ABSTRACT
Clinical SAS® programmers come from diverse background. As programmers step into this new field, they would have enough working knowledge about SAS techniques and how to program tables, listings, and graphs. However, as in any other field, there are lots of everyday activities, terminologies and processes that a programmer should be aware of in order to be successful and will learn on the job over a period of time, depending on the work environment. This paper is primarily targeted towards programmers who are relatively new to the field of clinical programming and the objective is to provide an early introduction to the various aspects of clinical programming.

CLINICAL TRIALS
By now, you must have heard about FDA and its consumer watchdog division, called CDER (Center for Drug Evaluation and Research), whose job is to evaluate new drugs before they are marketed. The process of development and approval of new drugs are generally complicated, expensive and time-consuming and involves many scientists and professionals with varying expertise.

Once a company identifies a compound as promising, a series of pre-clinical trials will be conducted and the results of those studies as well as future plans justifying clinical trials are submitted to the FDA. Upon approval from the agency, the company will start testing on humans.

Clinical Trials are well controlled research studies of a new drug or device to answer scientific questions regarding the safety and efficacy of the drug that is being tested. They progresses in a series of phases.
Phase I trials – This first-in-human studies are conducted on a small group of healthy volunteers or individuals with the target disease, to establish a dosage level and to study the safety profile.
Phase II trials – Intended to evaluate the effectiveness and to continue safety testing on individuals with the target disease, on a larger group of people.
Phase III trials – To further obtain additional safety and efficacy data on a large group of demographically diverse population. Usually, involves multiple trials conducted at many sites either nationwide or worldwide.
Phase IV trials – These are long term studies conducted after the drug has been approved and marketed. Purpose is to study the side effects as a result of prolonged use.

As you can imagine, there will be enormous amount of data that will be collected from all these clinical trials at various stages of development and stored in various databases. Clinical scientists and other team members need to obtain results and draw inferences from these studies - and that is where you come into the picture. Programmers extract the data from the database and present them in a condensed table form or print out all the details as a listing, or present them as a graph, as deemed necessary. This is put in a very simple way – in real life, it is a lot more complex and involves extensive amount of work.

The data that is collected during clinical trials are broadly grouped into either safety or efficacy. Safety data which includes adverse events, labs, vitals, etc. are collected to monitor the patient safety while participating in the clinical trials, and has always been a primary focus in the development of new pharmaceutical products as it helps to relate the benefits to risk. Data collected to demonstrate the effectiveness of the investigational product are grouped under efficacy and statisticians will be primarily using this data to draw statistical inferences.
Once a programmer is assigned to work on a study, there are certain documents that need to be reviewed to gain good understanding of the study design before starting to dig through the data. Among them, Study Protocol, Case Report Form (CRF) and Statistical Analysis Plan (SAP) or its equivalent are the most important.

Study Protocol describes among other things, the way the study is designed, whether it is open-label or blinded study, treatments assigned, number of subjects, duration, visit schedule, etc. Since the study protocol is a reference document for the clinical trial, it will include a lot more detailed information than what a programmer is looking for. It is worthwhile to spend time in understanding the study protocol and it will help tremendously while coding.

A CRF is a printed or electronic document designed to collect all protocol required information on each subject in a clinical trial. This data will then be transferred into a database and will be made available to SAS programmers as ‘raw’ datasets. Annotating the CRFs with variable names will provide a link between the variables in the dataset and the fields on the form. In most companies, CRFs are annotated electronically since it is mandatory for electronic submission of datasets to the agency.

SAP details the analyses planned for the study. Study protocol will have brief sections describing the analyses plans, but will be documented in detail in the SAP. It describes the various population for analyses, when and how to start counting adverse events, baseline and treatment windows, what data to include for interim and final analyses, how to handle outlier data points, etc. Also, it will include a list of displays (listings, tables and graphs) that need to be generated for the study along with the formatting details. Since this document provides all the necessary information in greater details, it becomes imperative for a programmer to spend significant amount of time in reading and understanding this document before starting to write codes. In addition to SAP, statisticians normally would provide the specifications for deriving the efficacy datasets, which would be subsequently used for analyses.

Other important documents include (1) Clinical Data Management Plan – describing how the data is collected, entered, stored, checked etc. (2) Data Transfer Specifications for data that comes from external vendor – ex. some lab data.

Once a clinical programmer starts working on projects, he or she would come across data relating to adverse events and concomitant medications. Invariably at that point you will start hearing about dictionaries used to code the terms, their version etc. So, what are dictionaries and how are they related to clinical data? Dictionaries are used to map verbatim or clinicians terms for an adverse event or drug name to a standard term. In the industry, the most commonly used dictionaries while processing clinical trials data are MedDRA and WHODrug.

An adverse event is any unwanted or undesirable health effect that occurs in a participant during a clinical trial. When data are collected during trial, patients report the symptoms in ‘free text’ form. For example, patients might report ‘slight headache’ or ‘stress headache’ or ‘sinus headache’ etc. In order to aggregate these terms into medically meaningful groupings for tabulation and analyses purposes, we need to use a standard lexicon.

MedDRA (Medical Dictionary for Regulatory Activities) is a standardized international medical terminology used by biopharmaceutical industry and it is the de facto standard for analysis and reporting of adverse events. In MedDRA, events are organized by System Organ Class (SOC – ex. Cardiac Disorders), divided into High-Level Group Terms (HLGT – ex. Cardiac Arrhythmias), High-Level Terms (HLT – ex. Rate and Rhythm Disorders NEC), Preferred Terms (PT – ex. Arrhythmia) and finally into Lower-Level Terms (LLT – ex. Arrhythmia, Dysrhythmias). Each MedDRA term is assigned an 8 digit numeric code. Once the verbatim adverse events texts that
are collected in the CRF are mapped to the dictionary, each event will be assigned a MedDRA event and code, at 5 different hierarchical levels. In some companies, the mapping gets done at the database level, while in some others the mapping will be done in SAS environment. Once the events are mapped, programmers would summarize the events based on the counting rules established by the company. A typical AE summary table would summarize only the PT and SOC. MedDRA versions are updated and released twice a year – in March and September. Data on medications other than the investigational drugs taken by subjects prior and during study period are normally recorded. It is very common that the same medication would be referred to by different names, and there is a need to classify all similar medications into one while tabulating medication usage and that is where drug dictionary proves useful. The World Health Organization – Drug (WHO-DRUG) dictionary is the most commonly used one in the industry. This dictionary uses the Anatomical Therapeutic Chemical (ATC) classification system to group the thousands of drugs into meaningful categories. The ATC system is hierarchical and includes five levels of granularity. The first level of the code indicates the anatomical main group and consists of one letter – Example: A – Alimentary Tract and Metabolism. The second level of the code indicates the therapeutic main group and consists of two digits- Example: C03 Diuretics. The third level of the code indicates the therapeutic/pharmacological subgroup and consists of one letter - Example: C03C High-ceiling diuretics. The fourth level of the code indicates the chemical/therapeutic/pharmacological subgroup and consists of one letter - Example: C03CA Sulfonamides. The fifth level of the code indicates the chemical substance and consists of two digits- Example: C03CA01 Furosemide.

ROLE OF CLINICAL PROGRAMMERS
Clinical programmers generate output/reports to support numerous activities that happen during drug development process such as CSRs (Clinical Study Report), DSMBs (Data Safety Monitoring Board), Annual IND (Investigational New Drug) Reports, IBs (Investigators' Brochure), FDA requests, safety updates to the agency, etc. As a beginner, you might get confused with all these acronyms and might not quite understand what they refer to. The following section briefly explains the most common terms or acronyms that are used in the industry.

One of the most critical of the documents submitted to regulatory bodies in the drug development process for marketing authorization application is the Clinical Study Report (CSR). It is an extensive and complete documentation of the integrated full report of efficacy and safety data for an individual study of an investigational product. By reading the CSR, one can understand why and how the study was conducted, the types of data collected and analyzed, and the nature and extent of the conclusions that may be drawn from the results. This document includes numerous safety and efficacy tables, listings and figures - clinical programmers play a pivotal role in generating these tabulations in discussion with statisticians, clinicians and medical writers.

Data Safety Monitoring Board (DSMB) or Data Monitoring Committee (DMC) is an independent committee composed of clinical research experts, statisticians, etc. Their role is to regularly review clinical trials data while the study is in progress to ensure the safety of participants. A DSMB may recommend that a trial, or part of a trial, be stopped or modified if there are safety concerns. The DSMB looks at the results of un-blinded analyses that are not available to the investigators or to anyone else. Protocol programmers who are assigned to the blinded study will be responsible to generate all the reports as agreed upon by the study team in a blinded fashion. In addition, there will be an un-blinded programmer and an un-blinded statistician who will be responsible for breaking the blind and provide the un-blinded reports to the board members. The DSMB reports will normally display a non-specific treatment codes such as A, B, C to prevent any inadvertent un-blinding of study team members. Treatment decodes will then be provided to the DSMB members in a secure manner.

Once a sponsor files for an IND, it is required to submit an annual report to the agency at yearly intervals within 60 days of the anniversary date when the IND became effective. Some of the required information that are generated by the clinical programmers and published in the IND Annual Report includes, subject demography, summary tabulations of AEs and SAEs within a
given reporting period, number of subjects discontinued due to AEs, death, etc. Kristi Wiser in her paper titled ‘IND Annual Reporting at a Glance’ provides a concise overview of the IND Annual Reporting process.

IB is a compilation of clinical and non-clinical data on the investigational product and the purpose is to provide the investigators and others involved in the trial a clear understanding of all aspects of the study. Programmers will generate safety reports periodically to be updated in the IB since the sponsors are responsible for keeping the information in the IB up to date.

When an NDA comes in, the FDA has 60 days to decide whether to file it so that it can be reviewed. The FDA expects to review and act on at least 90 percent of NDAs no later than 10 months after the applications were received for standard drugs and six months for priority drugs. During this review period FDA could request additional data, analyses, or reports from the sponsor. Also, safety update reports are submitted to the agency at 4 months (120 days) after the initial application, and at any time when the agency requests for an update. Again, programmers play a major role in generating these tabulations for submission to the agency.

DATA STANDARD
There are numerous clinical trials conducted globally by the pharmaceutical industry and the data is stored in various proprietary database structures with no industry wide standards. This lack of globally accepted data standards cost the pharmaceutical industry millions of dollars per annum. Clinical Data Interchange Standards Consortium (CDISC) is an organization or data standards group focused on developing platform-independent data standards for the interchange of clinical information within the pharmaceutical market. Specifically, CDISC is developing standard data models to support the electronic acquisition, exchange, submission, and archiving clinical trial data and metadata (data about data). Implementation of CDISC standards would result in significant reductions in time and cost.

There are different data models with CDISC such as SDTM, ADaM, Lab Model, ODM, etc., and you will gain better understanding of those as you start working on those. As a beginner, you might want to initially focus on learning more about the two data models that are of great importance viz. Study Data Tabulation Model (SDTM), and Analysis Data Model (ADaM).

Use of the SDTM in particular has been endorsed by FDA since 2004, and use of the other standards has been rapidly growing in acceptance in recent years. SDTM data is study data that has been formatted according to very specific standards. The regulatory agency has announced their intention to make SDTM required by regulation in the near future. Many companies have already started submitting data to the FDA in SDTM format. An organization might build an entire system starting from data collection in an SDTM standard. However, in many cases data were already collected using the existing system and this data is then mapped to SDTM standards. As a result of these changes in the clinical trials submission process, it is imperative for a programmer to be well versed with SDTM concept.

ADaM provides standards for creating analysis datasets used for statistical review and analysis. The primary purpose is provide the regulatory authorities with a clear description of the content, structure, and usage of datasets and variables used by the sponsor in support of statistical analyses. These analysis datasets in the model were designed in such a way that they are ‘one proc away’ from an analysis table or graph and could be used directly in a statistical procedure to get the results.

There are numerous papers written on CDISC and the various data models which would help in developing a deeper understanding of the concept – a compilation of which can be seen at www.lexjansen.com/cdisc.

CONCLUSION
Gaining technical knowledge in SAS coding is a good beginning, but that is only one half of what it takes to become a successful clinical programmer. The other half includes gaining knowledge
about the industry guidelines, standards and trends. There is a wealth of information available online in the form of technical papers published in various conference proceedings - which could be a good starting point for a programmer who is relatively new to the industry.

REFERENCES

CONTACT INFORMATION
Ramesh Ayyappath
Merck & Co., Inc
2015 Galloping Hill Road
Kenilworth
NJ 07033
E-mail: ramesh.ayyappath@spcorp.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other Countries. ® indicates USA registration.