Navigating Clinical Trial Data Analysis with R

An ***Introduction*** to CDISC, R Programming by Example, Dealing with Missing Values, and Reporting with Quarto

Joshua J. Cook

M.S. DS., M.S. CRM., ACRP-PM, CCRC

NORTH CAROLINA PharmaSUG SDE
OCTOBER 17, 2024
- VIRTUAL SEMINARS -

My Background

- 2016 – FLVS
- 2021 – B.S., Biomedical Sciences
- 2023 – M.S., Clinical Research Management, ACRP Certifications
- 2024 – M.S., Data Science
- 2025 – M.D./Ph.D. Matriculation 🤞
  - M.D. – Oncology
  - Ph.D. – Cell Biology

Why did I make this workshop?

- **To develop a new training pathway for people like me**
  - I wasn't taught CDISC / clinical trial programming in school
  - Instead, I attended numerous professional conferences and training seminars on open-source programming and clinical trial programming
  - Multiple programming languages
  - Hard if not coming from a strong SAS background
- I wanted this to be a **truly introductory experience.**

Overview of this Workshop

With Example Code!

- Introduction to Clinical Trials & Data Analysis
- Module 1: Importance of Standardization and Compliance
- Module 2: Analyzing Clinical Trial Data
- Module 3: Handling Missing Data
- Module 4: Reporting and Reproducible Research
- Final Q&A and Closing Remarks

Introduction to Clinical Trials & Data Analysis

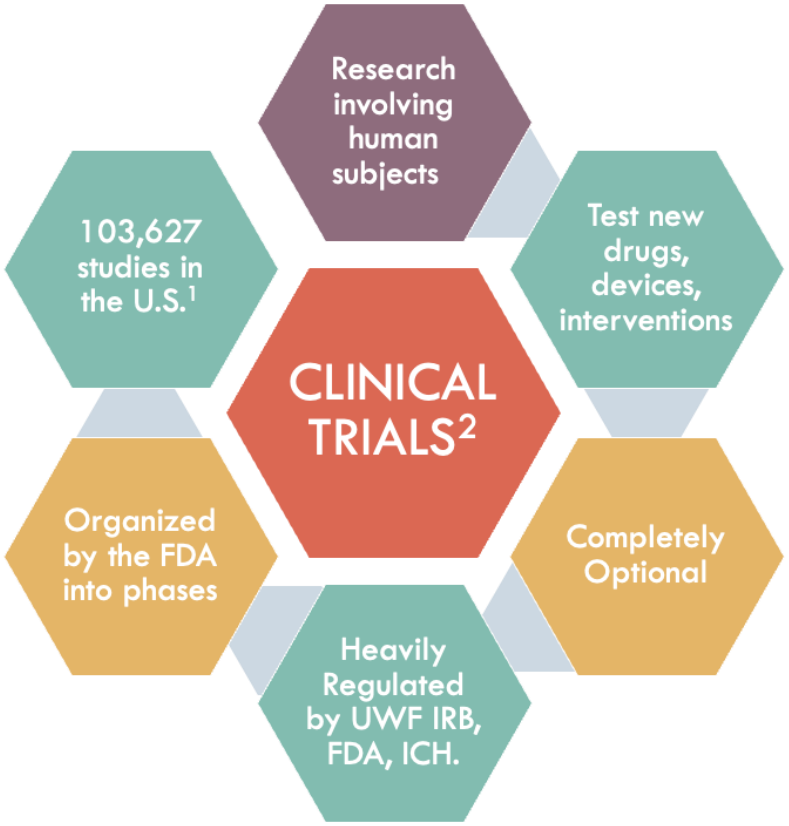
Clinical Trials

Clinical Trials are research studies that involve **humans.**

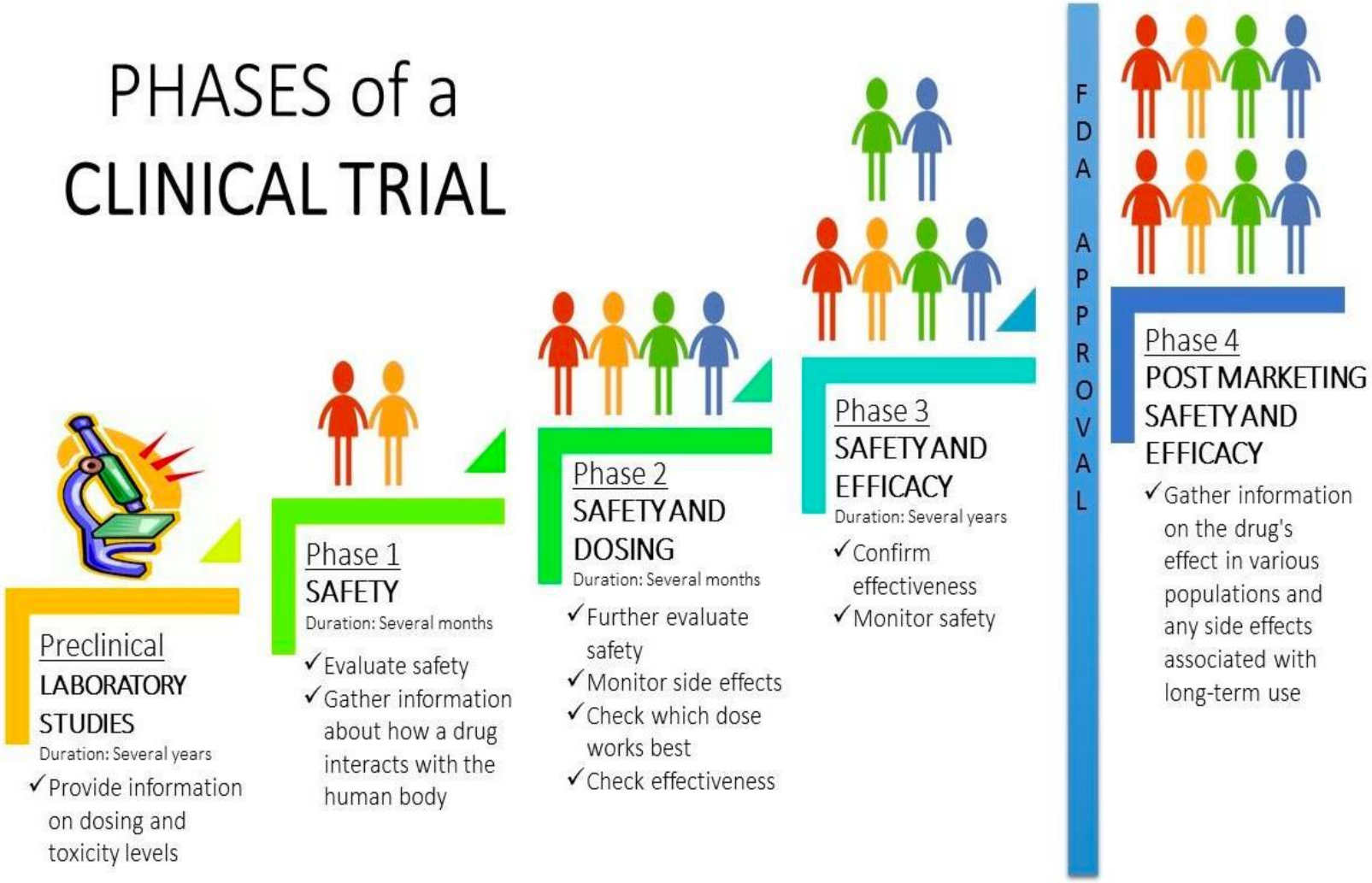
- Interventional v. observational
- Phases I-IV
- Drug/device approval happens after Phase III
- Typically funded by pharmaceutical sponsors or research institutions (via grants)
- **Offered free-of-charge to study participants.**
- **Completely voluntary**

Defined in 45 C.F.R. §46.102

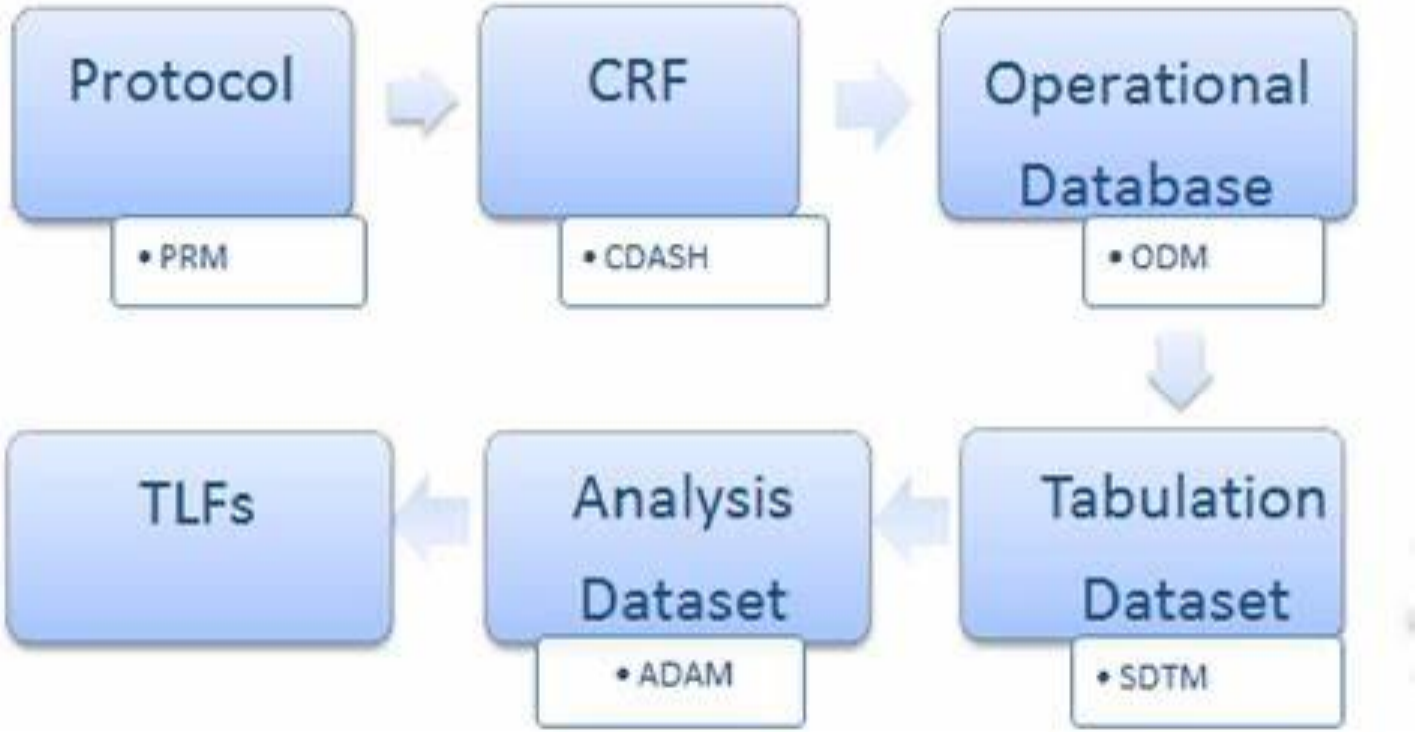
https://www.ecfr.gov/on/2018-07-19/title-45/subtitle-A/subchapter-A/part-46



Clinical Trials



Introduction to Clinical Trial Data Analysis
Turning Raw Data Into Actionable Insights



https://medium.com/@sairabhanusk1/cdisc-1d4e3d17aa9a

Introduction to Clinical Trial Data Analysis

Turning Raw Data Into Actionable Insights

- Clinical trials generate large, complex datasets, requiring systematic analysis to draw meaningful conclusions. In a nutshell, clinical trial data analysis involves several key steps:
  - **Data Collection –** in the clinic, typically through Case Report Forms (CRFs) within the Electronic Data Capture Systems (EDCs) and Electronic Medical Records (EMRs)
  - **Data Cleaning –** {janitor}, {dplyr}, {tidyr}
  - **Data Transformation –** {sdtm}, {admiral}, {data.table}, {mice}
  - **Statistical Analysis –** {gtsummary}, {survival}, {lme4}
  - **Reporting and Submission –** {quarto}, {officer{, {xportr}, {rtf}

Key Components of Clinical Trial Data Analysis

- **Protocol Development**: Defines what data needs to be collected and how it should be analyzed
- **Randomization and Blinding**: Ensures unbiased results by randomly assigning participants to different treatment groups and preventing knowledge of group assignments.
- **Handling Missing Data**: Missing data can skew the results if not handled appropriately. Techniques like multiple imputation help fill in gaps without introducing bias.
- **Statistical Power**: Ensuring the trial is adequately powered to detect treatment effects, meaning you have enough participants to make the results meaningful.

Chow, S., Liu, J., Chow, S., & Chow, S. (2013). *Design and analysis of clinical trials* :Concepts and methodologies.

Module 1: Importance of Standardization and Compliance

What is CDISC?

- CDISC, or the **Clinical Data Interchange Standards Consortium**, is a global organization that sets the standards for clinical trial data to improve data quality and streamline regulatory submissions.
- Essentially a **common language** that allows different organizations (like sponsors, contract research organizations, and regulatory agencies) to understand and share clinical data more effectively.

CDISC. (2021). *CDISC Standards: Overview*. Clinical Data Interchange Standards Consortium. Retrieved from https://www.cdisc.org/standards

Why CDISC Matters

- **Regulatory Requirement**: CDISC standards are required for FDA submissions in the U.S. and increasingly by other regulatory bodies worldwide (e.g., EMA - EU, PMDA - Japan).
  - **Efficiency**: By using standard data models like ADaM and SDTM, you save time and reduce errors during data analysis and submission - a big problem prior to CDISC.
  - **Consistency**: Standardized data improves data quality and facilitates comparisons between trials.
  - **Reusability**: Data structured in a standardized way can be reused for secondary analyses, such as meta-analyses or real-world evidence studies.
  - **Transparency**: Clear documentation of data derivation improves transparency for regulators and auditors.

Why Standardization and Compliance Are Critical

- **Regulatory Submission**: For a drug to be approved by the FDA, the data must follow CDISC standards. If not, it can delay or even invalidate a submission.
- **Data Sharing and Collaboration**: Standardized data can be shared across teams, organizations, and countries without confusion. This enhances collaboration in large, global clinical trials.
- **Automation**: Tools like R and Pharmaverse automate many steps in the process, such as transforming data into CDISC-compliant formats. This reduces errors and accelerates the workflow.

Key CDISC Standards

- **CDASH (Clinical Data Acquisition Standards Harmonization):**
  - **Purpose**: Defines the basic standards for **collecting clinical trial data** at the data acquisition level. CDASH ensures that data is collected in a way that can be easily mapped to SDTM and ADaM datasets for regulatory submission.
  - **Key Concepts**: Standardized Case Report Forms (CRFs) across clinical trials. Harmonized data fields for efficient and consistent data collection. CDASH variables align with SDTM domains, simplifying the transition from raw data to SDTM-compliant datasets.
  - **Example**: In a clinical trial, the CDASH standard dictates how demographics (e.g., participant age, sex), vital signs (e.g., blood pressure), and adverse events are recorded on CRFs. These are then easily mapped to the respective SDTM domains like **DM** (Demographics) and **AE** (Adverse Events).

CDISC. (2023). *CDASH: Clinical Data Acquisition Standards Harmonization*. Clinical Data Interchange Standards Consortium. Retrieved from https://www.cdisc.org/standards/foundational/cdash

Key CDISC Standards

CDASH Principles

Principle: Metadata will be organized into logical groupings of related concepts

Principle: Traceability for related variables will be reflected in the variable metadata, especially in the variable names

Principle: Data collection standards must be fit-for-purpose for all stakeholders

Principle: For concepts that are the same in both data collection and tabulation, the same controlled terminology shall be used

Principle: Question Text and Prompt must accurately match the CDASH definition of the variable

CDASH Example

cdisc

Site Number

Subject Number

Form DM - Demographics

1 DM - Demographics

1.1	Birth Date (DD-MMM-YYYY)		***BRTHDAT***
1.2	Age		***AGE***
1.3	Age Unit	Years	***AGEU***
1.4	Sex	[F] Female [M] Male [U] Unknown [UNDIFFERENTIATED] Undifferentiated	***SEX***
1.5	Ethnicity	[HISPANIC OR LATINO] Hispanic or Latino [NOT HISPANIC OR LATINO] Not Hispanic or Latino [NOT REPORTED] Not Reported [UNKNOWN] Unknown	***ETHNIC***
1.6	Race	[AMERICAN INDIAN OR ALASKA NATIVE] American Indian or Alaska Native [ASIAN] Asian [BLACK OR AFRICAN AMERICAN] Black or African American [NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER] Native Hawaiian or Other Pacific Islander [WHITE] White [NOT REPORTED] Not Reported [UNKNOWN] Unknown [OTHER] Other Specify Other Race	***RACE*** ***RACEOTH***

Key CDISC Standards

CDASH Example:

Section											
OID	Name	Repeating	Description	Order No.	Mandatory	Aliases	Condition	IsReferenceData	Repeating Information	SASDatasetName	Domain
CDASH_2-1_IG_2	DM - Demographics	No	DM - Demographics [en]	1	Yes						DM
Questions											
OID	Name	Text	DataType	Order No.	Mandatory	Terminology	Length	Significant Digits	Units	Description	Aliases
IT.BRTHDAT	BRTHDAT	What is the subject's date of birth? Birth Date [en]	date	1	No					BRTHDAT [en]	BRTHDAT [CDASH] Record the date of birth to the level of precision known (e.g., day/month/year, year, month/year, etc.) in this format (DD-MON-YYYY). [completionInstructions] BRTHDAT is the collected field used for recording the full birth date. The sponsor may choose to database the date of birth as a single variable (BRTHDAT), or as separate variables for each component of the date/time (BRTHYY, BRTHMO, BRTHDD, BRTHTIM). The

Key CDISC Standards

- **SDTM (Study Data Tabulation Model)**:
  - **Purpose**: Organizes raw clinical data (from CDASH CRFs/e-CRFs) into **standard tables**. These tables are used for organized and standardized submission to regulatory bodies AND to to create ADaM datasets.
  - **Key Concepts**: Standardized tables for clinical trial domains like Demographics (DM), Adverse Events (AE), and Lab Results (LB).
  - **Example**: Patient demographics, adverse events, lab results are all structured in specific datasets.

CDISC. (2021). SDTM: *Study Data Tabulation Model*. Clinical Data Interchange Standards Consortium.
Retrieved from https://www.cdisc.org/standards/foundational/sdtm

Key CDISC Standards

SDTM Principles

Principle: Determine SDTM class (before IG domain)

Principle: Align with SDTM variable definition (before IG domain)

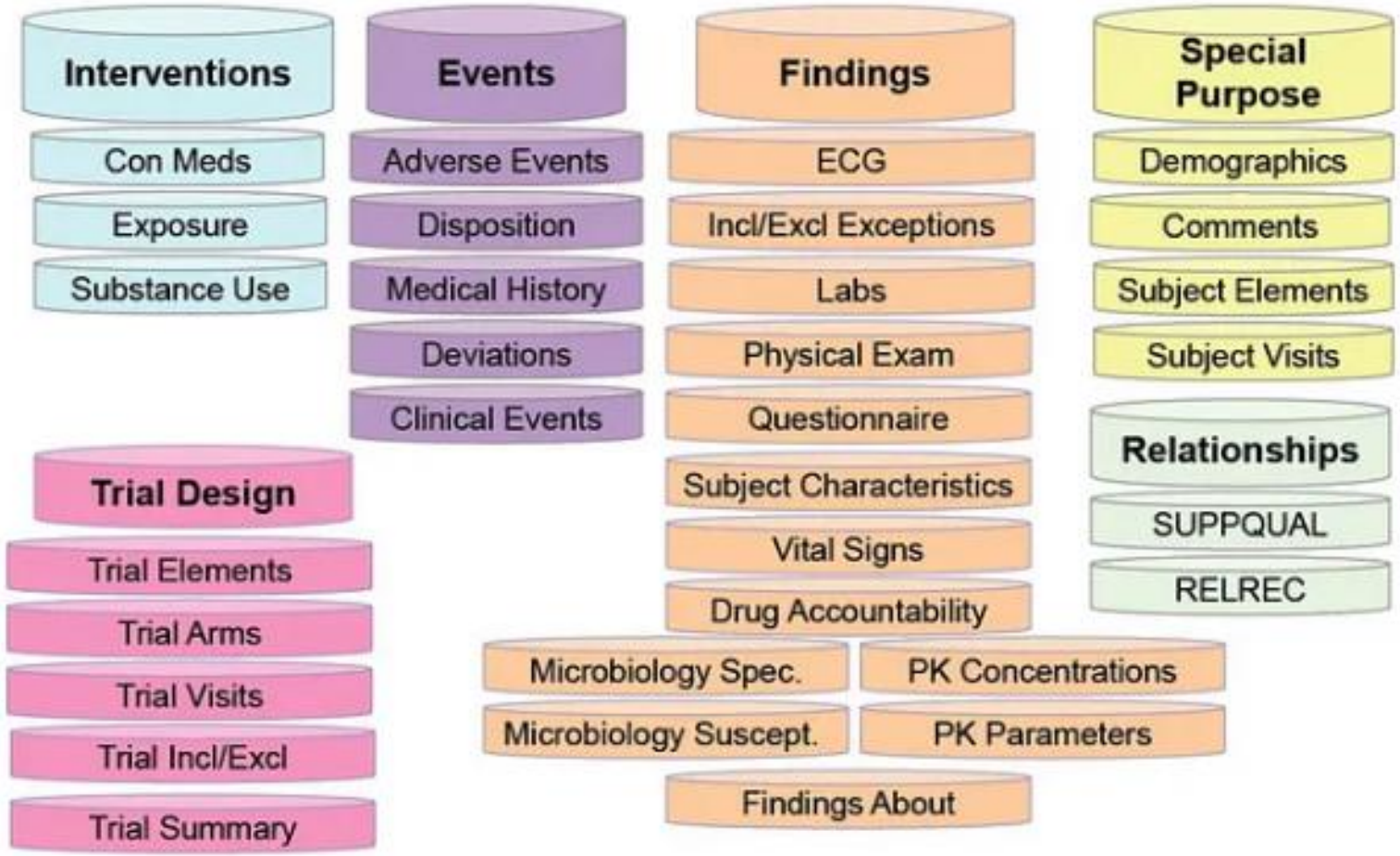
Principle: Align semantics (before IG domain)

Principle: Represent a concept in the same IG domain

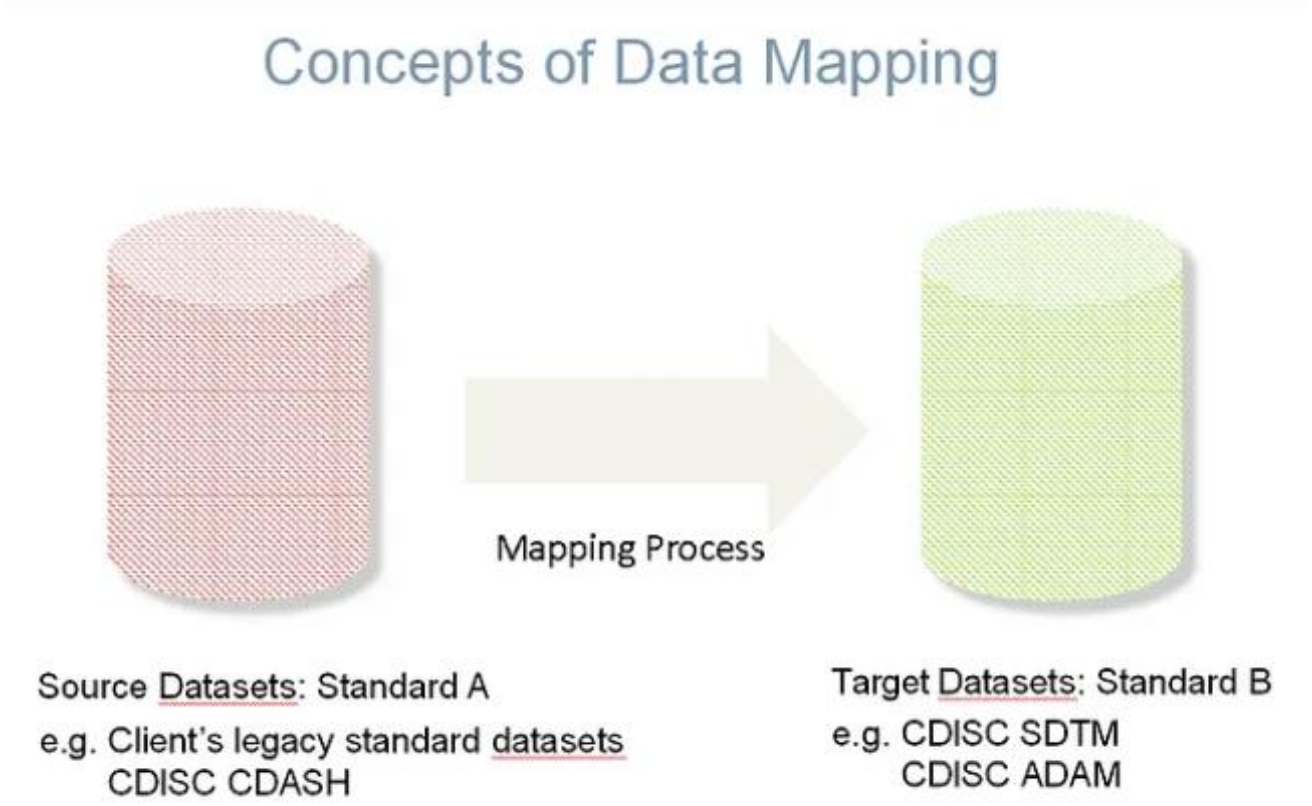
Principle: Preserve the original meaning but standardize the representation

Principle: Consider the impact of changes

SDTM Example

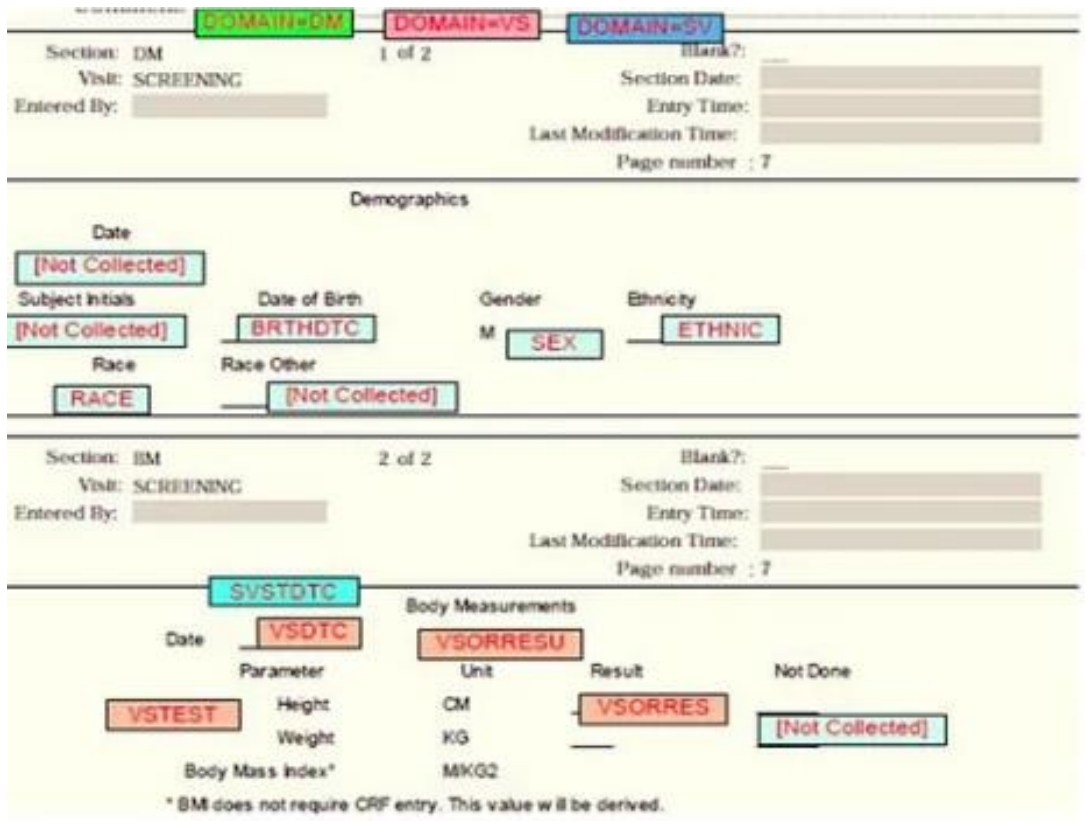


SDTM Example



Key CDISC Standards

- **SDTM Steps:**
  - **SDTM Annotation** – typically provided as a PDF of the CRF
  - **Mapping Table Creation** – with derivation rules
  - **Trial Design Information (*REQUIRED)** – study protocol, SAP, annotated CRFs, SDTMIG
  - **Update SDTM Metadata Repository**
  - **Convert Source -> SDTM (1 script per conversation)**
  - **<u>VALIDATION – peer review, external validators</u>**



SDTM Example

AE - Examples for Adverse Events Domain Model

Example 1

This is an example of data from an AE CRF that collects AE terms as free text. The first study drug was administered to the subject on October 13, 2006 at 12:00. Three AEs were reported. AEs were coded using MedDRA, and the sponsor's procedures include the possibility of modifying the reported term to aid in coding. The CRF is structured so that seriousness category variables (e.g., AESDTH, AESHOSP) are checked only when AESER is answered "Y."

Rows 1-2: Show the following: (1) an example of modifying the reported term for coding purposes. The modified value is in AEMODIFY. (2) An example of the overall seriousness question AESER answered with an "N" and corresponding seriousness category variables (e.g., AESDTH, AESHOSP) left blank.

Row 3: Shows an example of the overall seriousness question AESER answered with a "Y" and the relevant corresponding seriousness category variables (AESHOSP and AESLIFE) answered with a "Y". The other seriousness category variables are left blank. This row also shows an example of AEENRF being populated because the AE was marked as "Continuing" as of the end of the study reference period for the subject ***[see Section 4: 4.1.4.7, Use Of Relative Timing Variables]***.

Row	STUDYID	DOMAIN	USUBJID	AESEQ	AETERM	AESTDTC	AEENDTC	AEMODIFY	AEDECOD
1	ABC123	AE	123101	1	POUNDING HEADACHE	2005-10-12	2005-10-12	HEADACHE	Headache
2	ABC123	AE	123101	2	BACK PAIN FOR 6 HOURS	2005-10-13T13:05	2005-10-13T19:00	BACK PAIN	Back pain
3	ABC123	AE	123101	3	PULMONARY EMBOLISM	2005-10-21			Pulmonary embolism

Row	AEBODSYS	AESEV	AESER	AEACN	AEREL
1 (cont)	Nervous system disorders	SEVERE	N	NOT APPLICABLE	DEFINITELY NOT RELATED
2 (cont)	Musculoskeletal and connective tissue disorders	MODERATE	N	DOSE REDUCED	PROBABLY RELATED
3 (cont)	Vascular disorders	MODERATE	Y	DOSE REDUCED	PROBABLY NOT RELATED

Row	AEOUT	AESCONG	AESDISAB	AESDTH	AESHOSP	AESLIFE	AESMIE	AESTDY	AEENDY	AEENRF
1 (cont)	RECOVERED/RESOLVED							-1	-1	
2 (cont)	RECOVERED/RESOLVED							1	1	
3 (cont)	RECOVERING/RESOLVING				Y	Y		9		AFTER

Questions on concepts?

We are about to start programming!

Setting Up

- Option 1: Local Machine
- Install R – programming language
- Install RStudio Desktop – Integrated Development Environment (IDE)

1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.

DOWNLOAD AND INSTALL R

2: Install RStudio

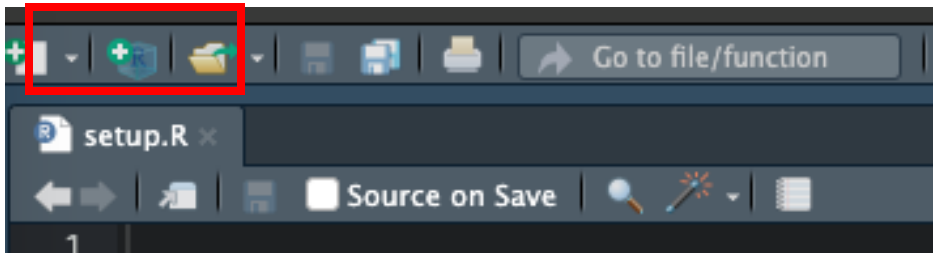
DOWNLOAD RSTUDIO DESKTOP FOR MACOS 12+

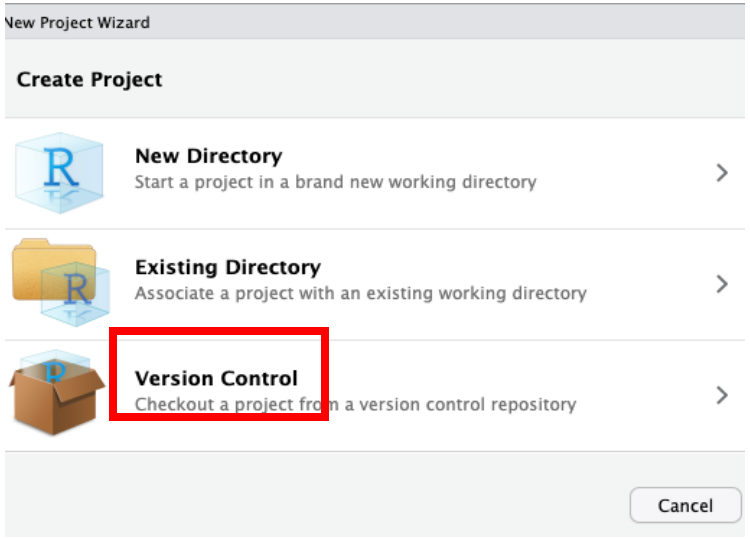
This version of RStudio is only supported on macOS 12 and higher. For earlier macOS environments, please download a previous version.

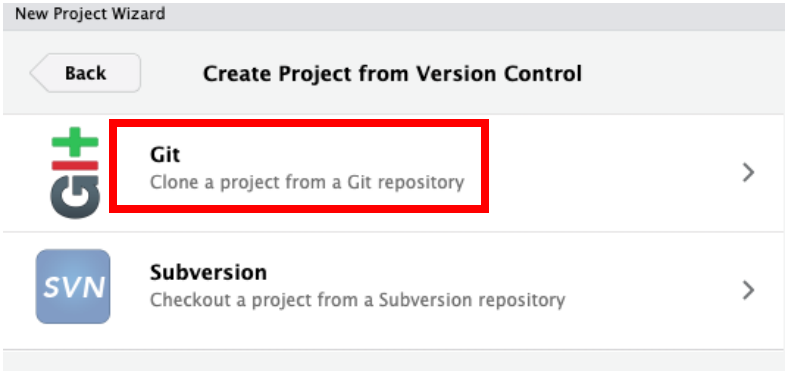
Size: 621.00 MB | SHA-256: 54D722FD | Version: 2024.09.0+375 | Released: 2024-09-23

Setting Up

- Option 1: Local Machine
- Clone GitHub repository to local machine
- https://github.com/jjc54/pharmasug_sde_2024

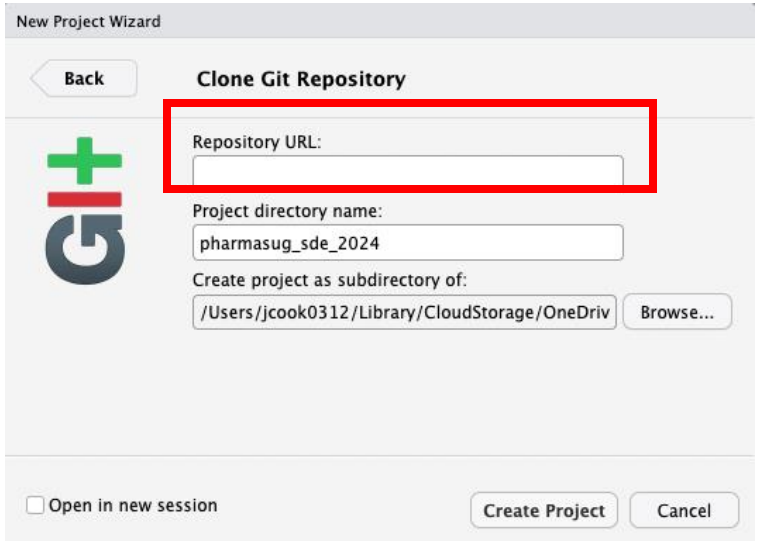


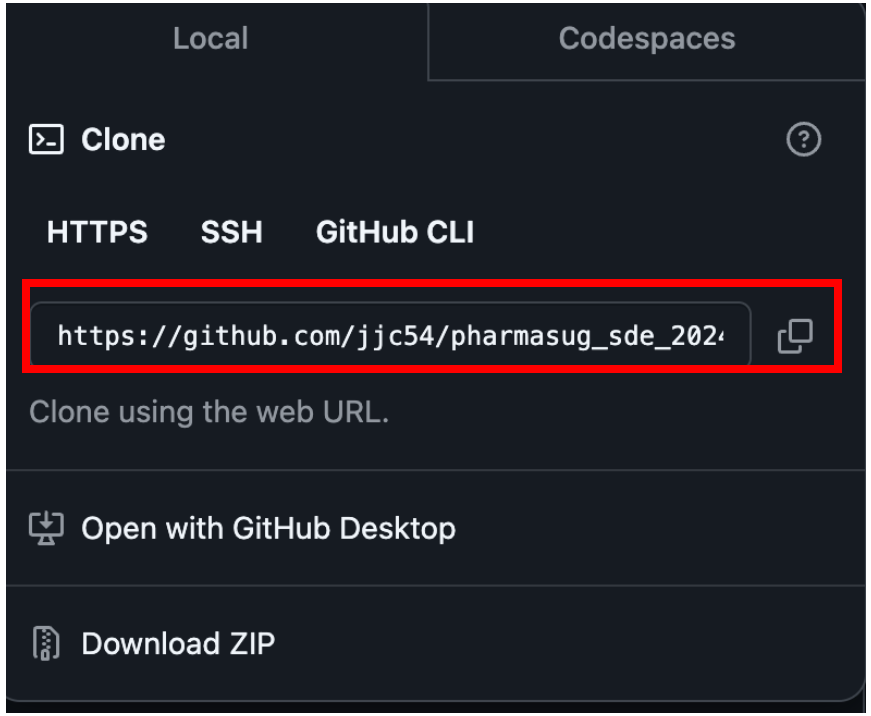




Setting Up

- Option 1: Local Machine
- Clone GitHub repository to local machine
- https://github.com/jjc54/pharmasug_sde_2024.git





Setting Up

Option 2: Posit Workbench

Link posted on the day of the workshop!

Introducing the R tidyverse and dplyr Package

Tidyverse Overview:

- **Tidyverse**: A collection of R **packages designed for data science**.
- **Key Features**:
  - Integrates tools for data manipulation, visualization, and modeling.
  - Ensures a consistent, intuitive syntax across different packages.

dplyr Package:

- **Purpose**: Part of the Tidyverse, **specialized for data manipulation**.
- **Features**:
  - Provides functions like filter(), mutate(), select(), and summarize() to streamline data wrangling.
  - Designed for efficient data frame operations, focusing on readability and ease of use.
- **Benefits**:
- **Simplifies Code**: Clear syntax for transforming datasets.
- **Interoperability**: Works seamlessly with other Tidyverse packages.

Hands-On Example: Mapping Raw Data to SDTM

- **Overview**: Demonstrate **mapping CDASH-compliant CRF raw data to SDTM** format using R.
- **Libraries**: Use {tidyverse} for effective data manipulation.
- **Mock Data**: Simulate demographic data (e.g., AGE, SEX, RACE, RACEOTH).
- **Data Understanding**: Ensure standardization in demographic fields.
- **Mapping Process**: Create DOMAIN, USUBJID, and derive RACERECOD for standardized race categories.
- **Metadata**: Document SDTM variables for submissions.
- **Hands-on Learning**: Practice real-world compliance and data standardization.

The CRF

cdisc

Site Number

Subject Number

Form DM - Demographics

1 DM - Demographics

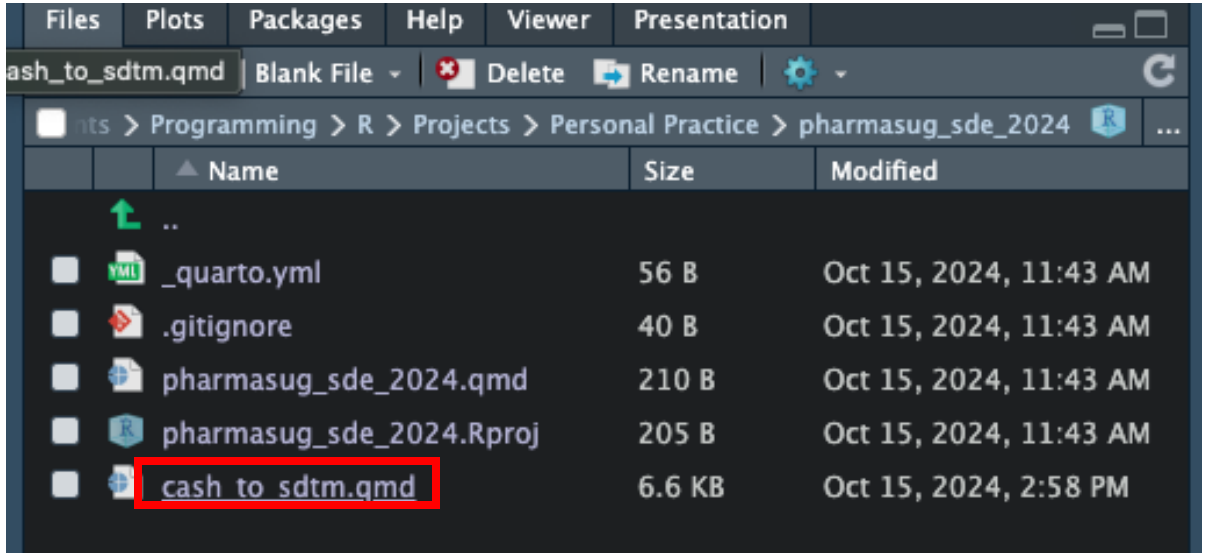
1.1	Birth Date (DD-MMM-YYYY)		***BRTHDAT***
1.2	Age		***AGE***
1.3	Age Unit	Years	***AGEU***
1.4	Sex	[F] Female [M] Male [U] Unknown [UNDIFFERENTIATED] Undifferentiated	***SEX***
1.5	Ethnicity	[HISPANIC OR LATINO] Hispanic or Latino [NOT HISPANIC OR LATINO] Not Hispanic or Latino [NOT REPORTED] Not Reported [UNKNOWN] Unknown	***ETHNIC***
1.6	Race	[AMERICAN INDIAN OR ALASKA NATIVE] American Indian or Alaska Native [ASIAN] Asian [BLACK OR AFRICAN AMERICAN] Black or African American [NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER] Native Hawaiian or Other Pacific Islander [WHITE] White [NOT REPORTED] Not Reported [UNKNOWN] Unknown [OTHER] Other Specify Other Race	***RACE*** ***RACEOTH***

CDASH Example

Section											
OID	Name	Repeating	Description	Order No.	Mandatory	Aliases	Condition	IsReferenceData	Repeating Information	SASDatasetName	Domain
CDASH_2-1_IG_2	DM - Demographics	No	DM - Demographics [en]	1	Yes						DM
Questions											
OID	Name	Text	DataType	Order No.	Mandatory	Terminology	Length	Significant Digits	Units	Description	Aliases
IT.BRTHDAT	BRTHDAT	What is the subject's date of birth? Birth Date [en]	date	1	No					BRTHDAT [en]	BRTHDAT [CDASH] Record the date of birth to the leve of precision known (e.g., day/month/year, year, month/year, etc.) in this format (DD-MON-YYYY) [completionInstruc tions] BRTHDAT is the collected field used for recording the full birth date. The sponsor may choose to database the date of birth as a single variable (BRTHDAT), or as separate variables for each component of the date/time (BRTHYY, BRTHMO, BRTHDD, BRTHTIM). The

Hands-On with R

- Navigate to RStudio/Posit Workbench
- Open **cash_to_sdtm.qmd** for this section of the workshop!



Questions?

5-Minute Break!

Module 2: Analyzing Clinical Trial Data

Key CDISC Standards

- **ADaM (Analysis Data Model)**:
  - **Purpose**: Structures data for **statistical analysis**. While SDTM handles raw data, ADaM prepares the data for analysis (e.g., deriving variables).
  - **Key Concepts**: Analysis-ready datasets, derivations, and variables that are crucial for efficacy and safety assessments. Key to traceability in the analysis process going all the way back to data collection on the CRFs.
  - **Example**: If you want to compare blood pressure before and after treatment, ADaM datasets help calculate and organize this information for analysis.

CDISC. (2021). *ADaM: Analysis Data Model*. Clinical Data Interchange Standards Consortium.
Retrieved from https://www.cdisc.org/standards/foundational/adam

Key CDISC Standards

ADaM Principles

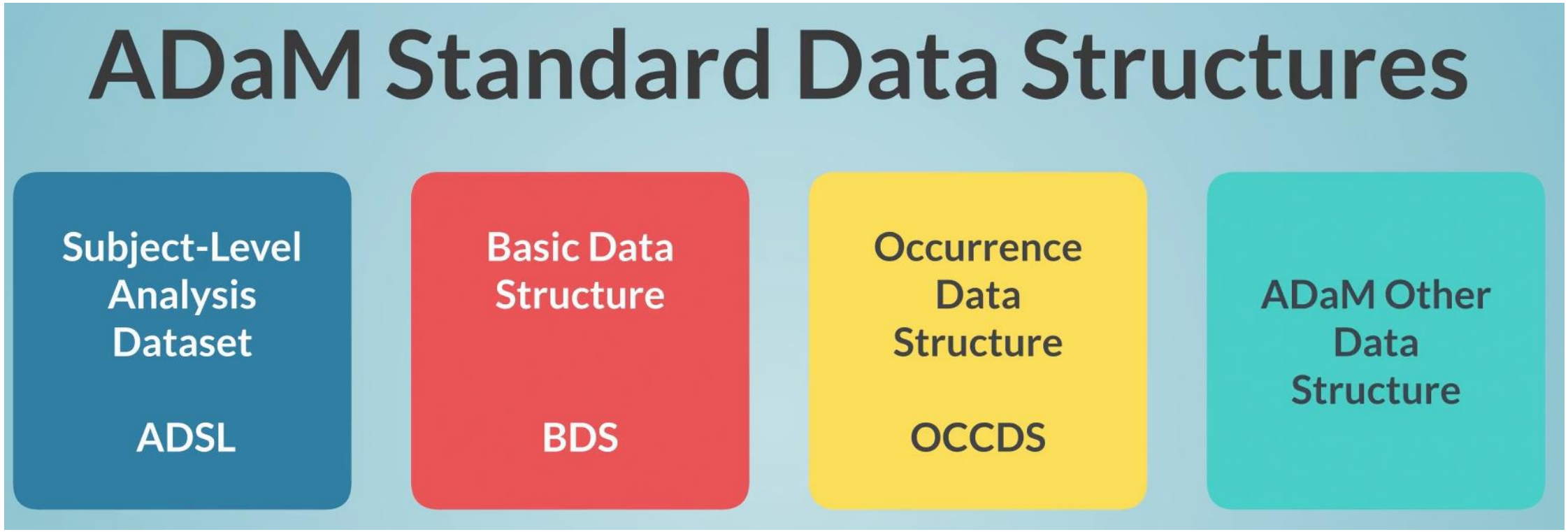
Principle:	Usability
Principle:	Clarity and consistency
Principle:	ADaM documents include metadata
Principle:	Support end-to-end data flow within CDISC
Principle:	Continuous improvement of standards, focusing on priorities

ADaM Example

Four types of ADaM metadata analysis

Datasets	Variables
Parameters	Results

ADaM Example



ADaM Example

Key CDISC Standards

- **Define-XML**:
  - **Purpose**: Metadata file that explains how SDTM and ADaM datasets are structured, providing **essential context** for the data. **REQUIRED** by FDA for every SDTM/ADaM model.
  - **Key Concepts**: Variable definitions, controlled terminology, and dataset annotations.
  - **Example**: Define-XML acts as a data dictionary for regulators, making it easier to interpret submitted data.

CDISC. (2021). *Define-XML: Metadata Submission Guideline*. Clinical Data Interchange Standards Consortium.
Retrieved from https://www.cdisc.org/standards/data-exchange/define-xml

Key CDISC Standards

Define-XML Principles

Principle: Align with foundational standards

Principle: ODM provides basis for end-to-end XML-based interoperability

Principle: Agile development

Principle: Leverage XML technology as much as possible

Principle: Choose solutions that minimize implementation costs

Define-XML Example

	variable	label	type
1	STUDYID	Study Identifier	Char
2	DOMAIN	Domain Abbreviation	Char
3	USUBJID	Unique Subject Identifier	Char
4	SUBJID	Subject Identifier for the Study	Char
5	BRTHDAT	Birth Date of the Subject	Date
6	AGE	Age of the Subject	Num
7	AGEU	Age Unit	Char
8	SEX	Sex of the Subject	Char
9	ETHNIC	Ethnicity of the Subject	Char
10	RACE	Race of the Subject	Char
11	RACEOTH	Other Race (if applicable)	Char
12	RACERECOD	Recoded Race including corrected RACEOTH	Char

Data Structure and Compliance in Clinical Trials

1. From Collection to Submission:

- **Raw Data**: Collected during the trial in its native format following CDASH (e.g., electronic case report forms, medical records).
- **SDTM Mapping**: First step is to map raw data into SDTM-compliant datasets, creating standardized tables for regulatory submission.
- **ADaM Creation**: SDTM data is then transformed into ADaM datasets, which are analysis-ready.
- **Define-XML**: Provides metadata and explanations for the datasets, making it clear to regulators how data was derived and calculated.

Data Structure and Compliance in Clinical Trials

2. Example Flow:

- **Raw Data**: Collected demographics for all trial participants.
- **SDTM**: CDASH-compliant CRF data is structured into DM (Demographics), AE (Adverse Events), and VS (Vital Signs) domains.
- **ADaM**: New datasets are created that show baseline blood pressure, change from baseline, and whether the adverse events are related to treatment.
- **Define-XML**: Document that explains the structure of the SDTM and ADaM datasets, including the variables, derivations, and coding used.

Data Structure and Compliance in Clinical Trials

3. CDISC Validation & Compliance Checks:

- Regulatory agencies, such as the FDA, require that SDTM and ADaM datasets conform to CDISC standards. Non-compliance can lead to delays or rejection of submissions.
- Tools like **OpenCDISC Validator** can help automate these compliance checks before submission, ensuring that the structure, terminology, and metadata follow the standards.

Introducing the R Pharmaverse and admiral Package

R Pharmaverse Overview:

- **Pharmaverse**: A collaborative, open-source community focused on developing **R packages for clinical reporting.**
- **Purpose**: Offers a validated, consistent ecosystem of tools for use across clinical trial workflows.
- **Website**: pharmaverse.org

{admiral} Package:

- **Purpose**: **Facilitates the creation of ADaM (Analysis Data Model) datasets**, ensuring CDISC compliance.
- **Features**:
  - Contains templates and functions to simplify common ADaM derivations.
  - Developed collaboratively, widely used for regulatory submissions.
- **Benefits**:
  - **Efficiency**: Reduces manual coding effort.
  - **Compliance**: Supports CDISC ADaM standards directly.

pharmaverse

admiral

NORTH CAROLINA Pharma SUG SDE
OCTOBER 17, 2024
- VIRTUAL SEMINARS -

Hands-On Example: Mapping SDTM to ADaM using R

- **Overview:** Transform **SDTM-compliant clinical trial data to ADaM format** using R.
- **Required Libraries:** Use {tidyverse} for data manipulation and {admiral} for ADaM creation.
- **Input Data:** SDTM DM dataset from previous steps.
- **Data Understanding:** Derive analysis-ready variables for compliance with ADaM standards.
- **Transformation Process:**
  - Derive AGEGR1 (age groups).
  - Create SAFFL (safety population flag).
- **Metadata Creation:** Document ADaM variables for regulatory submissions.
- **Hands-on Learning:** Practical experience in preparing analysis-ready datasets.

A Review

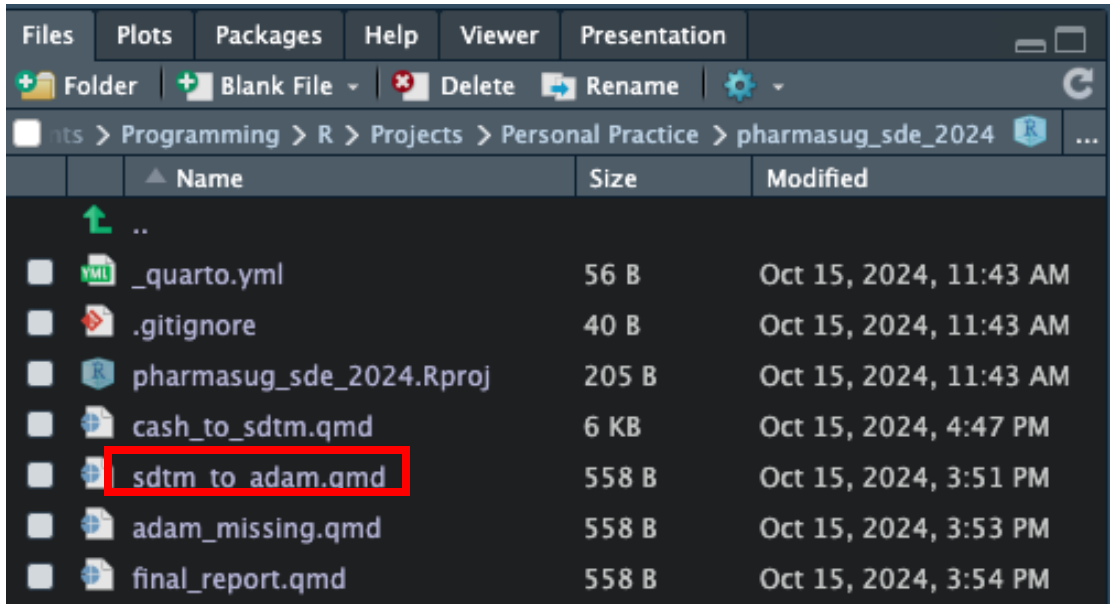
sdtm_dm

Filter

	STUDYID	DOMAIN	USUBJID	SUBJID	BRTHDAT	AGE	AGEU	SEX	ETHNIC	RACE	RACERECOD	RACEOTH
1	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB001	SUB001	1991-08-07	53	Years	F	NOT HISPANIC OR LATINO	NOT REPORTED	NOT REPORTED	NA
2	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB002	SUB002	1991-09-24	60	Years	U	HISPANIC OR LATINO	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	NA
3	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB003	SUB003	1948-03-04	32	Years	F	NOT HISPANIC OR LATINO	UNKNOWN	UNKNOWN	NA
4	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB004	SUB004	1945-01-15	49	Years	U	UNKNOWN	AMERICAN INDIAN OR ALASKA NATIVE	AMERICAN INDIAN OR ALASKA NATIVE	NA
5	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB005	SUB005	1949-03-24	24	Years	M	HISPANIC OR LATINO	NOT REPORTED	NOT REPORTED	NA
6	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB006	SUB006	1971-11-11	26	Years	UNDIFFERENTIATED	HISPANIC OR LATINO	NOT REPORTED	NOT REPORTED	NA
7	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB007	SUB007	1953-01-12	58	Years	U	NOT HISPANIC OR LATINO	OTHER	OTHER	NA
8	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB008	SUB008	1958-06-20	40	Years	UNDIFFERENTIATED	NOT REPORTED	BLACK OR AFRICAN AMERICAN	BLACK OR AFRICAN AMERICAN	NA
9	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB009	SUB009	1984-02-26	44	Years	UNDIFFERENTIATED	NOT REPORTED	NOT REPORTED	NOT REPORTED	NA
10	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB010	SUB010	1947-07-19	77	Years	M	HISPANIC OR LATINO	WHITE	WHITE	NA
11	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB011	SUB011	1998-11-02	70	Years	M	UNKNOWN	OTHER	OTHER	NA
12	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB012	SUB012	1974-08-05	24	Years	U	NOT HISPANIC OR LATINO	UNKNOWN	BLACK OR AFRICAN AMERICAN	Black American
13	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB013	SUB013	1965-03-18	70	Years	UNDIFFERENTIATED	HISPANIC OR LATINO	BLACK OR AFRICAN AMERICAN	BLACK OR AFRICAN AMERICAN	NA
14	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB014	SUB014	1967-12-09	44	Years	M	UNKNOWN	UNKNOWN	UNKNOWN	NA
15	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB015	SUB015	1977-06-01	55	Years	M	NOT HISPANIC OR LATINO	NOT REPORTED	NOT REPORTED	NA
16	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB016	SUB016	1947-11-27	51	Years	U	HISPANIC OR LATINO	ASIAN	ASIAN	NA
17	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB017	SUB017	1956-11-21	80	Years	U	NOT REPORTED	UNKNOWN	WHITE	Caucasian
18	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB018	SUB018	1991-11-19	30	Years	UNDIFFERENTIATED	NOT REPORTED	WHITE	WHITE	NA
19	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB019	SUB019	1966-05-25	42	Years	F	UNKNOWN	WHITE	WHITE	NA
20	CDASH_DEMO_01	DM	CDASH_DEMO_01-SUB020	SUB020	1974-03-21	55	Years	M	UNKNOWN	OTHER	OTHER	NA

Hands-On with R

- Navigate to RStudio/Posit Workbench
- Open **sdtm_to_adam.qmd** for this section of the workshop!



Questions?

5-Minute Break!

Module 3: Handling Missing Data

Introduction to Missing Data in Clinical Trials

- **Definition**: Missing data refers to instances **where information is not recorded or available. VERY common in clinical trials.**
- **Types of Missing Data**:
  1. **MCAR (Missing Completely at Random)**: Missingness is unrelated to observed or unobserved data.
  2. **MAR (Missing at Random)**: Missingness depends on observed data but not on unobserved data.
  3. **MNAR (Missing Not at Random)**: Missingness depends on unobserved data itself.

Impact of Missing Data on Clinical Trials

Statistical Impact:

- Loss of statistical power.
- Potential biases in the estimation of treatment effects.

Regulatory Concerns:

- Regulatory bodies, like the FDA and EMA, emphasize complete and accurate data.
- Missing data may affect the integrity of clinical trial findings.

Illustrative Example:

- Loss of key outcome data affecting the results' reliability.

Common Causes of Missing Data

Patient Factors:	Operational Factors:	Example in Clinical Trials:
• Withdrawal from the study. • Non-compliance (e.g., missed appointments).	• Data entry errors. • Instrument failure (e.g., faulty device measurements).	• Patients experiencing side effects may drop out, resulting in systematic missingness.

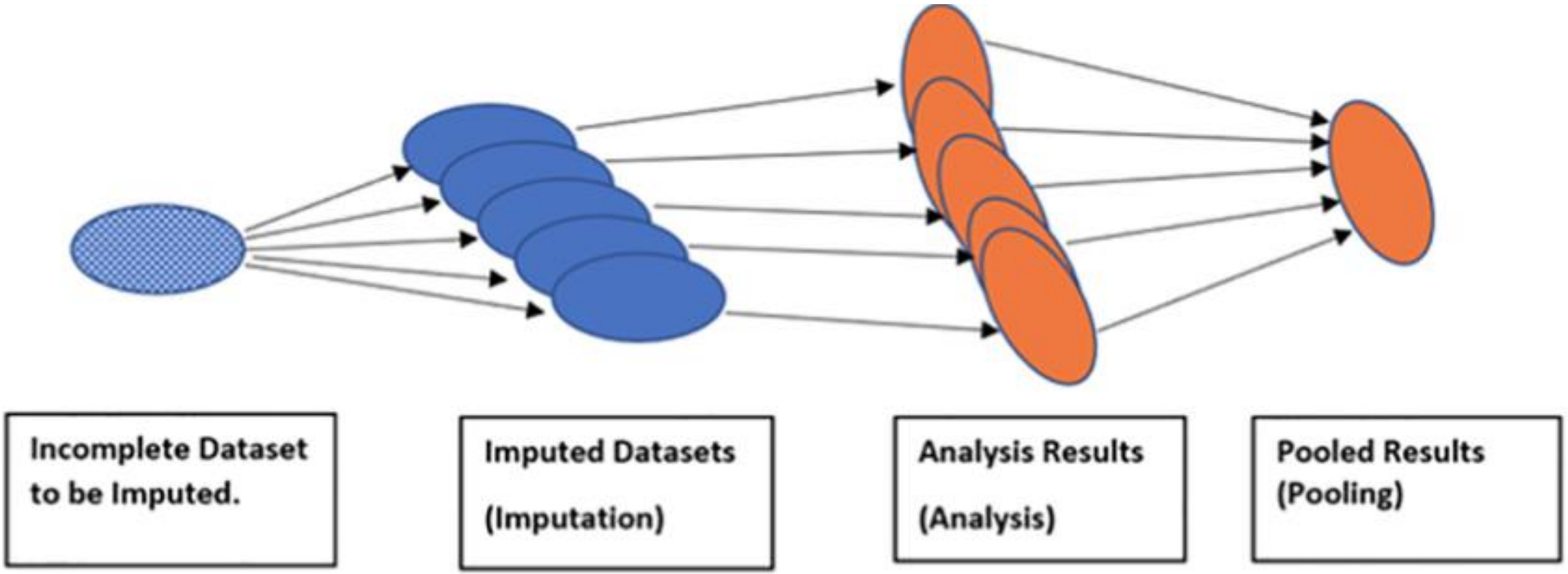
Introduction to Multiple Imputation

- **Concept**: **Multiple Imputation (MI)** replaces missing data points with a set of plausible values, thereby creating multiple complete datasets.
- **Why Multiple?**: It accounts for the uncertainty of missing values by using variability between imputations.
- **Benefits**: Reduces bias; More accurate parameter estimates compared to single imputation methods.

A Peer-Reviewed Tutorial

Key Steps in Multiple Imputation

- **Imputation Phase**: Generate multiple imputed datasets using observed data.
- **Analysis Phase**: Analyze each imputed dataset separately.
- **Pooling Phase**: Combine results across imputed datasets to derive estimates.



Introduction to the {mice} Package in R

- **What is mice?**: mice stands for **M**ultivariate **I**mputation by **C**hained **E**quations.
- Widely used package for handling missing data in R.
- **Advantages**:
  - Handles different types of missing data effectively.
  - User-friendly syntax and customizable methods for imputation.

{mice}

Implementing Multiple Imputation with {mice}

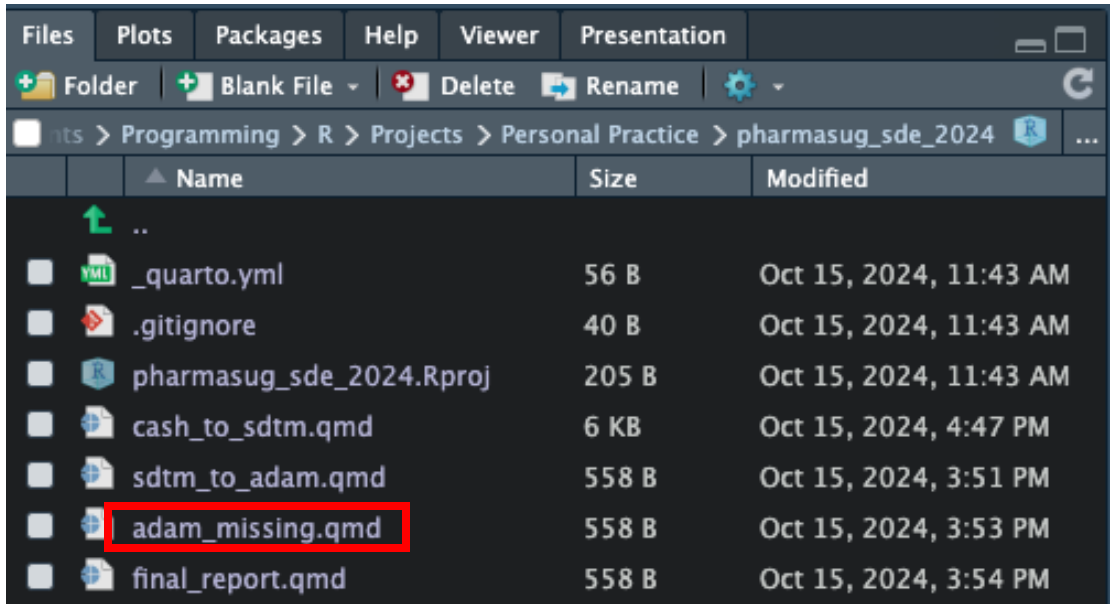
- **Step-by-Step Walkthrough**:**Inspect Missingness**:
  - Use md.pattern() to explore missing data patterns.
- **Perform Imputation**:
  - Apply mice() to create multiple complete datasets.
- **Analyze Imputed Datasets**:
  - Fit models to each dataset using standard R functions.
- **Pool Results**:
  - Use pool() to combine the results for final inferences.

Hands-On Example: Imputing Missing Data Using Clinical Trial Data

- **Overview**: Transform SDTM-compliant clinical trial data to ADaM format using R and **handle missing data with multiple imputation**.
- **Required Libraries**: Use {tidyverse} for data manipulation, {admiral} for ADaM dataset creation, and {mice} for handling missing data.
- **Input Data**: **SDTM DM Dataset**: Start with the SDTM dataset we previously generated.
- **Hands-on Exercise**:
  1. **Generate Missing Data**:
     - Randomly set some AGE and SEX values to NA in the original SDTM DM dataset.
  2. **Attempt to Map to ADaM Format:**
     - Redo the ADaM mapping as before to see the new missing values and their effects.
  3. **Apply Multiple Imputation with mice**:
     - Use the mice package to impute missing values.
  4. **Recreate ADaM Dataset**:
     - After imputations, derive AGEGR1 and SAFFL in the imputed dataset.
  5. **Compare Results**:
     - Compare the original ADaM dataset (with missing values) to the imputed dataset to understand the effect of missing data handling.

Hands-On with R

- Navigate to RStudio/Posit Workbench
- Open **adam_missing.qmd** for this section of the workshop!



Questions?

5-Minute Break!

Module 4: Reporting and Reproducible Research

Introduction to Quarto

- **What is Quarto?**:
  - Quarto is a next-generation, open-source scientific and technical publishing system; maintained by Posit.
  - Built to **extend the functionality of R Markdown.**
- **Capabilities**:
  - Supports multiple programming languages: R, Python, Julia, etc.
  - Enables creating documents, presentations, dashboards, and interactive content.
  - Quarto **Project Types** refer to a type of organization of files (ex: report, website, presentation) while Quarto **Project Formats** refer to the output file type of a document (ex: HTML, PDF, DOCX)
- **Advantages Over R Markdown**:
  - More powerful for cross-language workflows.
  - Enhanced options for layout, citations, and reproducibility.

An Introduction to Quarto - PharmaSUG 2024

Advantages of Quarto for Clinical Reporting

Reproducibility: Supports literate programming, combining code, output, and explanations in one document.

Multi-format Support: Outputs to HTML, PDF, Word, presentations, and more.

Cross-language Integration: **Can use both R and Python** in the same document, ideal for clinical data analyses involving multiple tools.

Comparison: Quarto vs R Markdown

- **Language Support**:
  - **Quarto**: Multi-language (R, Python, Julia).
  - **R Markdown**: Primarily R-focused, limited Python support.
- **Flexibility**:
  - **Quarto**: Enhanced theming, layout customization, and cross-format publishing.
  - **R Markdown**: Good for basic reports but less versatile.
- **Use Cases**:
  - **Quarto: ideal for clinical trials**, research papers, dynamic dashboards.
  - **R Markdown:** former industry standard, but no longer being updated

Building Reproducible Reports for Clinical Trial Data

- **Scenario**: Creating a report summarizing key clinical trial data and results.
- **Features**:
  - Data import and analysis in R.
  - Interactive elements (e.g., parameterized reports).
  - Output format: HTML for easy sharing with stakeholders. Can simultaneously output to PDF or other formats.
- **Steps**:
  1. **Import Data**: Load clinical trial dataset in R.
  2. **Perform Analysis**: Derive statistics (e.g., demographics, primary outcomes).
  3. **Generate Report**: **Embed** both code and results, allowing reproducibility. This is dynamically pulling code and results from other files. **The report is just a shell.**

Customization and Advanced Features

- **Layout and Theming**:
  - Customize themes for better readability.
  - Use YAML options to control layout (format: html).
- **Code Execution**: Customize chunk options (eval, echo) to control what appears in the final report.
- **Interactivity**: Use parameters to make reports dynamic (e.g., filter by different sites or demographics).

Quarto Guide
Beautiful Examples

Best Practices for Reproducible Reporting

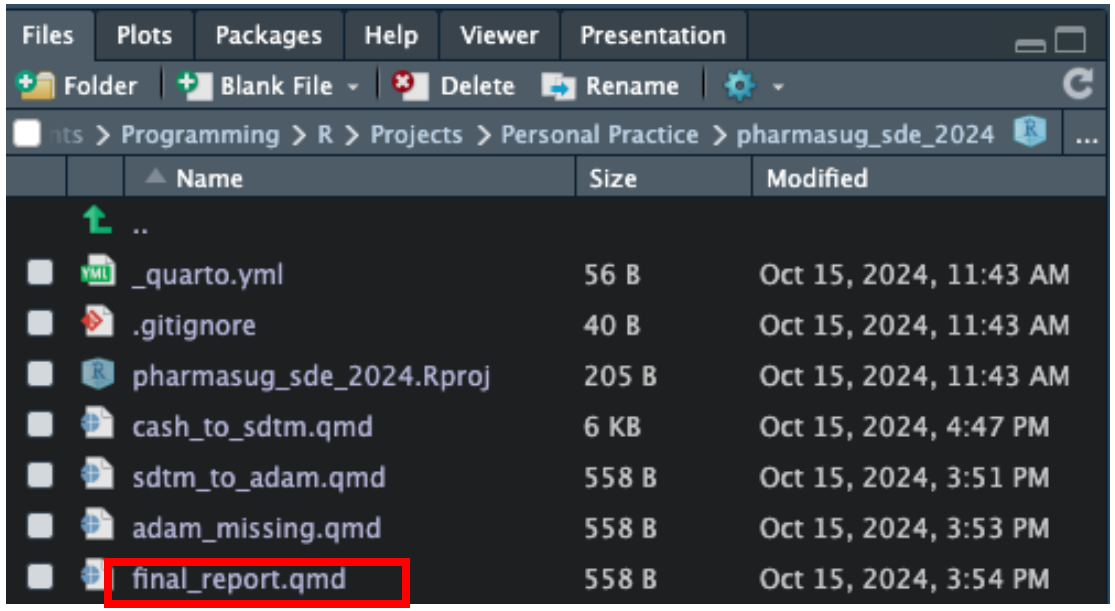
- **Modularize Code:** have separate folders for data, code, and output. Embed these together in your final report.
- **Use Version Control**: Track changes with Git to manage edits to files.
- **Document Everything**: Clearly document each step using comments to make your report understandable to stakeholders.
- **Organization Tips:**
  - Don’t reuse/overwrite object names
  - Be descriptive with object names
  - Comment out quality checks
  - Use rm() to remove older objects that you don’t need

Hands-On Example: Imputing Missing Data Using Clinical Trial Data

- **Overview**: Use the **complete CDISC-compliant SDTM and ADaM datasets to create a final report using Quarto**. This report will include tables, listings, and figures (TLFs), which are commonly required for regulatory submissions.
- **Required Libraries**: {tidyverse} for data manipulation and figures, {gt} and {gtsummary} for listings and tables, and {quarto} for generating the final reproducible report.
- **Input Data**: **Complete SDTM and ADaM Datasets**

Hands-On with R

- Navigate to RStudio/Posit Workbench
- Open **final_report.qmd** for this section of the workshop!



Final Questions & Closing Remarks

Recommended Resources

- R for Data Science (2e) - Tidyverse Primer
- CDASH Primer
- CDASHIG
- SDTM Primer
- SDTMIG
- ADaM Primer
- ADaMIG
- Pharmaverse Primer
- Admiral Primer
- mice Documentation
- Quarto Primer
- An Introduction to Quarto - PharmaSUG 2024
- Quarto Guide
- Beautiful Quarto Examples

References

- U.S. Department of Health and Human Services. 2024. “45 CFR 46.102: Definitions.” *Code of Federal Regulations*, Title 45, Part 46.
- Chow, S., Liu, J., Chow, S., and Chow, S. 2013. *Design and Analysis of Clinical Trials: Concepts and Methodologies*.
- Clinical Data Interchange Standards Consortium (CDISC). Year. *CDISC Standards Overview*. CDISC.
- Wickham, H., and Grolemund, G. Year. *R for Data Science (2e) - Tidyverse Primer*.
- Pharmaverse. 2024. *Pharmaverse Overview*. Retrieved from https://pharmaverse.org.
- admiral. 2024. *admiral: ADaM Creation in R*. Retrieved from https://pharmaverse.github.io/admiral.
- van Buuren, Stef. 2024. *mice: Multivariate Imputation by Chained Equations*. Retrieved from https://cran.r-project.org/package=mice.
- Quarto. 2024. *Quarto Guide*. Retrieved from https://quarto.org.
- Iannone, Richard, et al. 2024. *gt: Grammar of Tables in R*. Retrieved from https://gt.rstudio.com.
- Sjoberg, Daniel D., et al. 2024. *gtsummary: Presentation-Ready Data Summary and Statistical Tests*. Retrieved from https://www.danieldsjoberg.com/gtsummary.

BUT WAIT… There's More

Join me virtually on November 1st at R / Pharma for a **FREE workshop** on:

- *"Visualizing Clinical Trial Data: Foundational Principles for Data Visualization, Best Practices, and Programming Techniques in R by Example."*
  - Key conceptual practices from visualization experts like **Stephen Few** and **Kirk Paul Lafler.**
  - Apply these principles to create compelling survival, failure, and swimmer plots, as well as other essential graphs using R by example.
  - Drawing from the acclaimed *"SAS Graphics for Clinical Trials by Example"* by **Harris** and **Watson**.
  - Gain practical skills in utilizing R packages like ggplot2 and survminer to produce high-quality graphics that meet the pharmaceutical industry's rigorous standards.

NORTH CAROLINA Pharma SUG SDE
OCTOBER 17, 2024
- VIRTUAL SEMINARS -

Acknowledgements and Contact

Joshua J. Cook thanks **Margaret Hung**, **Gary Moore**, **Pradeep Bangalore**, **Priscilla Gathoni**, and his mentors, **Richann Watson**, **Louise Hadden, Dr. Achraf Cohen, Dr. Swann Adams, Kirk Paul Lafler, Troy Martin Hughes, and Dr. Lisa Mendez.**

Contact **Josh**:
jcook0312@outlook.com

NORTH CAROLINA Pharma SUG SDE
OCTOBER 17, 2024
- VIRTUAL SEMINARS -