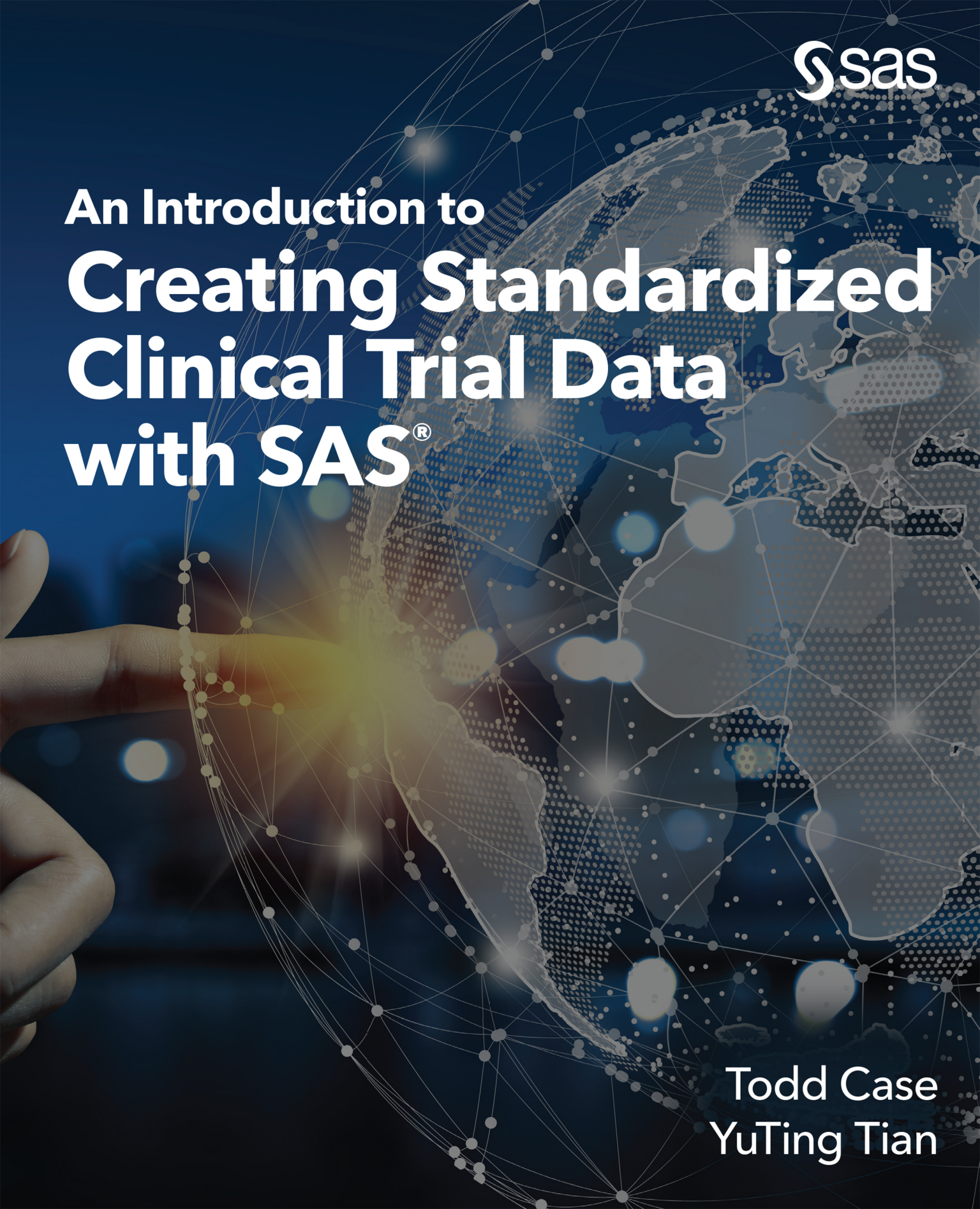

sas
An Introduction to
Creating Standardized Clinical Trial Data with SAS®
Todd Case
YuTing Tian

The correct bibliographic citation for this manual is as follows: Case, Todd and YuTing Tian. 2022. *An Introduction to Creating Standardized Clinical Trial Data with SAS®.* Cary, NC: SAS Institute Inc.

An Introduction to Creating Standardized Clinical Trial Data with SAS®

ISBN 978-1-955977-90-6 (Hardcover)
ISBN 978-1-955977-95-1 (Paperback)
ISBN 978-1-955977-96-8 (Web PDF)
ISBN 978-1-955977-97-5 (EPUB)
ISBN 978-1-68580-026-0 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

August 2022

Contents

About This Book ..... **vii**
What Does This Book Cover? vii
What Are the Prerequisites for This Book? vii
What Should You Know about the Examples? vii
SAS OnDemand for Academics viii
Acknowledgments viii
We Want to Hear from You viii

Chapter 1: Understanding the Industry **1**
1.1 Statistical Programmer Work Process 1
1.2 Drug Approval Process 2
1.3 Clinical Trial Study Design 3
1.4 CDISC Standard Data Structures 4
1.5 Important Documents Summary 4

Chapter 2: Getting Started from the Case Report Form **7**
2.1 eCRF Portal 7
2.2 Electronic CRFs (eCRFs) 8
2.2.1 Demographics 8
2.2.2 Disposition 9
2.2.3 Adverse Events 10
2.2.4 Exposure 12
2.2.5 Concomitant Medications 14
2.2.6 Electrocardiogram 16
2.2.7 Lab 18
2.3 Annotating the eCRF 20
2.3.1 Annotating Unique eCRF Pages 21
2.3.2 Appearance of Annotations 21
2.4 Annotated CRF Practices 22

Chapter 3: Study Data Tabulation Model (SDTM) **29**
3.1 Variable "Roles" 29
3.2 SDTM Standard Domains 29
3.3 SDTM Core Variables 30
3.4 Clinical Trial Schedule of Assessments 31
3.5 Creating a New Domain 31

3.6 Model for SDTM Generation 32
3.6.1 Demographics (DM) 32
3.6.2 Disposition (DS) 46
3.6.3 Adverse Events (AE) 53
3.6.4 Exposure (EX) 66
3.6.5 Concomitant Medications (CM) 73
3.6.6 Electrocardiogram Test Results (EG) 80
3.6.7 Laboratory Test Results (LB) 94
3.7 Trial Design Domains 107
3.7.1 Trial Summary Data Set (TS) 108
3.7.2 Trial Arm Data Set (TA) 110
3.7.3 Trial Element Data Set (TE) 112
3.7.4 Trial Visit Data Set (TV) 114
3.7.5 Trial Inclusion/Exclusion Data Set (TI) 115

Chapter 4: Analysis Data Model (ADaM) 119
4.1 ADaM Standard Structures 119
4.1.1 Variable "Roles" 119
4.1.2 Standard ADaM Structure and Domains 120
4.1.3 Standard ADaM Variables 120
4.2 Subject-Level Analysis Data Set (ADSL) 120
4.2.1 Structure of ADSL 121
4.2.2 ADSL Specification 122
4.2.3 ADSL Programming 125
4.3 Basic Data Structure (BDS) 130
4.3.1 ECG Test Results Analysis Data Sets (ADEG) 130
4.3.2 Lab Test Results Analysis Data Sets (ADLB) 143
4.4 Occurrence Data Structure (OCCDS) 155
4.4.1 Adverse Event Analysis Data Set (ADAE) 155
4.4.2 Concomitant Medication Analysis Data Set (ADCM) 163

Chapter 5: Case Report Tabulation Data Definition (Define-XML) 171
5.1 Structure of Define-XML 171
5.2 The Process of Creating Define-XML 173
5.2.1 Create Metadata Spreadsheet and Create Define-XML Components 173
5.2.2 Create XPT Files 176
5.2.3 Link for External Documents 177
5.2.4 Construct Define.XML 177
Conclusion 177

Appendix 179
A.1 Raw Data Spreadsheet 179
RAW.TS 179
RAW.TA 180
RAW.TE 180

RAW.TV 181
RAW.TI 181
RAW.DM 181
RAW.DS 182
RAW.AE 182
RAW.EX 184
RAW.CM 185
RAM.EG 185
RAM.LB 187
A.2 SDTM Programming Section 188
SDTM.DM 188
SDTM.SUPPDM 191
SDTM.DS 193
SDTM.AE 194
SDTM.EX 197
SDTM.CM 199
SDTM.EG 201
SDTM.LB 205
SDTM.TS 209
SDTM.TA 210
SDTM.TE 211
SDTM.TV 212
SDTM.TI 213
A.3 ADaM Programming Section 214
ADaM.ADSL 214
ADaM.ADEG 216
ADaM.ADLB 221
ADaM.ADAE 225
ADaM.ADCM 227

About This Book

What Does This Book Cover?

The purpose of this book is to introduce standardized clinical trial data to anyone interested in understanding the pharmaceutical industry and how that data is collected and created.

This book introduces the concept of standardized clinical data, technical terms, and programming practices in the pharmaceutical industry as well as clear and concise explanations with numerous practical examples. We include basic knowledge of the pharmaceutical industry as well as SAS programming practices used in the industry.

This book does not cover how to create define.xml, although we do introduce it to the reader.

What Are the Prerequisites for This Book?

The only prerequisite for this book is an interest in the pharmaceutical industry.

What Should You Know about the Examples?

This book includes SAS code and simulated data for the reader to gain hands-on experience with standardized clinical data. Visit the author's page at http://support.sas.com/case to access the example code and data.

Software Used to Develop the Book's Content

SAS Version 9.4 was used to develop the content and examples in this book.

Example Code and Data

This book includes data and complete programs used to create simulated standardized clinical trial data. Visit http://support.sas.com/case to access the example code and data.

An example to derive sex in the Demographics domain is demonstrated below:

```
/*Derive SEX*/
if SEX_="Female" then SEX="F";
else if SEX_="Male" then SEX="M";
else if SEX_="Unknown" then SEX="U";
else if SEX_="Undiff erentiated" then SEX="UNDIFFERENTIATED";
```

SAS OnDemand for Academics

 This book is compatible with SAS OnDemand for Academics. If you are using SAS OnDemand for Academics, then begin here: https://www.sas.com/en_us/software/on-demand-for-academics.html.

Acknowledgments

Thank you to CDISC and the technical reviewers who provided feedback: Margaret Hung, Matt Becker, Peter Eberhardt, Laura Elliott, William Kuan, and Crystal Cheng.

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Chapter 1: Understanding the Industry

In the pharmaceutical industry, there is a mandate to create standardized clinical data using very specific rules. These rules are created and governed by the Clinical Data Interchange Standards Consortium (CDISC). In this book, we describe and illustrate how to create these required CDISC data sets with SAS code. A statistical programmer should be familiar with the CDISC rules required to create standardized clinical trial data sets. After reading this book, readers will be able to understand CDISC standardized clinical data structures, as well as how to create it.

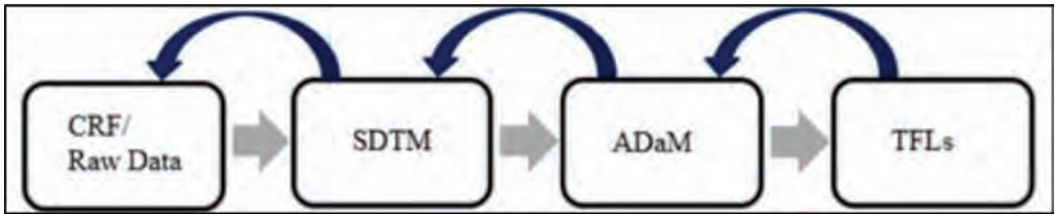
1.1 Statistical Programmer Work Process

In the pharmaceutical industry, the primary goal of a statistical programmer is to create standard data efficiently in order for the clinical trial biostatistician to perform their analysis. A simplified illustration of the process workflow for the statistical programmer is shown in Figure 1.1.

Figure 1.1 shows that the work process starts from the Case Report Form (CRF), which is designed for a specific study to collect clinical trial raw data from a site. Often, studies are global – having sites in countries all over the world. The Data Management group creates the CRF by working with the statistical programmer and other functions to ensure that the appropriate data is collected for the purpose of that study.

After the CRF is created and data is entered into it by the sites, the statistical programmer uses this data to create CDISC Study Data Tabulation Model (SDTM) domains to group collected information from the CRF in a way that facilitates standardization. The statistical programmer then creates CDISC Analysis Data Model (ADaM) data sets from the SDTM domains to support clinical trial analysis.

Figure 1.1: Statistical Programmer Process Workflow



Note: When we refer to SDTM, we use the term domain, and for ADaM, we use the term data set. To be crystal clear, both models generate standardized clinical data using SAS.

Creating SDTM and ADaM data sets ensure that data will meet the criteria to be accepted by regulatory agencies such as the United States Food and Drug Administration (FDA). Finally, the statistical programmer generates the Tables, Figures, and Listings (TFLs), which are used to support analysis presented in the Clinical Study Report (CSR). The CSR is used to provide evidence to regulatory agencies about the safety and efficacy of the study drug.

Note: This workflow actually represents a much more complicated process. We intentionally keep it at a level where the reader can just focus on how the statistical programmer receives the raw data and creates the standardized clinical data (SDTM and ADaM).

1.2 Drug Approval Process

The FDA's Center for Drug Evaluation and Research (CDER or CBER for Biologics) reviews the SDTM and ADaM data created in the previous section to assess the drug's safety and efficacy. There are many stages of development and clinical trials as the drug approval process advances. The following are the most critical drug development terms and milestones:

- **Pre-clinical Studies**: Research often using animals to find out if a drug is likely to be safe in humans.
- **Investigational New Drug (IND) Application**: Facilitates permission to start clinical trials in humans if the pre-clinical study results are promising.
- **Phase I Clinical Trial**: First in human (FIH) study of a new drug, often looking at dose ranges, drug-drug interactions, food effect, etc.
- **Phase II Clinical Trial**: Explore and determine efficacy of a new drug.
- **End of Phase II Meeting**: Regulatory agency (for example, FDA) and sponsor agree on design of Phase III study.
- **Phase III Clinical Trial**: Large-scale clinical trial that confirms efficacy and safety that, if successful, will be reviewed by regulatory agencies for marketing approval.
- **Pre-NDA/BLA Review Meeting**: Discuss strategy for potential approval of the IND, format and content of the anticipated application, including labeling, risk evaluation and mitigation strategy, data structure and accessibility of data for submission.
- **New Drug/Biologic Application (NDA/BLA)**: New drug application that can lead to market approval, which allows the drug to be legally marketed.
- **Drug Labeling Review**: Identify drug contents, information, and specific warnings for administration, storage, and disposal.
- **Facility/Sponsor Inspection**: Regulatory agency visits the sponsor, sites, or manufacturing facilities to evaluate trial conduct and compliance with the protocol and other regulatory requirements. This is often performed after submission of an NDA/BLA.

- **Phase IV Clinical Trial**: Experiments to conduct the long-term safety of a new drug after the drug is approved and is on the market. These are often designed to meet approval or reimbursement in areas outside of the United States, Japan, and China. (All of these countries have their own regulatory agencies that require standardized data be submitted.)

Table 1.1: Summary Table for the Four Clinical Trial Phases

	Phase 1	Phase 2	Phase 3	Phase 4 (pro-market)
Participants	Healthy volunteers or patients	Patients	Patients	Patients
Number	20–100	Up to hundreds	300–3000	Large, diverse population
Length	From days up to several months	Several months–2 years	1–4 years	Several years
Goal	Safety and dosage	Efficacy and side effects	Efficacy and monitoring of adverse reactions (safety)	Long-term safety and efficacy
% Continuation	Around 70% of the drugs move to the next phase	Around 33% of drugs move to the next phase	Around 25–30% of drugs move to the next phase	

Source: FDA. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research

1.3 Clinical Trial Study Design

The clinical trial design is one of the most critical interventional trial components. It serves to optimize the clinical trial conduct and provide the most objective range of approaches to evaluate the therapy. There are several clinical trial design concepts in practice that the reader needs to know. We only list the common clinical trial designs; in practice each one of these can be used in combination with each other. In addition, there are many other nuances to each design. For example, a Phase III randomized trial is often placebo-controlled and double-blind and has an open label extension for safety purposes.

- **Randomized:** Participants are divided randomly into separate treatment (placebo) groups that compare the groups.
- **Placebo-controlled:** Placebo is given to one group of participants, while a therapy is given to another group. Placebo is designed to have no real effect.
- **Open-label:** Both the researchers and the participants know which treatment is being administrated.

- **Double-blind:** Neither the participants nor the researchers know which treatment is assigned and administered.
- **Parallel Design:** Patients are randomly assigned to a treatment and remain on that treatment throughout the duration of the entire trial.
- **Crossover Design**: All subjects switch treatment regimens during the course of the trial.

For more information, please check the Drug Study Designs Guidance for Institutional Review Boards and Clinical Investigators: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/drug-study-designs.

1.4 CDISC Standard Data Structures

CDISC is a global non-profit organization that develops data standards for the pharmaceutical industry. There are three distinct standard data models developed by CDISC for regulatory submissions that the reader needs to understand. The basic concepts for these three models are below. More details are provided in subsequent chapters.

- **Study Data Tabulation Model (SDTM):** Defines a standard structure for human clinical trial data tabulations that are sent to a regulatory authority such as the FDA as a part of the data submission package. The SDTM model is considered the 'raw' data for regulatory submission.
- **Analysis Data Model (ADaM):** Uses the SDTM domains to develop data sets for the purpose of summarizing and analyzing the clinical data. The ADaM model data sets generate all the analysis to support the trial.
- **Define-XML**: When sending SDTM and ADaM data sets to the regulatory authorities, it's critical to see the specifications and understand how to navigate the SDTM and ADaM data sets. DEFINE-XML provides a machine-readable version of how the SDTM and ADaM data sets were created, including any explanations about complex data derivations. This allows the FDA to work more efficiently with data submission.

1.5 Important Documents Summary

There are some key and important documents that are essential for statistical programmers to understand, use, or create in order to create standardized clinical data – SDTM and ADaM. We include TFLs as they are why we create SDTM and ADaM data sets. The documents are listed in the order they are created. There will be multiple iterations, and for the reader's sake, we don't feel it's necessary to talk about every single scenario. The only document that MUST be finalized before all other documents is the Protocol.

- **Protocol:** Detailed summary and guide of the study including study design, schedule of assessments, and analysis methods. Every subsequent document and the study conduct are based on the Protocol. It is reviewed and approved by Institutional Review Boards (IRBs), regulatory authorities, and sites.
- **Blank Case Report Form (CRF):** Used to collect all the information for every single patient in the study. The CRF is created by the data manager, then the statistical programmer, biostatisticians, and other functions review the CRF to confirm that all the data needed for analysis is captured. The CRF can only be finalized after the Protocol is finalized.
- **Statistical Analysis Plan (SAP):** Created by the study biostatistician and explains how the data is to be analyzed.
- **Table, Figure, and Listing templates (TFLs):** Created by the study biostatistician, these provide the content and detailed information to help statistical programmers create the actual TFLs once the SAP is stable.
- **SDTM Annotated Case Report Form (SDTM aCRF):** Annotated by the statistical programmer. Statistical programmers use the annotated CRF to generate and understand the structure of SDTM domains.
- **SDTM Specifications:** Provide details about how to generate the SDTM domains; they cover information about how to program all domains, including variables' lengths, labels, formats, as well as instructions on how to create each variable. They are created by the statistical programmer with the SDTM aCRF simultaneously as the two documents are highly correlated and dependent on each other.
- **ADaM Specifications:** Contain information about the analysis data sets from SDTM domains as well as new variables and derivations required for analysis purposes in ADaM data sets. These specifications are created by the statistical programmer. A stable SAP and TFL shells are required in order to generate ADaM specifications.
- **Define-XML:** Machine-readable version of specifications including the SDTM Define-XML document and ADaM Define-XML document. This also provides more detailed information about how the data was created.

Table 1.2: Summary Table of Important Documents

Document	Purpose	Statistical Programmers' Role	Time
Protocol	Detailed summary and guidance of the study	Study lead review	Before study starts
Blank CRF	Detailed data collection	Study lead review	Before study starts
SAP	Explain how the data is analyzed	Study lead review	Pre-Programming
TFL Templates	As a reference when creating tables, listings, and figures	Study lead review	Pre-Programming
SDTM aCRF	Link CRF with SDTM domains	Study lead creates	Pre-Programming

(*Continued*)

Table 1.2: (*Continued*)

Document	Purpose	Statistical Programmers' Role	Time
SDTM specifications	Explain the derivation of each variable in SDTM domains	Study lead creates	Pre-Programming
ADaM specifications	Explain the derivation of each variable in ADaM data sets	Study lead creates	Pre-Programming
Define-XML	Provides machine-readable version of specifications	Study lead creates	Study End