PharmaSUG 2018 - Paper DS-22

Implementation of ADaM Basic Data Structure on Genetic Variation Data for Pharmacogenomics Studies

Linghui Zhang, Merck & Co., Inc.

ABSTRACT

Genetic variations have essential impacts on drug exposure and response. Pharmacogenomics (PGx), deciphering the drug-gene relationship, have been widely applied in drug design, discovery, development, and labeling. To facilitate scientific progress in the field of PGx and the use of PGx data in drug development, FDA officially issued a few guidelines to assist pharmaceutical sponsors engaged in evaluating the role of genetic variations on drug response and to encourage voluntary PGx data submission. To that end, the Study Data Tabulation Model Implementation Guide: Pharmacogenomics/Genetics (SDTMIG-PGx) was published in 2015 to provide guidance on the implementation of SDTM PGx domains. However, there is no official guidance on implementation of Analysis Data Module (ADaM) on PGx data. There are even fewer discussions on ADaM implementation of PGx data. Due to the complicated nature of PGx data, it is not only technical but also scientific challenge to fit PGx data in ADaM structure. Nevertheless, this topic gets more and more important and urgent while FDA highlighted the advancing use of biomarkers and PGx as one of the key principles of Prescription Drug User Fee Act for fiscal years 2018-2022 (PDUFA VI).

This paper will illustrate the implementation of ADaM Basic Data Structure (BDS) on typical genetic variation data including single nucleotide polymorphism (SNP) and short tandem repeat (STR). The common difficulties and solutions while programming genetic variation data will be dissected minutely. This paper will also go beyond ADaM implementation – discuss the incorporation of genetic variations as baseline covariates into efficacy and PK analysis, and the adaption of data structure to the purpose of specific analysis.

INTRODUCTION

Pharmacogenomics (PGx), integrating pharmacology (the science of drugs) and genomics (the study of the full genetic complement of an organism), investigates how the inter-individual differences of genomic components affect individual responses to disease and to treatment (Weinshilboum, 2003). PGx offers the promise of achieving personalized treatment by utilizing accurate and reliable genomic information to maximize efficacy and minimize the adverse drug reactions. PGx is now widely applied throughout drug discovery, development and all phases of clinical trials. As the result, over 200 drugs approved by FDA have PGx information in the drug labeling by February 2018 (Table of Pharmacogenomic Biomarkers in Drug Labeling, FDA). Furthermore, FDA published several guidelines to facilitate scientific progress in the field of PGx studies and to facilitate the use of PGx data in drug development (Resources Related to Pharmacogenomics, FDA). Realizing the increasing needs of PGx data submission to regulatory agencies, the Clinical Data Interchange Standard Consortium (CDISC) released the Study Data Tabulation Model (SDTM) Implementation Guide for PGx/Genetics (referred to as PGxIG) to provide guidance on the implementation of SDTM on PGx/genomic biomarker data. Clinical trial programmers are interesting in diving into PGx studies and mapping biomarkers (Cherukuri, 2016) and genetic variations (Zhang, 2017) in SDTM PGx domains.

In contrast to the active practice on implementing SDTM on PGx data, the implementation of Analysis Data Module (ADaM) on PGx data is left behind. PGx is a new study class for clinical trials. Programmers have less knowledge of PGx data and need to know how the PGx data is used in statistical analysis. There is even no official guidance on ADaM implementation of PGx data. Given the complicated nature and lack of standards for PGx data, there are manh challenges to fit PGx data in ADaM structure. Nevertheless, there are many advantages to generate ADaM dataset for PGx data, such as facilitating the clear communications, supporting review tools used by regulatory agencies, reducing the learning curve for new data and new studies, thereby reducing the review duration, etc. Starting with a brief introduction on genetic variations and SDTM PGx domains, this paper will explore the implementation of

ADaM Basic Data Structure (BDS) on genetic variation data and discuss the strategic and practical considerations of creating ADaM PGx data in a clinical trial setting.

GENETIC VARIATION

Genetic variation is the alternate forms of genotypes causing the genotypic and phenotypic differences between individuals in a population, and between populations. Depending on the location and type of genetic variation and its impact on gene function, the consequences of genetic variation vary from neutral or benign phenotype to life-threaten genetic disorders. Genetic variation can influence an individual's response to certain drugs, susceptibility to environmental factors and risk of developing particular diseases. Based on the structure of genetic contents, genetic variation can be divided into different forms according to DNA and RNA characteristics (E15, ICH). The DNA sequence variations can be found across the genome in different structure levels (Figure 1). The concepts related genomics are vast, complicated and not the scope of the paper. The common sequence variations are illustrated graphically in Figure 1 to assist the understanding about the characteristics of genetic variations.

Figure 1. Common DNA sequence variations.

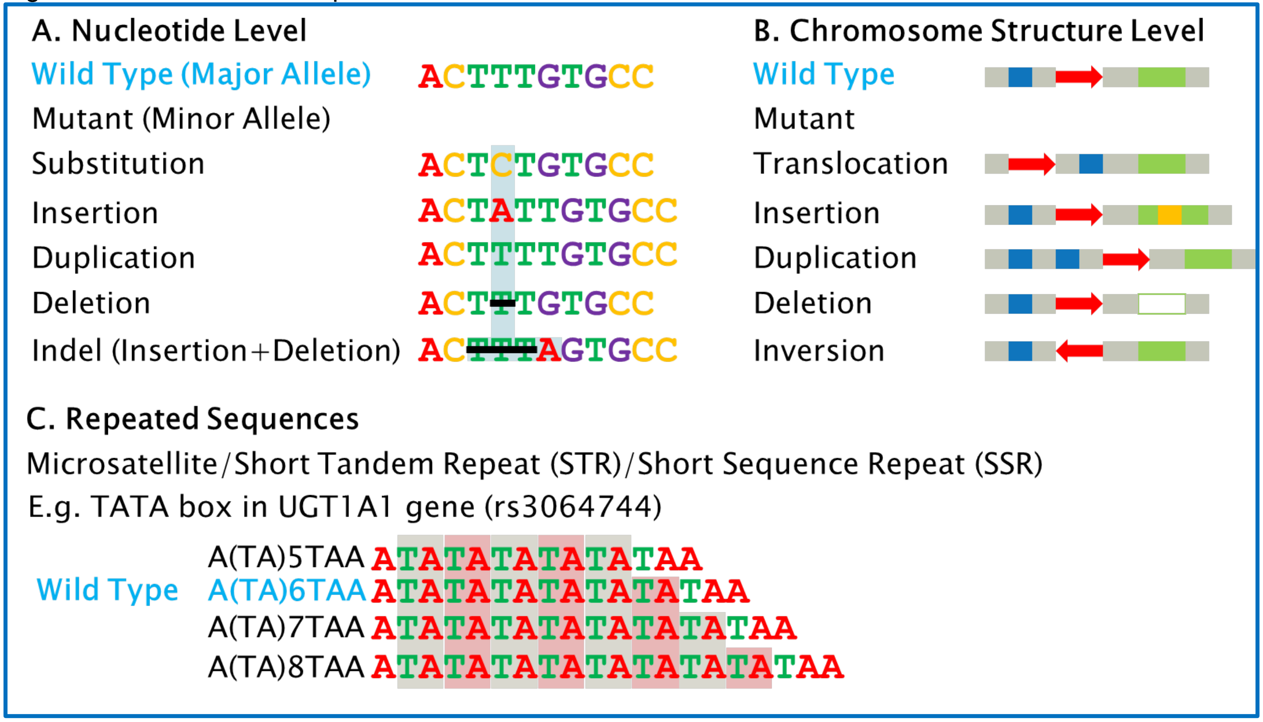

A. Nucleotide Level
Wild Type (Major Allele) ACTTTGTGCC
Mutant (Minor Allele)
Substitution ACTCTGTGCC
Insertion ACTATTGTGCC
Duplication ACTTTTGTGCC
Deletion ACTTTGTGCC
Indel (Insertion+Deletion) ACTTTAGTGCC
B. Chromosome Structure Level
Wild Type
Mutant
Translocation
Insertion
Duplication
Deletion
Inversion
C. Repeated Sequences
Microsatellite/Short Tandem Repeat (STR)/Short Sequence Repeat (SSR)
E.g. TATA box in UGT1A1 gene (rs3064744)
A(TA)5TAA ATATATATATATAA
Wild Type A(TA)6TAA ATATATATATATATAA
A(TA)7TAA ATATATATATATATATAA
A(TA)8TAA ATATATATATATATATATAA

Table 1. SNPs and STR used in this paper.

Gene (Full Name)	rs ID in dbSNP	Reference Sequence	Location within a gene (exon, intro, etc.)	Major Allele	Minor Allele	Nucleotide change(s)	Amino acid change(s)	Effects on transporter activity	Effects on drug Exposure
SLCO1B1 (solute carrier organic anion transporter family 1 member B1)	rs2306283	NM_006446.4	Exon	A	G	c.388A>G	N130D	Increase	Decrease AUC
	rs11045819	NM_006446.4	Exon	C	A	c.463C>A	P155T	Increase	Decrease AUC
	rs4149056	NM_006446.4	Exon	T	C	c.521T>C	V174A	Decrease	Increase AUC
UGT1A1 (UDP-glucuronosyltransferases family 1 member A1)	rs4148323	NC_000002.11	Exon	G	A	c.211G>A	G71R	Decrease	Increase AUC
	rs3064744	NC_000002.11	Promoter	(TA)6	(TA)5 (TA)7 (TA)8	-	-	Unknown	Unknown

The single nucleotide polymorphism (SNP) and short tandem repeat (STR) are two typical DNA sequence variations. SNP is a substitution, deletion, or insertion in a single-base nucleotide of a DNA sequence in at least 1% of the population (Figure 1A). STR, also known as microsatellite, is a class of DNA sequence repeats. In STR, a couple of nucleotides (usually 1 to 10bp) repeat several times and the repeated sequences are directly adjacent to each other. STR is typically in the non-coding regions. The structure and clinical relevance of SNPs and STR used in this paper are summarized in Table 1.

SDTM PGX DOMAINS

To support the increasing need in PGx studies and provide guidance on PGx data submission, the CDISC PGx team released the SDTM PGx IGv1.0 to guide the implementation of SDTM on gene expression data and genetic variation data. It is necessary to have the knowledge about the SDTM PGx domains. The PGxIGv1.0 provided guidance on the implementation of the SDTM for gene expression and genetic variation data for human and viral studies. Seven domains are used to carry data from three categories.

1) Data about biospecimen: BE (Biospecimen Events), BS (Biospecimen Findings), and RELSPEC (Related Specimens).
2) Data about PGx findings: PF (PGx/Genetics Findings) and PG (PGx/Genetics Methods and Supporting Information).
3) Data defining a genetic biomarker or assigning it to a subject: PB (Pharmacogenomics/Genetics Biomarker) and SB (Subject Biomarker).

Based on the FDA Data Standards Catalog, SDTM and ADaM are the study data and analysis data for clinical use, respectively. In this paper, the SDTM PGx Findings (PF) is the source for PGx analysis data, referred to as ADPF. Other data structures not listed in FDA Data Standards Catalog will not be considered as the source for ADPF given regulatory considerations. The implementation of SDTM PGx domains on genetic variation data (Zhang, 2017) can be referred to as an example of PGx domains. The sample records and variables in SDTM.PF used in this paper are shown in table 2.

Table 2. Example records in SDTM.PF used to generate ADPF.

Row	PFSEQ	PFTESTCD	PFTEST	PFGENRI	PFGENTYP	PFREFSEQ	PFCAT	PFSCAT	PFORRES
1	1	NUC	Nucleotide	SLCO1B1	GENE	NM_006446.4	GENETIC VARIATION	GENOTYPE	C/T
2	2	NUC	Nucleotide	SLCO1B1	GENE	NM_006446.4	GENETIC VARIATION	GENOTYPE	C/A
3	3	NUC	Nucleotide	SLCO1B1	GENE	NM_006446.4	GENETIC VARIATION	GENOTYPE	T/T
4	4	NUC	Nucleotide	UGT1A1	GENE	NC_000002.11	GENETIC VARIATION	GENOTYPE	(TA)6/(TA)6
5	5	NUC	Nucleotide	UGT1A1	GENE	NC_000002.11	GENETIC VARIATION	GENOTYPE	G/G

Row	PFORREF	PFGENLOC	PFSTRESC	PFRSNUM	PFSPEC	PFMUTYP	PFMETHOD	PFDTC
1	C	388	c.[388A>G];[=]	rs2306283	DNA	GERMLINE	MICROARRAY	2014-09-25T15:15
2	A	463	c.[463C>A];[463C>A]	rs11045819	DNA	GERMLINE	MICROARRAY	2014-09-25T15:15
3	T	521	c.[=];[=]	rs4149056	DNA	GERMLINE	MICROARRAY	2014-09-25T15:15
4	(TA)6	234668881	g.[234668881_234668882[6]];g.[234668881_234668882[6]]	rs3064744	DNA	GERMLINE	POLYMERASE CHAIN REACTION	2014-10-12T11:25
5	G	234669144	c.[=];[=]	rs4148323	DNA	GERMLINE	MICROARRAY	2014-09-25T15:15

ADAM

ADaM is a data standard and designed to facilitate the clear and unambiguous communication of the content and source of the datasets supporting the statistical analyses. There are four classes of ADaM datasets, but only three standard data structures (Table 3). ADaM OTHER class has no standard structure. The summary information of the three standard structures in Table 3 can be used as a lookup table to determine the class of ADaM for PGx data.

BEFORE IMPLEMENTATION OF ADAM ON GENETIC VARIATION DATA

The implementation of ADaM on PGx data must adhere to the fundamental principles. Analysis-ready and traceability are key considerations for ADaM implementation on PGx data. Before implementing ADaM on genetic variation data, the basic questions are:

1) Of the four ADaM structures, which ADaM structure is used to implement genetic variation data?

2) How to achieve the standard of traceability and analysis-ready?

To address these questions, the features of genetic variation data and statistical model should be considered carefully.

The two types of genetic variations used in this paper, SNP and STR are both germline characteristics describing differences between the DNA sequences. Like other subject-level features, e.g. race, gender, date of birth, SNP and STR are not caused by drug administration and will not be changed over time. Therefore, occurrence and incidence analysis are not applied on germline characteristics and OCCDS is not appropriate to present genetic variations. In terms of statistical models, SNP and STR will be used as subject-level covariates integrated into PK and efficacy analysis to evaluate the role of genetic variation on drug response. The statistical models can be mixed model, generalized linear model, logistic regression, etc. Genotypes can be used as categorical variables in these models. Thus, both ADSL and BDS can present SNP and STR and enable statistical analysis.

Table 3. Three standard data structures of ADaM datasets.

Class	Structure	Statisitical Analysis	Dataset(s)	Typical Variables
ADSL (Subject Level Analysis Data)	One record per subject	Descriptive information	ADSL	Subject-level variables: Demographic Disease status Treatment Baseline observations Population flags Dates of important events
BDS (Basic Data Structure)	One or more records per analysis parameter per observation per subject	ANOVA Linear regression Logistic regression Mixed model	ADEG ADLB ADVS ADPC	Record-level variables: Analysis parameters Analysis values Treatment/dose Timing variables Analysis indicators
OCCDS (Occurance Data Structure)	One record per SDTM collected term	Occurrence analysis Incidence analysis	ADAE ADCM ADMH	Events Start and end of date/time Analysis flags

ADSL OR BDS?

The BDS is favored than ADSL for a few reasons.

First, the vertical structure of BDS is more flexible and powerful to present the derivation of records. Take SNP as an example. In SDTM.PF, the DNA sequence is character variable shown as two-letter combination of A, T, G, C in PFORRES (Table 3). In analysis data, the character PFORRES can be recoded to numeric analysis value (AVAL) to incorporate into statistical models. In one study, different recoding rules can be applied to the same SNP based on genetic effect, statistical models, sample size, etc. For example, additive model assumes genetic effects are the sum of two individual alleles equally. Then the copy number of minor alleles used for statistical analysis can be 0, 1, 2 in analysis data. If genetic effect presents dominance model that the dominant allele masks function of the recessive allele and the dominant trait is caused by either one or two copies of the dominant allele, then AVAL is a binary value: 0 if homozygous for the recessive allele; 1 if at least one copy of the dominant allele. When two recoding rules are applied to one SNP under distinct genetic effects, two sets of records must be generated separately. BDS is well adapted to this scenario.

Second, in terms of analysis-ready, BDS supports the majority of statistical models (Table 2) and table products. The generation of BDS data is a creative process. Implementation of BDS can provide more flexibility and potential for exploratory analysis. Moreover, BDS supports common table layouts used in clinical study report (CSR), e.g. frequency, summary statistics, shift tables, etc.

Third, BDS enables the datapoint traceability that points directly to the specific predecessor record(s). The derivation for AVAL in PGx analysis can be a complex data manipulation path, it's very helpful to include original variables from source data (SDTM.PF) to trace back to the specific data values used as input for an analysis value. Like other ADaM datasets, the common variables assisting traceability are SRCDOM, SRCSEQ, SRCVAR, and --SEQ from SDTM. In addition, specific variables in SDTM.PF can be included in ADPF, such as PFORRES, PFORREF, PFSPEC, PFMETHOD, PFDTC, etc.

At last, adding PGx data in ADSL may cause the delay in finalizing analysis deliverables. In a typical analysis and report procedure, the order of programming execution is usually ADSL first, then BDS and OCCDS, and finally tables. Unlike clinical data that is collected by eCRF and captured by Electric Data Capture (EDC) system, PGx data usually is not documented in EDC because the PGx assays are often conducted by external vendors, not central lab. In addition, PGx form may not even be included in eCRF in exploratory studies. As the result, the format of PGx data is nonstandard and varies from vendors. The nonstandard formats can cause programming difficulties. Moreover, the delivery of PGx data can be delayed for many reasons, then further delays the finalization of ADSL if PGx data is implemented in ADSL dataset. Thus, it is better to present genotype data in a separeted dataset (e.g. ADPF) than in ADSL. Then the consequence of programming difficulties and the delay of raw PGx data will be limited in only a few ADaM datasets used in PGx analysis.

WHY NOT ADAM OTHER CLASS?

Another question is, can ADaM OTHER class be used to present genetic variation or other type of PGx data, since there is no ADaM implementation guidance for PGx data? The general rule of thrumb is that use standard structures if the standard structures can support the analysis (Troxell, 2015). There are many benefits of using standard structures. ADaM standard structures are preferred by regulatory agencies, pharmaceutical sponsors, clinical trial programmers, and many others using standard structures. Standard structures facilitate the clear communications, reduce the learning curve for new data and new studies, support review tools used by regulatory agencies. Therefore, standard structures should be used other than nonstandard structures. ADaM OTHER class is not a standard structure, thereby ADaM OTHER is not recommended unless the standard structures are not capable of supporting the analysis.

IMPLEMENTATION OF ADAM BDS ON GENETIC VARIATION DATA

The BDS structure usually contains several sets of variables enabling statistical analysis and traceability. This paper will focus on analysis parameters and analysis values relevant to PGx observations.

ANALYSIS PARAMETER VARIABLES

Analysis parameter variables (e.g. PARAM) describe the values being analyzed. The definitions of PARAM and PARAMCD must follow the rules as ADaM implementation guide suggested:

Unique, meaningful, and informative

One-to-one map of PARAM and PARAMCD

The length of PARAMCD value is eight characters or less

However, applying these mapping rules can be very challenge for PGx data because PGx observations are more complicated and not standardized compared with other clinical tests. Examples in Table 4 provided sorts of ideas to create PARAM and PARAMCD, but none of them can meet all the rules above.

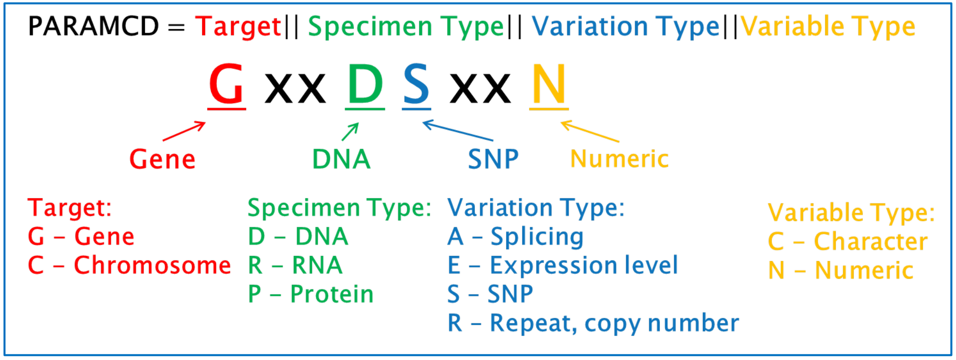
The PGx tests need to be uniquely identified to allow them to be combined for analysis. In addition, analysis parameters must include descriptive and qualifying information relevant to the analysis purpose of the parameter. In order to consistently create PARAM and PARAMCD based on ADaM rules for PGx data, a naming algorithm is proposed to incorporate type of biomarker, genetic specimen, variation, and variable (Figure 2).

Table 4. Examples of PARAMCD failed to follow ADaM rules for PARAMCD.

Derivation Rule	Example	ADaM Rules			
		Meaningful	Unique	Informative	≤ 8 char
RS reference number	RS11045819, RS2306283	✓	✓	✗	✗
Gene name + mutant loaction	SLCO1B1:c.[388A>G]	✓	✓	✓	✗
Chromosome number + chromosome loaction	Chr2:g.234668881-234668882	✓	✓	✓	✗
Study defined SNPs	SNP01, SNP02, …	✗	✓*	✗	✓
Study defined variations	VARNT01, VARNT02, …	✗	✓*	✗	✓

* PARAMCD is uniqe for each indivadule study, but may need to recode for the analysis cross studies

Figure 2. An example algorithm for creating PARAMCD.



PARAMCD = Target|| Specimen Type|| Variation Type||Variable Type

G xx D S xx N

Gene DNA SNP Numeric

Target:
G – Gene
C – Chromosome

Specimen Type:
D – DNA
R – RNA
P – Protein

Variation Type:
A – Splicing
E – Expression level
S – SNP
R – Repeat, copy number

Variable Type:
C – Character
N – Numeric

ANALYSIS VALUE VARIABLES

Under additive effect, genotype for each SNP was categorically defined as 0, 1 or 2 depending on the copies of the minor allele (Table 5).

Table 5. Derivation of AVAL for SNPs under additive genetic model.

PFRSNUM	PFORRES	PFORREF	Minor Allele	Copy Number of Minor Allele	AVAL	PFSTRESC
rs2306283	C/T	C	T	1	1	c.[388A>G];[=]
rs11045819	C/A	A	C	1	1	c.[463C>A];[463C>A]
rs4149056	T/T	T	C	0	0	c.[=];[=]
rs4148323	G/G	G	T	0	0	c.[=];[=]

Similarly, the genotype of TA repeats in the UGT1A1 promoter was categorically defined as 0, 1 or 2 for genotypes TA 6/6, TA 6/7 or TA 7/7 respectively (Table 6). Subjects with genotypes that fall outside of TA 6/6, TA 6/7 or TA 7/7 are rare and will not be included in the analysis of UGT1A1. The three major genotypes contribute over 96% of the study population.

Table 6. Derivation of AVAL for (TA)n repeat in UGT1A1.

PFORRES	PFORREF	Copy Number of (TA)7	AVAL
(TA)6/(TA)6	(TA)6	0	0
(TA)6/(TA)7	(TA)6	1	1
(TA)7/(TA)7	(TA)6	2	2
(TA)5/(TA)6	(TA)6	0	
(TA)5/(TA)7	(TA)6	1	
(TA)6/(TA)8	(TA)6	0	
(TA)7/(TA)8	(TA)6	1	

CASE STUDIES

In the case studies, the roles of genetic variations on efficacy and pharmacokinetics (PK) were evaluated by statistical analysis. Genetic variations are used as subject-level covariates.

CASE 1

In this study, the impacts of genetic variations in SLCO1B1 and UGT1A1 on PK exposure were investigated. The primary PK endpoints are AUC0-∞, steady state AUC0-τ and maximal concentration (Cmax). The genetic endpoints are three SNPs in SLCO1B1, one SNP in UGT1A1 and one SRT (TA repeat) in UGT1A1.

A linear mixed effect model was performed on the individual natural log transformed PK parameters (AUC0-∞, AUC0-τ, and Cmax). This model contains fixed effects of treatment (categorical), genotype (categorical, coded as 0, 1 or 2) for each variant separately, and a random subject effect. Covariates of disease status, weight, gender, and race will be included in the model as appropriate. For example, the model for AUC0-∞ was in the following format:

$\ln(\text{AUC0-}\infty_{ijk}) = \beta_0 + \beta_{1j} * I(\text{treatment}_i = j) + \beta_2 * I(\text{genotype}_i = 1) + \beta_3 * I(\text{genotype}_i = 2) + \beta_4 * \text{covariates}_i$ (e.g., disease status, age, weight, gender, and race) $+ S_i + \varepsilon_{ijk}$

i – i^{th} subject
j – j^{th} treatment
k – k^{th} observation
AUC0-∞_{ijk} – the AUC0-∞ value from subject i, treatment j, observation k
S_i – the random subject effect
e_{ijk} – the residual error

The sample records of ADPF and ADPP (PK parameter analysis dataset) are shown in Table 7 and Table 8.

Table 7. Example layout of ADPF for one subject. The AVAL was derived from SDTM.PF records shown in Table 2.

Row	PARAM	PARAMCD	AVAL	PFRSNUM	PFORRES	PFORREF
1	SLCO1B1 SNP rs2306283 Recode N	G01DS02N	1	rs2306283	C/T	C
2	SLCO1B1 SNP rs11045819 Recode N	G01DS01N	1	rs11045819	C/A	A
3	SLCO1B1 SNP rs4149056 Recode N	G01DS03N	0	rs4149056	T/T	T
4	UGT1A1 Repeat rs3064744 Recode N	G02DR01N	0	rs3064744	TA6/TA6	TA6
5	UGT1A1 SNP rs4148323 Recode N	G02DS01N	0	rs4148323	G/G	G

Table 8. Sample records in ADPP. Analysis genotypes (AGTGzVz) were derived from AVAL in ADPF shown in Table 7.

Parameter Code	*Parameter*	*Analysis Value*	*Aanalysis Genotype of SLCO1B1 rs2306283*	*Analysis Genotype of SLCO1B1 rs11045819*	*Aanalysis Genotype of SLCO1B1 rs4149056*	*Aanalysis Genotype of UGT1A1 rs3064744*	*Aanalysis Genotype of UGT1A1 rs4148323*
PARAMCD	**PARAM**	**AVAL**	**AGTG1V2**	**AGTG1V1**	**AGTG1V3**	**AGTG2V1**	**AGTG2V2**
AUCINF	AUC Infinity (h*ng/mL)	78540	1	1	0	0	0
LNAUCINF	Log(AUC Infinity (h*ng/mL))	11.2714	1	1	0	0	0

CASE 2

In this study, a similar linear mixed effect model was proposed to investigate genetic effect of rs2306283 in SLCO1B1 on PK profile. However, in certain subgroup analysis, there are insufficient sample size for subjects having either 1 or 2 copies of the minor allele (defined here as a sample size less than 5 subjects). Then genotype will instead be defined at the subject level as 0 (homozygous of major allele) or

1 (at least one copy of the minor allele). Therefore, two sets of analysis records were derived from the same set of original observations (Table 9).

Table 9. The different derivations of PARAM for the same SNP. Two sets of analysis parameters were generated for main group (PARAM="G01DS02A") and subgroup (PARAM="G01DS02B") analysis.

USUBJID	PARAM	PARAMCD	AVAL	PFORRES	PFORREF	Minor Allele	Copy Number of Minor Allele
002-001	SLCO1B1 SNP rs2306283 Recode N	G01DS02A	0	C/C	C	T	0
002-002	SLCO1B1 SNP rs2306283 Recode N	G01DS02A	1	C/T	C	T	1
002-003	SLCO1B1 SNP rs2306283 Recode N	G01DS02A	2	T/T	C	T	2
002-001	SLCO1B1 SNP rs2306283 Recode B	G01DS02B	0	C/C	C	T	0
002-002	SLCO1B1 SNP rs2306283 Recode B	G01DS02B	1	C/T	C	T	1
002-003	SLCO1B1 SNP rs2306283 Recode B	G01DS02B	1	T/T	C	T	2

Table 10. Example layout of ADPP. Analysis genotypes were derived from G01DS02A and G01DS02B shown in Table 9.

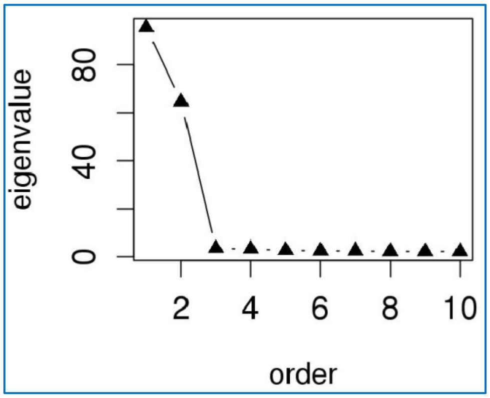
Unique Subject Identifier	*Parameter Code*	*Parameter*	*Analysis Value*	*Analysis Genotype rs2306283 Additive*	*Analysis Genotype rs2306283 Binary*
USUBJID	**PARAMCD**	**PARAM**	**AVAL**	**AGTG1V1A**	**AGTG1V1B**
002-001	LNAUCINF	Log(AUC Infinity (h*ng/mL))	11.2714	0	0
002-002	LNAUCINF	Log(AUC Infinity (h*ng/mL))	14.3997	1	1
002-003	LNAUCINF	Log(AUC Infinity (h*ng/mL))	12.0094	2	1

CASE 3

In an anti-HCV study, the primary efficacy endpoint was sustained viral response at 12 weeks (SVR12). Undetectable HCV for 12 or more weeks after the end of treatment is an SVR12. SVR12 was numerically defined as a binary endpoint of achieved (SVR12=1) or failure (SVR12=0).

For each subject, genome-wide SNPs (~0.85 million) were assayed by microarray. The 0.85M SNPs are considered as a huge set of variables with some redundancies and correlations, e.g. SNPs in linkage disequilibrium regions in chromosomes, SNPs measuring the same construct. Obviously, it is not feasible and over modeled to incorporate all of the 0.85M variables in one statistical model. To reduce the huge set of observed variables and account for most of the variance in the observed variables, the principal component analysis (PCA) was applied to these 0.85M SNPs to generate a small set of artificial variables (called principal component, PC). These PCs describe the overall ancestry related genetic structure of this patient population. The number of components generated in a PCA is equal to the number of observed variables being analyzed, but usually the first few components account for meaningful amounts of variance. In this study, the first three PCs contributed for over 90% of total variance as shown in scree plot (Figure 3) and were used in efficacy analysis.

Figure 3. Scree plot of PC versus its corresponding eigenvalue. The eigenvalues are ordered from largest to smallest. The eigenvalues of the correlation matrix equal the variances of the PCs. Scree plot is used to select the number of components to use based on the size of the eigenvalues. In this plot, the first three PCs account for over 90% of total variances.



A logistic regression model was performed on SVR12 to evaluate the relationship between the probability (p) of achieving SVR12 and genetic variations. The impact of three SNPs in SLCO1B1 and overall ancestry related genetic structure on the probability of achieving SVR12 was investigated. The model includes fixed effects of genotype (categorical coded as 0, 1, 2), the first three PCs obtained from the PCA, and other covariates (treatment, treatment duration, baseline HCV RNA, etc.). The model for the probability (p) of achieving SVR12 was in the following format:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 I(G=1) + \beta_2 I(G=2) + \sum_{k=1}^{3} \gamma_k PC_k + \sum_{k=4}^{p} \gamma_k X_k$$

I – i^{th} subject
G – genotype
PC – principal component
X – covariates of treatment, treatment duration, baseline HCV RNA, etc

The genotypes of three SNPs and first three PCs are included as subject-level variables in the efficacy analysis dataset (ADEFF, Table 11).

Table 11. Example layout of ADEFF. The genetic covariates are three SNPs in SLCO1B1 and the first three PCs from a PCA of 0.85M SNPs.

Unique Subject Identifier	*Parameter Code*	*Analysis Value*	*Analysis Genotype of rs2306283*	*Analysis Genotype of rs11045819*	*Analysis Genotype of rs4149056*	*Principal Component 1*	*Principal Component 2*	*Principal Component 3*
USUBJID	**PARAMCD**	**AVAL**	**AGTG1V2**	**AGTG1V1**	**AGTG1V3**	**PC1**	**PC2**	**PC3**
003-011	SVR12	0	1	0	0	-0.001629	-0.010175	-0.018122
003-012	SVR12	1	0	1	0	0.0148152	-0.010058	0.0009233
003-013	SVR12	1	0	2	1	0.0157934	-0.010744	0.0063899

The SNPs and PCs are genetic covariates in ADEFF for efficacy analysis. However, the inclusion of three PCs in ADEFF caused the traceability problems when PCs are not mapped in SDTM domains and SDTM.PF does not contain all of the 0.85M SNPs. As calculated values, PCs are not direct observations and should not be mapped to SDTM.PF, which is limited to genetic finding records. The PGx IGv1.0 doesn't suggest the implementation of derived values from genetic findings yet. In terms of total data fitness, it is not practical to list 0.85M records for each subject in SDTM.PF. In addition, it's not necessary to report all the SNPs since they are redundant and correlated. Sponsors are allowed to present only the interesting genetic findings according to PGx IGv1.0. Then the traceability between ADaM and SDTM is broken when PC is included in any of SDTM domains and only a few of 0.85M SNPs are reported by SDTM.PF.

In order to build bridge for PC between ADEFF and SDTM datasets, a custom SDTM domain, e.g. genetic parameters (GP, Table 12), can be generated to contain the three PCs, then a separated ADaM dataset (e.g. ADGP) can carry over PCs from SDTM. The dataset relationship between SDTM.PC (Pharmacokinetics Concentration) and SDTM.PP (Pharmacokinetics Parameters) can be a reference for the custom domain for genetic parameters and SDTM.PF. PP records are not direct observations and derived from PK concentrations in SDTM.PC, but can fit SDTM finding class well. As a reference from SDTM.PP, genetic parameters (e.g. three PCs derived from PCA) can be presented by SDTM finding class. Then the traceability between ADaM and SDTM is achieved by generating a customer SDTM domain.

Table 12. Example layout of a custom SDTM domain to present genetic parameters derived from SDTM.PF.

SUBJID	**--TEST**	**--TESTCD**	**--ORRES**
003-011	Principal Component 1	PC1	-0.0016294
003-011	Principal Component 2	PC2	-0.0101749
003-011	Principal Component 3	PC3	-0.0181217

CONCLUSIONS

The powerful ADaM BDS class provides sufficient flexibilities to present genetic variation data, enable datapoint traceability and support statistical analysis.

ACKNOWLEDGEMENTS

The author would like to thank the colleagues, Amy Gillespie, Qi Shen, Jing Su and Xiaohui Wang for their valuable suggestions.

REFERENCES

Weinshilboum, Richard. 2003. “Inheritance and Drug Response”. The New England Journal of Medicine,

348:529-537.

Table of Pharmacogenomic Biomarkers in Drug Labeling, FDA. https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm

FDA Resources Related to Pharmacogenomics. https://www.fda.gov/Drugs/ScienceResearch/ucm572736.htm.

Cherukuri, Kiran. 2016. “Transforming Biomarker Data into an SDTM based Dataset”. PharmaSUG 2016 DS15.

Zhang, Linghui. 2017. “Implementation of SDTM Pharmacogenomics/Genetics Domains on Genetic Variation Data”. PharmaSUG 2017 DS05.

ICH E15: Terminology in Pharmacogenomics.

CDISC. 2015. Study Data Tabulation Model Implementation Guide: Pharmacogenomics/Genetics.

Troxell, John. 2015. ‘What is the “ADAM OTHER” Class of Datasets, and When Should it be Used?’ PharmaSUG 2015 DS16.

CONTACT INFORMATION

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

Linghui Zhang
Merck & Co., Inc.
267-305-6747
linghui.zhang@merck.com