Part 11, Electronic Records; Electronic Signatures — Scope and Application". Accessed March 6, 2018. https://www.fda.gov/RegulatoryInformation/Guidances/ucm125067.htm

MedlinePlus, U.S National Library of Medicine, 2018. "Lipase test". Accessed March 6, 2018. https://medlineplus.gov/ency/article/003465.htm

i3 Statprobe. 2010. "STATISTICAL ANALYSIS PLAN". Accessed March 6, 2018. https://www.accessdata.fda.gov/Static/widgets/tobacco/MRTP/09%20appendix-2h-smna-smkng-cstn/sm-07-01/4.%20SAP%20and%20Rslts/SAP.pdf

National Institute on Drug Abuse, NIH. 2017. "Double-Blind, Placebo-Controlled, Dose Escalating, Safety and Pharmacology Study With Three Dosages of GBR 12909 in Cocaine Experienced Volunteers". Accessed July 30, 2017. https://datashare.nida.nih.gov/study/nida-cpu-0002

Clinical Data Interchange Standards Consortium (CDISC), 2012. "Study Data Tabulation Model (SDTM)". Accessed July 30, 2017. https://www.cdisc.org/standards/foundational/sdtm

Clinical Data Interchange Standards Consortium (CDISC), 2009. "Analysis Data Model (ADaM)". Accessed July 30, 2017. https://www.cdisc.org/standards/foundational/adam

Clinical Data Interchange Standards Consortium (CDISC), 2013. "SDTMIG v3.2". Accessed July 30, 2017. https://www.cdisc.org/standards/foundational/sdtmig

Clinical Data Interchange Standards Consortium (CDISC), 2009. “Analysis Data Model (ADaM) Implementation Guide”. Accessed July 30, 2017. https://www.cdisc.org/sites/default/files/members/standard/foundational/adam/adam_implementation_guide_v1.0.pdf

Pharmaceutical Research and Manufacturers of America (PhRMA), 2015. “The Biopharmaceutical Research and Development Process”. Accessed July 30, 2017. https://www.phrma.org/graphic/the-biopharmaceutical-research-and-development-process

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Venky Chakravarthy
BioPharma Data Services
venky@biopharmadataservices.com