

Welcome to the {iddoverse}: An R package for Converting IDDO-SDTM Data to Analysis Datasets

Rhys Peploe^{1,2}, Kasia Stepniewska^{1,2}, James Watson^{1,2,3} and Prabin Dahal^{1,2}

¹ Infectious Disease Data Observatory, University of Oxford, UK; ² Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK; ³ Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam

Background

- SDTM provides a coherent framework for the storage, standardisation and pooling of studies and clinical trials data.
- Using SDTM for heterogeneous historical studies requires customisation in the standard implementation at IDDO.
- However, the SDTM format is not as accessible for researchers in LMICs due to training, time and resource limitations.

SDTM - Study Data Tabulation Model

IDDO - Infectious Disease Data Observatory

LMICs - Low- and Middle-Income Countries

Objectives

- Creation of an open-source R package to convert SDTM data into analysis datasets, using IDDO implementation.
- Reduce the amount of duplicated work and prompt reproducible outputs.
- The effectiveness of packages like {admiral} (the Pharmaverse) are limited since our implementation is not compatible, and our audience are non-regular SDTM users.
- Provide additional features to improve data use, such as recalculating BMI results and including Child Growth Standards.

BMI – Body Mass Index

Process for PREP functions

- PREP_DM, PREP_LB_BL, PREP_VS_FU are examples of functions in the package.
- For each domain, they convert blanks to NA and results to upper case, character class.
- Filter variables of interest.
- Amalgamate data using standardised results or decoded terms, and where NAs exist fill with modified or original results.
- Pivot the domain wider, transforming the long data to a wide data format, with tests or terms as columns.
- Clean column names.

Package

Take a Laboratory (LB) domain, curated with our implementation of SDTM, subset to include essential information.

USUBJID	LBTESTCD	LBORRES	LBORRESU	LBSTRESC	LBSTRESN	LBSTRESU	VISITNUM	VISITDY	LBDY	EPOCH
RPTESTD_001	HGB	95	g/L	95	95	g/L	1	1	1	BASELINE
RPTESTD_001	PLAT	181000	10^6/L	181	181	10^9/L	1	1	1	BASELINE
RPTESTD_001	HGB	88	g/L	88	88	g/L	2	3	3	TREATMENT
RPTESTD_002	HGB	10100	mg/dL	101	101	g/L	1	1	2	BASELINE
RPTESTD_002	PLAT	100000	10^6/L				1	1	2	BASELINE
RPTESTD_002	HGB	99	g/L	99	99	g/L	2	3	4	TREATMENT

- Investigators manipulate the SDTM data into an analysis dataset, however, generated datasets are inconsistent and often cumbersome to reproduce.
- This process is heavily dependant on their coding ability and CDISC knowledge.
- The difference between VISITDY and LBDY, or LBSTRESN and LBORRES for example are occasionally misunderstood.
- With the {iddoverse}, users can transform individual domains separately, or combine them using the ANALYSE_ functions.
- Below is the transformed, truncated, dataset; providing a condensed, clear and easy to analyse data frame.
- This output can be achieved by executing one command in R. The code and data used in this poster are available in the {iddoverse} GitHub.

USUBJID	AGE	SEX	ARMCD	VISITNUM	VISITDY	DAY	HGB	HGB_UNITS	PLAT	PLAT_UNITS	HEIGHT	HEIGHT_UNITS	WEIGHT	WEIGHT_UNITS
RPTESTD_001	67	F	PBO	1	1	1	95	g/L	181	10^9/L	167	cm	NA	NA
RPTESTD_001	67	F	PBO	2	3	3	88	g/L	NA	NA	NA	NA	60	kg
RPTESTD_002	18	M	TRT	1	1	2	101	g/L	100000	10^6/L	143	cm	NA	NA
RPTESTD_002	18	M	TRT	2	3	4	99	g/L	NA	NA	NA	NA	42	kg

- DM, LB, VS domains merged into a single dataset, one row per person, per day.
- Units displayed to ensure minimal confusion; not assuming standardised units.
- Takes LBSTRESN results as default, when missing (i.e. RPTESTD_002 PLAT), other results and units (LBORRES/LBORRESU) are used to maximise data use.
 - DISEASE parameter pre-selects certain variables important to that theme.
 - VARS parameter allows additional variables to be included.

Conclusion & Future Work

- Generation of standardised analysis datasets, using custom SDTM implementation, is possible, demonstrating ability to speed up exploratory data analysis and making the SDTM format more accessible to researchers unfamiliar with the format.
- Consequently, if the IDDO repository data is more digestible, people will be more confident requesting and using data, leading to a greater quantity of high quality research being produced in the infectious disease community.
- By renaming column names to longer, descriptive title, the features are more understandable to a wider audience.
- Number of functions, diseases and domains in the package will be expanded, as well as, introducing additional features.
- Time complexity is an area for improvement, the package performs well for under 50,000 subjects, but struggles at larger quantities.



Check out our GitHub repository for the {iddoverse}







