PharmaSUG 2023 - Paper SD-98

An Introduction to the Development of a Resistance Dataset

Jenny Zhang, Merck & Co., Inc., Rahway, NJ, USA;
Shunbing Zhao, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

Antimicrobial resistance (AMR) is increasingly being recognized as a global threat to public health. Microbiology data provides important information that is used to guide clinical development of a new investigational drug.

We had an opportunity to develop a special resistance analysis dataset (ADRST) for an infectious disease clinical trial. In contrast to other typical ADaM datasets, ADRST was different, and it was challenging to make it analysis ready. In this paper, we will introduce key variables derived in the dataset following FDA guidance on Antiviral Product Development, and showcase SAS codes for genotypic data, providing positioning variables for hundreds of amino acid sequence data for PR (Protease) and RT (Reverse Transcriptase) endpoints. We will also highlight the challenges of creating PR and RT related specifications and how programming steps were implemented.

INTRODUCTION

HIV drug resistance is an example for AMR and is caused by changes in the genetic structure of HIV that affect the ability of medicines to block the replication of the virus. Increased use of HIV medicines has been accompanied by the emergence of HIV drug resistance, the levels of which have steadily increased in recent years. All antiretroviral drugs, including those from newer drug classes, are at risk of becoming partially or fully inactive due to the emergence of drug-resistant viruses.

FDA has drafted guidance document “Antiviral Product Development- Conducting and Submitting Virology Studies to the Agency” which assists sponsors in the development of antiviral drugs and biological products (i.e., therapeutic proteins and monoclonal antibodies) from the initial pre-IND through the new drug application (NDA) and post marketing stages.

The resistance dataset is one of the key datasets required for the antiviral product development submission. In the following sections of the paper, we will share details of creating the resistance dataset, key variables required as per the FDA guidance document. We will also share the challenges that the team faced when developing this dataset which had more than 800 variables.

RESISTANCE DATASET

The resistance dataset ADRST followed ADaM other data structure. This dataset can be categorized into four key sections - patient data, endpoint data, genotypic data, and phenotypic data, The dataset has one record (row) per patient per isolate (separation of a HIV virus strain).

Below are important variables required in the dataset as per FDA guidance document:

1) Patient data

Table 1

Variable Name	Variable Label	Type	Length
STUDYID	Study Identifier	Char	12
USUBJID	Unique Subject Identifier	Char	30
ARM	Description of Planned Arm	Char	100
AVISIT	Analysis Visit	Char	30
ADY	Analysis Relative Day	Num	8
EXTRTHIV	Concomitant HIV-1 Treatment Drugs	Char	50

ISOLDY	Isolate Sample Collection Relative Day	Num	8
ISOLID	Unique Identifier for Isolate	Char	50
ISOLDTC	Date of Isolate	Char	19
HIVGTSC	HIV-1 Clade at Screening	Char	12
HBVCOINF	Hepatitis B Co-infected	Char	1
HCVCOINF	Hepatitis C Co-infected	Char	1

Most of the patient data related variables are retained from ADSL dataset. A few of them are derived. For example, ISOLDY is an isolate sample collection relative day from the lab sample. ISOLID is a unique identifier for isolate and is derived using SUBJID, ISOLDY and assay name. Assay name is a test name typically for Genotypic test and Phenotypic test.
Example: SUBJID="12345", ISOLDY="111", ASSAY Name="GENOSURE", ISOLID="12345|111| GENOSURE".

2) End Point Data

Table 2

Variable Name	Variable Label	Type	Length
HIVVLBL	Baseline HIV RNA (cp/mL)	Num	8
LOGVLBL	Baseline HIV RNA (log10 copies/mL)	Num	8
HIVVL	HIV-1 RNA (copies/mL)	Num	8
LOGHIVVL	HIV-1 RNA (log10 copies/mL)	Num	8
HIVVLW16	HIV-1 RNA at Week 16	Num	8
HIVVLW24	HIV-1 RNA at Week 24	Num	8
HIVVLEOT	HIV-1 RNA (copies/mL) at End of TRT	Num	8
LOGVLEOT	HIV-1 RNA (log10 cp/mL) at End of TRT	Num	8
EFFICFL	Achieved HIV RNA <**5**0 at Wk48	Char	1
NONRECAT	Failure Category	Char	20
DISCTXFL	Discontinuation Flag	Char	20
VFFL	Protocol-Defined Virologic Failure Flag	Char	8

HIVVL, HIVVLW16 and HIVVLW24 are derived from SDTM Microbiology Specimen (MB) dataset.
Variable VFFL is derived from ADaM Virologic Failure Dataset (ADVF).
VFFL is a flag (Y or NULL) used to indicate the specific study visit in which the subject met the criteria for protocol-defined virologic failure (e.g., rebound, end of treatment).

3) Genotypic Data

Table 3

Variable Name	Variable Label	Type	Length
RTyxxx	Rev. Transcriptase Amino Acid Pos.	Char	30
PRyxxx	Protease Amino Acid Position	Char	30
GENOMET	Genotypic Assay Name	Char	30
TOTNNRTI	Total NNRTI Substitutions	Num	8
TOTNRTI	Total NRTI Substitutions	Num	8

TOTPI	Total PI Substitutions	Num	8
GENOFAIL	Genotypic Test Failed	Char	1
RESISTFL	Resistance Data Reported	Char	1
RESBLFL	BSL Resistance Data Reported	Char	1
RESEOTFL	Last Resistance Data Reported	Char	1

Genotypic resistance testing examines the genetic structure (genotype) of a patient's HIV. A blood sample is taken from the patient, and the HIV is analyzed for the presence of specific genetic mutations that are known to cause resistance to specific drugs.
Genotypic data should be provided for the HIV-1 target, one amino acid per column, with the wild-type (WT) amino acid in the column heading. Changes from WT standard sequence should be indicated in the row (i.e., RTyxxx or PRyxxx blank data indicates no change).

Genotype testing is performed at the time of diagnosis to detect transmitted resistant mutants, generally non-nucleoside reverse transcriptase inhibitor (NNRTI), nucleoside reverse transcriptase inhibitor (NRTI) and protease Inhibitors (PI) resistance.

NNRTI and NRTI resistance usually include reverse transcriptase (RT) mutation, PI resistance includes protease (PR) mutation.

For reverse transcriptase (RT) mutation related variables we utilized the RTyxxx format. For example, RTK065 - RT represents reverse transcriptase; K indicates amino acid code and “065” is the position of the amino acid code.

For protease (PR) mutation related variable, we utilized PRyxxx format. For example, PRL090, “PR” represents protease mutation, “L” indicates amino acid code and “090” is the position of the amino acid code.
Let us consider an example variable RTK103 with a value =”Q”. Here it means there is a mutation in HIV's reverse transcriptase gene. In position 103, the amino acid K has been replaced by amino acid Q.

CHALLENGES FOR CREATING RTYXXX AND PRYXXX VARIABLES

PR and RT amino acid sequence codes are obtained from clinical team. There are 560 amino acid code positions for RT, and 99 amino acid code positions for PR. If we have one variable for each amino acid code position, then there will be 659 variables that need to be created in the specification, and in the dataset. To keep it simple, in the specifications, there are only two rows created for variable RTyxxx and PRyxxx.
It is a difficult task to transform RTyxxx and PRyxxx into 560 RT variables and 99 PR variables.
We create 560 amino acid dummy variables for RT and 99 amino acid dummy variables for PR by using the below SAS code. We did not create a placeholder for value 0 in the variable names. For PR variable example, amino acid code=P and position=1 the variable name is PRP1.

```
data aaseq0(keep=vname val);
  length seq $2000 vname $10;
  val=' ';

** 99 amino acid sequence code for PR ** In the above sequence
  the first amino acid code is represented by “P, the second is “Q”,
  the third is “I”, all the way to 99 which is “F”;

seq='PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAI
GTVLVGPTPVNIIGRNLLTQIGCTLNF';

  do i=1 to length(seq);
```

```
    vname='PR'||substr(seq,i,1)||left(put(i,best.));
    output;
  end;


** 560 amino acid sequence code for RT.
 Since it is too long to have 560 position codes in one variable it is split
into 4 sequences (seq1 to seq4) **
**For RT, the first amino acids code is "P', the second is "I", the third is "S", all
the way to 560th which is "L".
;
seq1='PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKW
RKLVDFRELNKRTQDFWEVQLGIPHPAGLKQKKSVTVLDVGDAYFSVPLDKDFRKYTAFTI';

seq2='PSINNETPGIRYQYNVLPQGWKGSPAIFQCSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIE
ELRQHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKDSWTVNDIQKLVGKLNWASQIYAGIKVRQLC
KLLRGTKALTEVVPLTEEAELELAE';

seq3='NREILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMKGAHTNDVKQLTEAVQKIA
TESIVIWGKTPKFKLPIQKETWEAWWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIIGAETFYVDGAANRETKLG
KAGYVTDRGRQKVVPL';
seq4='TDTTNQKTELQAIHLALQDSGLEVNIVTDSQYALGIIQAQPDKSESELVSQIIEQLIKKEKVYLAWVPAHK
GIGGNEQVDKLVSAGIRKVL';

  seq=seq1||seq2||seq3||seq4;
  do i=1 to length(seq);
    vname='RT'||substr(seq,i,1)||left(put(i,best.));
    output;
  end;
run;

proc transpose data=aaseq0 out=aaseq_trans(drop=_:);
  id vname;
  var val;
run;
```

We can see in each column header, the dummy variable name PRP1, PRQ2, PRI3, PRT4, PRL5 to PRF99, and RTP1, RTI2, RTS3, RTP4, RTI5 to RTL560 are created.

Table 4

	PRP1	PRQ2	PRI3	PRT4	PRL5	RTP1	RTI2	RTS3	RTP4	RTI5
1										

Next, we need to add attributes into hundreds of these variables. First, we import the initial spec Excel file and generate a specification dataset spec_adrst_dev which includes variable names RTyxxx, PRyxxx. Second, we create global macro variables and use dummy data aaseq0 that we created previously to store the names of all RT and PR variables. See below code example.

```
data _null_;
  set spec_adrst_dev;
  if variable='RTyxxx' then do;
    call symputx('rtyxxx',cats(_n_-1));
    call symputx('rtyxxx_len',cats(length));
    call symputx('rtyxxx_label',label);
    call symputx('rtyxxx_type',type);
  end;
  else if variable='PRyxxx' then do;
    call symputx('pryxxx_len',cats(length));
```

```
    call symputx('pryxxx_label',label);
    call symputx('pryxxx_type',type);
  end;
run;
```

Then, type, length and label of all RT and PR variables can be assigned as elaborated below. With this specification dataset, we can use our standard macro to add attributes into the final dataset.

```
**Create Spec Dataset for RTyxxx and PRyxxx**;
data _aaseq0(keep=variable domain label type length);
  length label $40 type $4;
  set aaseq0(rename=(vname=variable) keep=vname);

  domain='ADRST';
  if substr(variable,1,2)='RT' then do;
    type="&rtyxxx_type";
    length=&rtyxxx_len;
    label="&rtyxxx_label "||compress(variable,,'kd');
  end;
  else if substr(variable,1,2)='PR' then do;
    type="&pryxxx_type";
    length=&pryxxx_len;
    label="&pryxxx_label "||compress(variable,,'kd');
  end;
run;
```

The below is the partial output for _aaseq0 dataset.

Table 5

	LABEL	TYPE	VARIABLE	DOMAIN	LENGTH
1	Protease Amino Acid Position 1	Char	PRP1	ADRST	30
2	Protease Amino Acid Position 2	Char	PRQ2	ADRST	30
3	Protease Amino Acid Position 3	Char	PRI3	ADRST	30
4	Protease Amino Acid Position 4	Char	PRT4	ADRST	30
5	Protease Amino Acid Position 5	Char	PRL5	ADRST	30
6	Protease Amino Acid Position 6	Char	PRW6	ADRST	30
7	Protease Amino Acid Position 7	Char	PRQ7	ADRST	30
8	Protease Amino Acid Position 8	Char	PRR8	ADRST	30
9	Protease Amino Acid Position 9	Char	PRP9	ADRST	30
10	Protease Amino Acid Position 10	Char	PRL10	ADRST	30

4). Phenotypic data

Table 6

Variable Name	Variable Label	Type	Length
zzzEC50	EC50 Values (uM)	Char	10
zzzECRF	EC50 Fold Change to Reference	Char	10
zzzECBL	EC50 Fold Change to Baseline	Char	12
PHENOMET	Phenotypic Assay Name	Char	50
PHENORF	Reference Strain	Char	20
PHENFAIL	Phenotypic Test Failed	Char	8

Phenotype tests assess which drugs can stop HIV from growing in a laboratory setting. They measure a virus's ability to grow in different concentrations of antiretroviral drugs and the ability of drugs to block viral replication in cell culture.

zzzEC50 (EC50 Values (uM)): zzz is a placeholder for the three-character abbreviation of drug used in phenotype assay., i.e., DOR, FTC, TFV, one column (one variable) for each drug.
Half maximal effective concentration (EC50) is a measure of the concentration of a drug. There is one column for each unique value of drug for all records with PFTESTCD='IC50S'.

zzzECRF (EC50 Fold Change to Reference): Fold change is defined as the ratio of test compound EC50 to the reference compound EC50. When a reference control is defined for the run or protocol, the fold change is determined automatically as part of the overall dose-response set of parameters.
There is one column for each unique value of drug for all records with PFTESTCD='IC50FCR'.

CONCLUSION

The HIV resistance dataset is a key dataset to analyze drug resistance. In addition to patient data, it contains various end point data with HIV-1 RNA lab test results and virological failure flag. The genotypic data that indicates the presence of specific genetic mutations can be generated by creating dummy variables and providing positioning information for over six hundred amino acid sequence data for PR and RT endpoints. Lastly, phenotypic data are incorporated to assess which drugs can stop HIV from growing. By designing an effective algorithm and manipulating the data structure modularly, we can generate it efficiently and accurately.

REFERENCE

World Health Organization. “HIV drug resistance”. 17 November 2022. Available at Fact Sheet: HIV Drug Resistance (who.int)

Guidance on Antiviral Product Development — Conducting and Submitting Virology Studies to the Agency: Guidance for Submitting HIV-1 Resistance Data. February 2014, https://www.fda.gov/media/88255/download

Renslow Sherer: “Genotype and Phenotype basics. 26July2008. Available at Genotype and phenotype basics (thebody.com)

ACKNOWLEDGMENTS

Thanks Abhilash Vasu Chimbirithy at Merck for his contributions to this paper.

CONTACT INFORMATION

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

Jenny Zhang
Merck & Co., Inc.
Jenny.zhang@merck.com
267-305-1023

Shunbing Zhao
Merck & Co., Inc.
Shunbing.zhao@merck.com
732-594-3976