PharmaSUG 2017 - Paper DS05

Implementation of STDM Pharmacogenomics/Genetics Domains on Genetic Variation Data

Linghui Zhang, Merck & Co., Inc., Upper Gwynedd, PA USA

ABSTRACT

Pharmacogenomics (PGx) explores how gene expression and genetic makeup affect individual responses to drugs. It has been recognized since the early 1960s that inherited variation contributes to drug responses. Currently PGx studies are widely used at various stages of drug development and labeling, such as stratifying patients in clinical trials, improving drug safety and optimizing doses. Although the Guidance for Industry: Pharmacogenomics Data Submissions was officially issued by FDA in 2005, due to the complexity and specificity of PGx data, the Clinical Data Interchange Standard Consortium (CDISC) published the clinical data standard for genomics and biomarker data, Study Data Tabulation Model Implementation Guide: Pharmacogenomics/Genetics (SDTMIG-PGx) in May 2015. Since PGx is a still new topic to clinical trial programmers, the type, quality and analysis of PGx data, as well as biospecimen collection and handling techniques for PGx studies, are not broadly discussed in the programming community. Moreover, new domains and variables are developed to capture PGx data in SDTMIG-PGx. Therefore, it is a challenge to programmers to implement SDTM PGx domains.

This paper will provide a high level introduction on the type and application of PGx data and the strategies and practical considerations of creating SDTM PGx domains. By using the single nucleotide polymorphism (SNP), a type of well-known genetic variation as an example, the details of mapping human SNPs to SDTM PGx domains will be illustrated.

INTRODUCTION

Pharmacogenomics (PGx) investigates how the inter-individual differences of genomic components affect patient responses to disease and to treatment (Weinshilboum, 2003). PGx has been widely recognized as the fundamental steps toward personalized medicine (Xie & Frueh, 2005). More and more pharmaceutical agents are now embarking on PGx programs in drug discovery, development and post-marketing surveillance. To provide guidance on PGx data submission, the Clinical Data Interchange Standard Consortium (CDISC) PGx team recently released the Study Data Tabulation Model (SDTM) Implementation Guide for PGx/Genetics (referred to as PGxIG1.0) and provided guidance on the implementation of the SDTM datasets for PGx/genomic biomarker data. This paper briefly introduces the concepts related to PGx, and reviews the structure and variables of SDTM PGx domains in PGxIG1.0. An example on the implementation of SDTM PGx datasets will be provided to illustrate the transformation on human single nucleotide polymorphisms (SNPs, pronounced "snips") in genome-wide association study (GWAS). At the end, this paper will discuss the strategies and practical considerations of creating SDTM PGx domains for PGx data in clinical trial setting.

GENOMIC BIOMARKERS, PHARMACOGENOMICS AND PHARMACOGENETICS

As a part of biomedical studies, pharmacogenomics is not new in science. In the early 1960s, pharmacologists recognized that inherited variation in genes contributes to pharmacokinetics in plasma and urine, e.g. genes encoding the enzymes butyryl-cholinesterase and N-acetyltransferase. Dr. Werner Kalow first proposed the concept and systemic study of pharmacogenetics (Kalow, 1962). Historically, pharmacogenomics, pharmacogenetics and genomic biomarker are often used interchangeably. The term “pharmacogenetics” was first used to describe the study of genetic variation in genes associated with drug metabolism, while the term “pharmacogenomics” usually refers to the study of genome-wide genetic and pharmacologic interactions. Pharmacogenetics was most concerned with drug safety, but pharmacogenomics will deal with drug efficacy as well. The term “biomarker” generally refers to a measurable indicator of some biological state or condition.

To ensure the consistent use of pharmacogenomics terminologies in regulatory documents, International Conference on Harmonisation (ICH) published the harmonized guideline E15 in November 2007, and

then the FDA released the Guidance for Industry on E15 in April 2008. ICH definitions of a genomic biomarker, pharmacogenomics and pharmacogenetics are detailed below:

- A **genomic biomarker** is a measurable **DNA and/or RNA characteristic** that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions.
- **Pharmacogenomics (PGx)** is the study of variations of **DNA and RNA characteristics** as related to drug response.
- **Pharmacogenetics (PGt)** is a subset of PGx, which is the study of variations in **DNA sequence** as related to drug response.

It must be noted that the three concepts refer to not only human variants but also microbial variants and variants from animal samples. In addition, the three concepts include but are not limited to germline or somatic variants. The germline variant affects every cell in an organism and is passed on to offspring, while the somatic variant occurs in a single cell in somatic tissue and is not transmitted to progeny. In terms of genomics and genetics, the variation is the differences of DNA and RNA characteristic within and among populations. The field of genomics is vast and complicated, and is not the topic of this paper. Detailed information on DNA and RNA characteristics can be found in E15. The readers can refer to Appendixes in PGxIG 1.0 for the glossary and abbreviations used in this paper, such as gene, protein, allele, intron and coding sequence.

PGx promises to improve drug safety and efficacy by developing personalized, genetic-based strategies to identify the right drug and dose for each patient (Ventola, 2013). That is, one-size-fit-all drug model has being pushed aside by customized prescriptions. PGx has been widely used throughout drug discovery and development, such as determining drug responders, avoiding adverse drug response, and optimizing dose. Furthermore, PGx information has been added to the drug labeling in many FDA-approved drugs (FDA website, Table of Pharmacogenomic Biomarkers in Drug Labeling). For example, Simvastatin is a statin drug to treat dyslipidemia and to prevent atherosclerosis-related complications such as stroke and heart attacks. The FDA recommends 5-80mg daily Simvastatin dosage. However, genetic variation (rs4149056 c.521T>C, p.V174A) of gene SLCO1B1 modestly increases the risk of myopathy even at lower simvastatin doses (40mg daily) (Niemi, 2010). The patients with this variation are recommended with a lower dose of Simvastatin or use an alternative statin.

SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

Single nucleotide polymorphism (SNP) is the most common type of genetic variation among people. SNP is a substitution, deletion or insertion in a single-base nucleotide of a DNA sequence (Bentley, 2000). A DNA sequence is formed from a chain of nucleotides consisting of four different nitrogenous bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). A SNP occurs if a base is replaced with other base, removed, or added in a DNA strand in at least 1% of the population. For example, the base G may appear in most individuals at a specific base position in the human genome, but the position is occupied by base A in a minority of individuals.

SNPs have been found normally throughout a person's DNA. Most SNPs are genetically neutral or benign and don't affect health or development, but some SNPs can change an individual's response to certain drugs, susceptibility to environmental factors and risk of developing particular diseases. For example, SNPs rs2383206 and rs10757274 are two variants located in introns (non-coding sequence) of gene CDKN2B-AS1 (CDKN2B antisense RNA), and significantly enhance the risk of heart disease. About one in every four Caucasians is thought to be the carrier of the two SNPs, and their risk of coronary heart disease is increased by 30 to 40%. The example of SNP alternating drug response is SNP rs1142345 (c.719A>G, p.Tyr240Cys), a missense mutation (causing nonfunctional protein) in gene TPMT (thiopurine methyltransferase). TPMT protein is an enzyme in the metabolism of thiopurine drugs (e.g. 6-mercaptopurine [6-MP]). This SNP leads to TPMT enzyme deficiency, and consequently decreases inactivation of 6-MP. Individuals who have two abnormal copies (homozygous deficient) are tended to develop anemia, leukopenia and infections due to 6-MP. Thus, dose range of 6-MP should be adjusted based on TPMT genotype.

STUDY DATA TABULATION MODEL IMPLEMENTATION GUIDE: PHARMACOGENOMICS/GENETICS

The PGxIG1.0 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:

Data about Biospecimens

BE – The Biospecimen Events domain captures the details regarding actions taken that affect a specimen or alter its status, such as transportation, freezing, thawing and aliquoting. To track these actions, BE domain also includes the date/time of the action occurred and the site, laboratory and vendor that are accountable for the specimen.

BS – The findings related to specimen handling, the characteristics of biospecimens and extracted samples are collected in the Biospecimen Findings domain. Data may include specimen mass, volume, shipping and storage conditions, and the quantity and quality of extracted samples (e.g. purity and integrity of the DNA or RNA samples).

RELSPEC – the Related Specimens domain documents the hierarchy of specimen relationships.

Data about Genetic Observations

PF – Data about genetic observations is included in Pharmacogenomics/Genetics Findings domain.

PG – Pharmacogenomics/Genetics Methods and Supporting Information domain holds the methods, algorithms and parameters used in the analysis.

Data that define a genetic biomarker or assign it to a subject

PB – The Pharmacogenomics/Genetics Biomarker domain is independent from study subjects. This domain defines the biomarker and contains known associations between observed biomarkers and medical conclusions (e.g., disease diagnosis, drug resistance).

SB – the Subject Biomarker domain connect biomarkers observed in PF domain with the medical conclusions documented in the PB domain for each subject.

BE belongs to the general observation class of Events; BS, PF, and PG are findings. RELSPEC, PB and SB are special-purpose domains and do not conform to any of the CDISC three observational classes of findings, events or interventions. The category of PGx data and class for SDTM PGx domains are summarized in Table 1.

Table 1. Data category and domain class for PGx datasets.

Class / Category	General Observation Class		Special-Purpose
	Event	Finding	
Data about biospecimens	BE	BS	RELSPEC
Data about genetic observations		PF, PG	
Data that define a genetic biomarker or assign it to a subject			PB, SB

PGXIG1.0 IS DESIGNED TO ACCOMMODATE THE FEATURES OF PGX STUDY

Since PGx is a part of biomedical studies, the data collected within PGx study has many characteristics of biological data. Actually, the PGx data is multiple dimensional and highly complex. PGxIG1.0 is structured to accommodate these characteristics.

First, PGx data is heterogeneous. Depending on the objective of the PGx study and computational solutions used in the study, PGx data may contain sequence, images, annotations, values in various formats not limited to plain text, binary, tubular data. In terms of database schemas, flat file database, object-oriented database and relational database are all used to manage PGx data. Consider sequence

data, the very common data of genetic variation, as an example, there're quite a few commonly used formats for DNA and protein sequences, such as FASTA, Multiple Sequence Alignment (MSA), ClustalW, GenBank, EMBL and SWISSPROT/TrEMBL. Given the heterogeneity of PGx data, new variables are specified to comply with the dimensions and forms related to genetic variation and gene expression data. Such as --GENRI and --GENTYP are used to annotate the genetic region of interest, PFGENLOC and PFGENSR indicate the numeric location and the portion of the genetic variation within the sequence. Variable PFRSNUM is assigned to carry rs identifier from dbSNP database or cluster identifier from refSNP database because these two identifiers are most widely used to refer SNPs and being used in commercial SNP arrays. Variable --REFSEQ is given to accession number in RefSeq database to refer genomic DNA, RNA transcripts, and proteins.

Second, PGx data is quite large. PGx studies are fueled by advances in the development of a variety of high-throughput technologies (Fox, et al. 2009; Chan, 2005), such as cDNA microarray, ChIP-chip (chromatin immunoprecipitation ["ChIP"] with DNA microarray ["chip"]), SNP microarray, etc. By applying these high-throughput technologies, usually thousands to millions of genes or biomarkers can be detected in a single assay. For example, Genome-Wide Human SNP Array 6.0 manufactured by Affymetrix features 1.8 million genetic markers, including 0.9 million SNPs and 0.9 million probes for the detection of copy number variation. In a clinical trial enrolled with several hundred subjects, billions of genetic records can be acquired via the SNP Array 6.0 above. As a result of high volume and high variability, original PGx data is highly noisy. It is crucial to apply statistical analysis on original PGx data to get meaningful results. Moreover, only a small portion of the differently expressed genes or significant genetic variations are well known on the function or are identified to relate with response to drug. Those with statistical significance but unclear function might be good candidates for exploratory studies in the future. However, they can't be associated to clinical conclusions if the function is less known. Thus, PGxIG1.0 allows sponsor to only report the "interesting" findings, e.g. the significant biomarkers associated with drug response. Unlike the regulatory submission of clinical data which almost entire data should be submitted, the report of PGx data is result-orientated in genome-wide studies. This policy is due to the unexplained variation of PGx data. Only the well-identified biomarkers with statistical significance are considered to report. For the studies that genomic biomarker is used to stratify patients, such as optimizing dose based on genotypes to maximize efficacy, avoiding adverse drug response in patients with specific genotypes, genetic findings on these genomic biomarkers must be reported.

Third, quality and accuracy are less controllable with many forms of PGx data. The application of PGx study in clinical trial setting is being challenged by data inconsistencies. Inconsistencies arise as existence of various experimental platforms, sample preparation protocols and data analysis methods. Even different experimental runs with different technicians in the same lab may not be completely consistent with each other. To deal with these inconsistencies, highly reliable protocols, quantitative and unbiased measurements need to be developed. Therefore, PGxIG1.0 emphasizes to report the information on biospecimen handling, sample preparations and the analysis methods. Seven domains are created to hold data about genetic observations, biospecimens procedures and analysis methods to generate genetic findings, and references on genetic biomarkers. The BE and BS domains hold the actions and findings about biospecimen handling and sample preparations. Because the results of PGx tests can be very instrument-dependent and protocol-dependent, and sensitive to sample handling, it becomes particularly important to document such procedures, devices and conditions related to PGx sample preparations. PGx methodologies and supporting information are recorded in the PG domain to provide the information on algorithms, software and parameters which result in the PGx findings.

PGX MAPPING PROCEDURES

In a study involving a statin drug, three SNPs in gene SLCO1B1 (solute carrier organic anion transporter family member 1B1) associated with drug absorption, disposition, metabolism and excretion (ADME) process are investigated by SNP microarray.

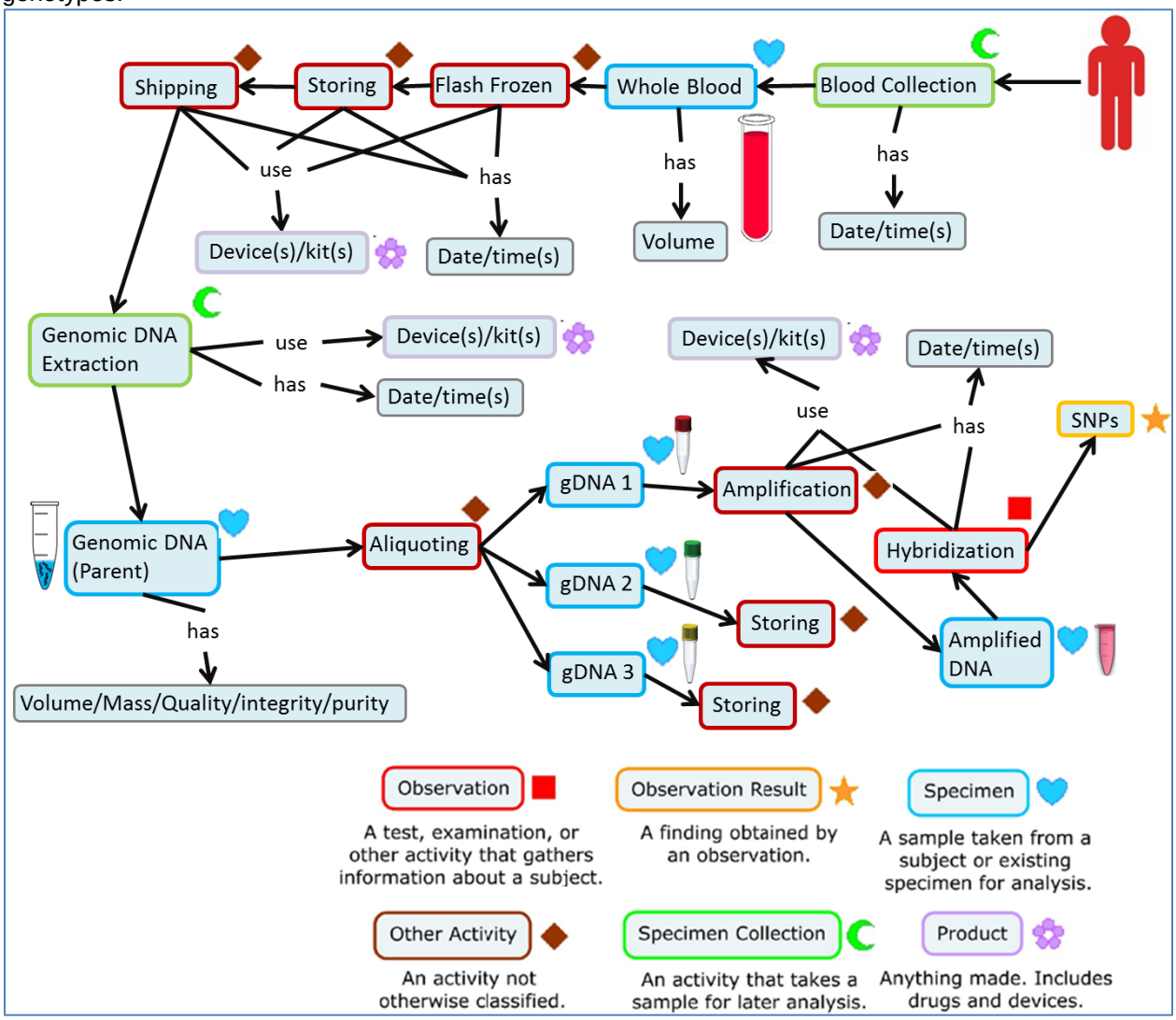
The SLCO1B1 protein transports numerous compounds including hormones, toxins, and drugs from the blood into the liver for removal from the body. For example, SLCO1B1 transports the statin drugs, which are used to lower cholesterol level in blood to prevent heart attack and stroke. Many studies indicate SNP rs4149056 (c.521T>C, p.V174A) in SLCO1B1 is associated with response to statins. The rs4149056 C allele generates an amino acid change from Valine (V) to Alanine (A) at residue 174 which has decreased

transporter function. Therefore, the patients with C allele tend to build up higher drug concentration in plasma than the person having T allele and have modest increased risk in myopathy even at lower statin doses. It has been suggested by other studies that rs2306383 (c.388A>G, p.N130D) in SLCO1B1 has increased uptake activity. In the following example of clinical trial on a statin, rs4149056, rs2306383 and rs11045819 in SLCO1B1 are detected by microarray to evaluate the associations. The PGx information of the three SNPs is summarized in table 2.

Table 2. SNPs interested in the clinical trial on a statin drug.

Gene	rs ID in dbSNP	Nucleotide change(s)	Amino acid change(s)	Effects on transporter activity	Effects on drug concentration in plasma
SLCO1B1	rs2306283	c.388A>G	N130D	Increase	Decrease
	rs11045819	c.463C>A	P155T	Increase	Decrease
	rs4149056	c.521T>C	V174A	Decrease	Increase

Figure 1. Sample handling and preparation process from collection of whole blood to identification of genotypes.



To extract genomic DNA for SNP microarray assay, whole blood was collected at the screening visit, flash frozen immediately to avoid DNA degradation then stored at -80˚C before shipping to Q lab. The genomic DNA was extracted from whole blood at Q lab. The purity and integrity of genomic DNA were detected to ensure the high quality of samples for subsequent tests. Genomic DNA was then aliquoted for three sub-samples. One aliquot was amplified then conducted with microarray assay. Two other aliquots were frozen immediately and stored in -80˚C freezer for further analysis. The journey from collection of blood sample to identification of genotype was illustrated in Figure 1 by a concept map with classification key used in PGxIG1.0. From whole blood to genomic DNA aliquots, totally six samples were collected, extracted and aliquoted and four levels of specimen were generated. The specimen linkage is shown in Figure 2 and RELSPEC is shown in Table 3.

Figure 2. Specimen relationship.

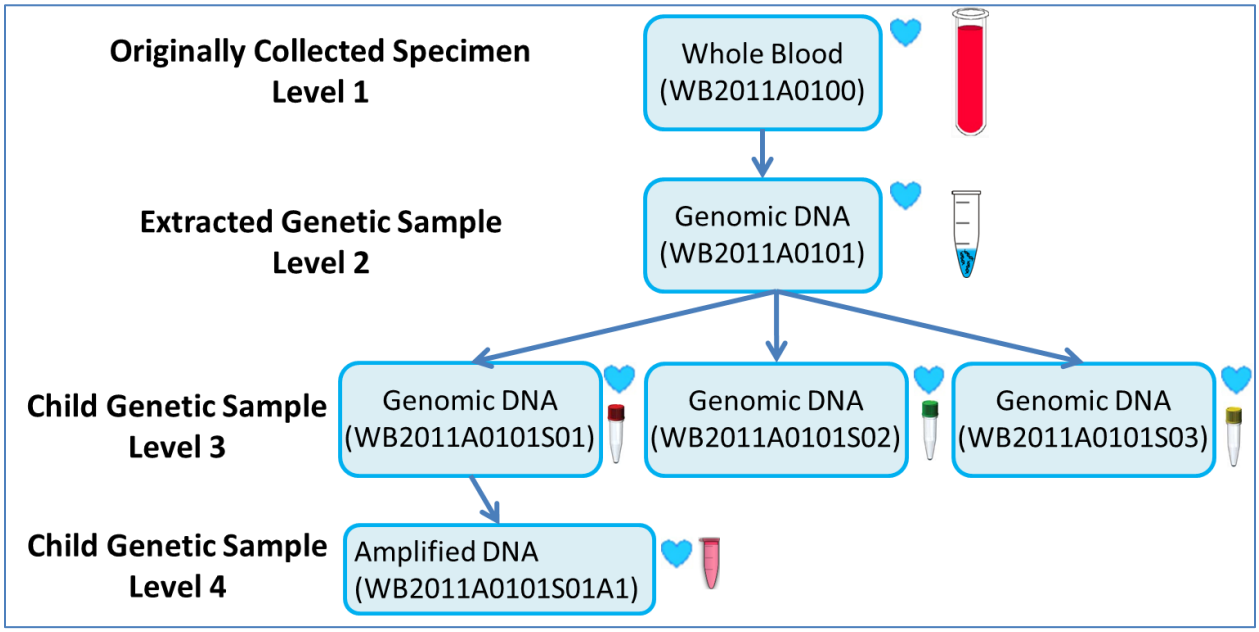


Table 3. Example of RELSPEC domain.

STUDYID	USUBJID	REFID	SPEC	PARENT	LEVEL
ABC-1234	ABC-1234-100001	WB2011A0100	BLOOD		1
ABC-1234	ABC-1234-100001	WB2011A0101	DNA	WB2011A0100	2
ABC-1234	ABC-1234-100001	WB2011A0101S01	DNA	WB2011A0101	3
ABC-1234	ABC-1234-100001	WB2011A0101S02	DNA	WB2011A0101	3
ABC-1234	ABC-1234-100001	WB2011A0101S03	DNA	WB2011A0101	3
ABC-1234	ABC-1234-100001	WB2011A0101S01A1	DNA	WB2011A0101S01	4

BE DOMAIN

BE domain documents the actions taken to the biospecimen (Table 4). Three attributors BETERM, BEDECOD and BECAT are used to carry the topic, description and category of these actions. The traceability information about biospecimen and extracted samples are also collected. Such as date/time (BESTDTC, BEENDTC), accountable party (BEPARTY, BEPRTYID), devices. Unlike other Events domains, BEDTC does not hold the date/time of data collection. Instead, it holds the date/time of specimen collection, in alignment with the use of --DTC for specimen-related findings. BEDTC values for extracted or otherwise derived specimens are copied from that of the parent specimen.

BS DOMAIN

The observations/findings regarding the characteristics of biospecimens and extracted samples such as mass, volume, quality of extracted sample, specimen condition and the integrity of the genomic DNA samples are stored in BS domain (Table 5).

PF DOMAIN

The genetic findings and test methods to identify the finding are stored in PF domain. This PF data in Table 6 shows three SNPs in gene SLCO1B1 for one subject. Each variation is presented as one record per finding per observation per biospecimen per subject in the findings class dataset.

PFTEST and PFTESTCD indicate the variants were identified by nucleotide test, which is one of methods to determine genetic variations. Therefore, “GENETIC VARIATION” is assigned to the category for PGx test (PFCAT) and “GENOTYPE” is assigned to subcategory (PFSCAT). The genetic region of interest (PFGENRI) is gene SLCO1B1 and the type of PFGENRI (PFGENTYP) is GENE apparently. Because the variation is inheritable, the mutation type (PFMUTYP) is germline. Otherwise, it is somatic and this is very common in the finding from tumor tissues. There are quite many methods to detect genetic variations, such as PCR, microarray, sequencing, restriction fragment length polymorphism, denature high performance liquid chromatography. In this study, microarray is used to detect genome-wide SNPs. Thus, the PFMETHOD is set to SNP microarray.

PG DOMAIN

The PGx methods and support information are collected in the PG domain (Table 7). The PG domain is modeled as one record per observation per biospecimen per run per subject. In this example, PG and PF are 1:1 matched. Unlike the method recorded in PF domain, the methods in PG domain are applied to the entire test, not an individual result.

Table 7. Example of records and variables in PG domain.

ROW	SPDEVID	PGREFID	PGTESTCD	PGTEST	PGGENRI	PGGENTYP	PGCAT	PGSCAT
1	LB034MWS02	WB2011A0101S01A1	NUC	Nucleotide	SLCO1B1	DNA	GENETIC VARIATION	GENOTYPE
2	LB034MWS02	WB2011A0101S01A1	NUC	Nucleotide	SLCO1B1	DNA	GENETIC VARIATION	GENOTYPE
3	LB034MWS02	WB2011A0101S01A1	NUC	Nucleotide	SLCO1B1	DNA	GENETIC VARIATION	GENOTYPE

ROW	PGORRES	PGSTRESC	PGMETHOD	PGXFN	PGNAM	PGRUNID	PGDTC
1	GA	c.[388A>G];[=]	MICROARRAY	AFFY11A202.TXT	Q LAB	AFFY11A202	2010-04-01T11:50
2	AA	c.[463C>A];[463C>A]	MICROARRAY	AFFY11A202.TXT	Q LAB	AFFY11A202	2010-04-01T11:50
3	TT	c.[=];[=]	MICROARRAY	AFFY11A202.TXT	Q LAB	AFFY11A202	2010-04-01T11:50

PB DOMAIN

PB data relates a set of genetic variations to an inference about that set of genetic variations (i.e., a medical statement) and is structured with one record per genetic biomarker.

The medical statement (PBSTMT) describes the medical conclusion of the genetic variation for use of a drug, (in PBDRUG) or the diagnosis of a medical condition (in PBDIAG). PBSTMT can be inferred from the presence of a single genetic variant or all genetic variations in a set must be present for an inference to be drawn. PBMRKR identifies the individual variant. PBMRKRID is used to group genetic variation records which belong to a set and which form the basis for medical statement inference.

Table 8. Example of records and variables in PB domain.

ROW	STUDYID	PBSEQ	PBMRKRID	PBMRKR	PBGENRI	PBGENTYP	PBDRUG	PBSTMT
1	ABC-1234	1	C463A	c.463C>A	SLCO1B1	DNA	XStatin	Reduced Transporter Activity
2	ABC-1234	2	A388G	c.388A>G	SLCO1B1	DNA	XStatin	Reduced Transporter Activity
3	ABC-1234	3	T521C	c.521T>C	SLCO1B1	DNA	XStatin	Increased Transporter Activity

Table 4. Example of records and variables in BS domain.
Rows 1-3: The whole blood sample was collected; flash frozen and stored while frozen.
Row 4: The biospecimen was transported to Q lab.
Row 5: The whole blood sample was stored at -80˚C until the extraction.
Rows 6-7: Sample was thaw and submitted to the extraction procedure to get the genomic DNA.
Rows 8-10: The genomic DNA was aliquoted for three sub-samples.
Rows 11-12: The freezing of the second and third DNA aliquots.
Rows 13-14: Amplify the first DNA aliquot, then microarray hybridization is conducted to amplified DNA.

ROW	SPDEVID	BESEQ	BEREFID	BETERM	BEDECOD	BECAT	BEPARTY	BEPRTYID	BEDTC	BESTDTC	BEENDTC
1		8	WB2011A0100	Collected	COLLECTED	COLLECTION	SITE	ST036	2010-04-01T11:50	2010-04-01T11:50	
2		9	WB2011A0100	Flash Frozen	FLASH	FLASH FROZEN	SITE	ST036	2010-04-01T11:50	2010-04-01T11:50	2010-04-01T11:55
3	ST036A101-01	10	WB2011A0100	Stored in Freezer	STORED	STORING	SITE	ST036	2010-04-01T11:50	2010-04-01T11:55	2010-04-02T09:50
4	C0029	11	WB2011A0100	Shipped	SHIPPED	TRANSPORT	Q LAB	LB034	2010-04-01T11:50	2010-04-02T09:50	2010-04-03T08:50
5	LB034FZR15	12	WB2011A0100	Stored in Freezer	STORED	STORING	Q LAB	LB034	2010-04-01T11:50	2010-04-03T08:50	2010-04-04T09:50
6		13	WB2011A0100	Thaw	THAW	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T09:50	2010-04-04T09:55
7	QIAamp51106	14	WB2011A0101	Extracted	EXTRACTED	EXTRACTION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T09:55	2010-04-04T11:20
8		15	WB2011A0101S01	Aliquoted	ALIQUOTED	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	
9		16	WB2011A0101S02	Aliquoted	ALIQUOTED	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	
10		17	WB2011A0101S03	Aliquoted	ALIQUOTED	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	
11	LB034FZR04	18	WB2011A0101S02	Stored in Freezer	STORED	STORING	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	
12	LB034FZR04	19	WB2011A0101S03	Stored in Freezer	STORED	STORING	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	
13	AFFYKIT90758	20	WB2011A0101S01A1	Amplified	AMPLIFIED	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T11:20	2010-04-04T13:11
14	LB034MWS02	21	WB2011A0101S01A1	Hybridized	HYBRIDIZED	PREPARATION	Q LAB	LB034	2010-04-01T11:50	2010-04-04T13:20	2010-04-05T07:11

Table 5. Example of records and variables in BS domain.

Rows 1-3: The blood sample collection and measurement of the volume, then sample was flash frozen on dry ice and stored at -80˚C before sample was shipped to laboratory.
Rows 4-5: The volume and concentration of the genomic DNA extracted from whole blood sample.
Rows 6-8: Shows the purity and integrity of DNA. The A260/A280 and A260/A230 ratios are used to assess the purity of DNA. A260/280 and A260/230 values greater than 1.8 are typically suitable for analysis. Lower ratios may indicate contamination with protein or reagents used during the extraction process. DNA Integrity Number (DIN) is quality measurement calculated by a special algorithm and used to determine the integrity of DNA. The lower value of DIN indicates degradation of genomic DNA during the sample processing and genomic DNA with low DIN is not appropriate for microarray assay.
Row 9: Show the volume of the first DNA aliquot
Row10: Show the concentration of the amplified DNA from first DNA aliquot.
Rows 11-13: Show quality measurements in microarray hybridization.
Rows 14-17: Show the volume and storage of the 2nd and 3rd aliquots.

ROW	BSREFID	BSTESTCD	BSTEST	BSCAT	BSSPEC	BSORRES	BSORRESU	BSMETHOD	BSDTC	BSNAM
1	WB2011A0100	VOLUME	Volume	SPECIMEN MEASUREMENT	BLOOD	1	mL	OBSERVATION	2010-04-01T11:50	SITE
2	WB2011A0100	FFRZTMP	Flash Frozen Temperature	SPECIMEN HANDLING	BLOOD	-80	C	OBSERVATION	2010-04-01T11:50	SITE
3	WB2011A0100	FFRZMAT	Flash Frozen Material	SPECIMEN HANDLING	BLOOD	DRY ICE			2010-04-01T11:50	SITE
4	WB2011A0101	VOLUME	Volume	SPECIMEN MEASUREMENT	DNA	150	ul	OBSERVATION	2010-04-01T11:50	Q LAB
5	WB2011A0101	CONC	Concentration	SPECIMEN MEASUREMENT	DNA	56.9	ng/ul	SPECTROPHOTOMETRY	2010-04-01T11:50	Q LAB
6	WB2011A0101	A260A230	A260/A230	QUALITY CONTROL	DNA	1.89	ng/ul	SPECTROPHOTOMETRY	2010-04-01T11:50	Q LAB
7	WB2011A0101	A260A280	A260/A280	QUALITY CONTROL	DNA	1.98	ng/ul	SPECTROPHOTOMETRY	2010-04-01T11:50	Q LAB
8	WB2011A0101	DIN	DNA Integrity Number	QUALITY CONTROL	DNA	9.6		ELECTROPHORESIS	2010-04-01T11:50	Q LAB
9	WB2011A0101S01	VOLUME	Volume	SPECIMEN MEASUREMENT	DNA	50	ul	OBSERVATION	2010-04-01T11:50	Q LAB
10	WB2011A0101S01A1	CONC	Concentration	SPECIMEN MEASUREMENT	DNA	102.6	ng/ul	SPECTROPHOTOMETRY	2010-04-01T11:50	Q LAB
11	WB2011A0101S01A1	DISHQC	Dish QC	SPECIMEN MEASUREMENT	DNA	0.95		CALCULATION	2010-04-01T11:50	Q LAB
12	WB2011A0101S01A1	STEP1CR	Step 1 callrate	SPECIMEN MEASUREMENT	DNA	98.67	%	CALCULATION	2010-04-01T11:50	Q LAB
13	WB2011A0101S01A1	STEP2CR	Step 2 callrate	SPECIMEN MEASUREMENT	DNA	99.36	%	CALCULATION	2010-04-01T11:50	Q LAB
14	WB2011A0101S02	VOLUME	Volume	SPECIMEN MEASUREMENT	DNA	50	ul	OBSERVATION	2010-04-01T11:50	Q LAB
15	WB2011A0101S02	FFRZTMP	Flash Frozen Temperature	SPECIMEN HANDLING	DNA	-80	C	OBSERVATION	2010-04-01T11:50	Q LAB
16	WB2011A0101S03	VOLUME	Volume	SPECIMEN MEASUREMENT	DNA	50	ul	OBSERVATION	2010-04-01T11:50	Q LAB
17	WB2011A0101S03	FFRZTMP	Flash Frozen Temperature	SPECIMEN HANDLING	DNA	-80	C	OBSERVATION	2010-04-01T11:50	Q LAB

Table 6. Example of records and variables in PF domain. The variations found in SLCO1B1 gene from one subject are shown.

ROW	PFREFID	SPDEVID	PFTESTCD	PFTEST	PFGENRI	PFGENTYP	PFREFSEQ	PFCAT	PFSCAT	PFORRES	PFORREF
1	WB2011A0101S01A1	LB034MWS02	NUC	Nucleotide	SLCO1B1	DNA	NM_006446.4	GENETIC VARIATION	GENOTYPE	GA	AA
2	WB2011A0101S01A1	LB034MWS02	NUC	Nucleotide	SLCO1B1	DNA	NM_006446.4	GENETIC VARIATION	GENOTYPE	AA	CC
3	WB2011A0101S01A1	LB034MWS02	NUC	Nucleotide	SLCO1B1	DNA	NM_006446.4	GENETIC VARIATION	GENOTYPE	TT	TT

ROW	PFGENLOC	PFSTRESC	PFRSNUM	PFXFN	PFNAM	PFSPEC	PFMUTYP	PFMETHOD	PFRUNID	PFDTC
1	388	c.[388A>G];[=]	rs2306283	AFFY11A202.TXT	Q LAB	DNA	GERMLINE	MICROARRAY	AFFY11A202	2010-04-01T11:50
2	463	c.[463C>A];[463C>A]	rs11045819	AFFY11A202.TXT	Q LAB	DNA	GERMLINE	MICROARRAY	AFFY11A202	2010-04-01T11:50
3	521	c.[=];[=]	rs4149056	AFFY11A202.TXT	Q LAB	DNA	GERMLINE	MICROARRAY	AFFY11A202	2010-04-01T11:50

SB DOMAIN

The Subject Biomarker (SB) domain is a special purpose domain that associates the subject with the medical conclusions contained in the PGx Biological State (PB) domain based on the genetic mutation observations reported in the Pharmacogenomics Findings (PF) domain.

The Subject Biomarker (SB) domain contains the statement about a subject's clinical state, or about the response of a subject's pathogen to a treatment and connects the genetic findings in the PF domain and clinical statements defined within the PB domain. Thus, a statement is made based on one or more subject findings. The SB data can originate from a case report form (when only the biomarker is collected) or can be derived at a lab (when a more complete set of data are generated).

SBMRKRID should match a value of PBMRKRID in the PB domain. The PBMRKR variable is used to identify the individual variations that belong to the group identified by PBMRKRID in the PB domain. These then get linked via the same values for SBMRKRID and PBMRKRID.

SBSTMT is the statement that is inferred from genetic variation data for the subject. The clinical statement is about either a drug (PBDRUG) or a medical condition (PBDIAG) as designated in the PB domain.

Table 9. Example of records and variables in SB domain.

ROW	SBREFID	SBMRKRID	SBGENRI	SBGENTYP	SBNAM	SBDTC
1	WB2011A0101S01A1	C463A	SLCO1B1	DNA	Q LAB	2010-04-01T11:50
2	WB2011A0101S01A1	A388G	SLCO1B1	DNA	Q LAB	2010-04-01T11:50

DISCUSSIONS

THE STRATEGY FOR IMPLEMENTING SDTM PGX DATASETS

Unlike clinical data, there's no commonly used database management system to deal with PGx data in genome-wide study. Depending on the type of PGx data and software solutions, the PGx data can be structured in various formats and manipulated under different computational environments. Therefore, PGx data is not collected in Electronic Data Capture (EDC) system, e.g InForm and RAVE, which are widely used to manage clinical trial data. As a result, the downstream software solutions and SAS macros transforming EDC extracts to CDISC datasets can't be employed to create SDTM PGx datasets. This brings up many questions:

What data should be used to create SDTM PGx datasets?

How to track the transformations of PGx data?

How to create SDTM PGx datasets?

When should the generation of SDTM PGx datasets occur during the clinical trial process?

This paper can't answer these questions yet, but presents a few ideas about the strategy for implementing SDTM PGx datasets. Based the numbers of biomarkers investigated in a clinical trial, two typical scenarios should be considered, genome-wide study versus the study focusing on a few genomic biomarkers.

In genome-wide study, the large size of PGx data can be acquired by high-throughput technologies and multiple analysis steps are necessary to get meaningful PGx findings. Then further analysis on joint data from both PGx and clinical study is used to understand how genetic variations explain clinical outcomes. At last, the biomarkers associated with the response to drug can be determined after comprehensive analysis on entire study data. In this case, retrospect strategy might be appropriate to implement SDTM PGx datasets because the clinical significant PGx markers can't be identified until all analyses are completed. In retrospect strategy, the generation of PGx domains is result-orientated. Therefore, tremendous programming efforts will be expected to create SDTM-compliant PGx datasets.

For the study investigating only a few biomarkers by classical small-scale approaches, such as PCR, sequencing, ELISA, Some of PGx findings might be collected in EDC. For example, genotypes and

mutants of virus were once mapped to viral resistance (VR) domain based on Virology Therapeutic Area Data Standards User Guide (VR-UG, published by CDISC in 2012). Now the findings about virus genotype can be considered to transfer to PF domain according to PGxIG 1.0. Two new variables, PFNSPCES and PFNSTRN (Non-Host Species and Strain) are added to indicate the species and strain of a microorganism. Then the genetic variations from microorganism and host can be held together in PF domain. In this situation, EDC extract of PGx data can be mapped to SDTM format by directly utilizing well-developed software solutions and SAS macros or adapting these tools to PGx setting. In the case of PGx data is not recorded in EDC but outsourced, the creation of SDTM data may be held until the PGx findings are available or follow the retrospect strategy which is used for genome-wide study.

In general, the strategy for implementing SDTM PGx data is based on the objective of PGx study. Once PGxIG is adopted within an organization, efficiencies will be fueled if PGx data transformation can be streamlined.

CAN PGX DATA FIT INTO EDC?

The implementation of CDISC datasets for clinical data has received significant attention in most recent years. Many tools and processes are developed to support the need. Since the implementation of CDISC datasets on clinical data is well streamlined from EDC extract in pharmaceutical industry, there're many advantages if PGx data is collected in EDC. The implementation can be planned in advance and then the creation of SDTM can be conducted as soon as initial PGx findings are available. Data issues can be caught and solved before database lock. The transformation of PGx data can then be well tracked. Challenges and opportunities arise when evaluating whether PGx data can be moved to EDC. While this paper can't answer the feasibility of this question, as the use of these SDTM domains mature, more discussion will occur along with new options to help simplify and streamline the creation of SDTM data.

ACKNOWLEDGEMENTS

The author would like to thank the colleagues in Merck, Amy Gillespie, Huei-Ling Chen and Ellen Asam for their valuable suggestions.

REFERENCES

Weinshilboum, Richard. 2003. "Inheritance and Drug Response". The New England Journal of Medicine, 348:529-537.

Xie, Hong-Guang and Frueh, Felix. 2005. "Pharmacogenomics steps towards personalized medicine" Future Medicine. 2:325-337.

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

Available at: https://www.cdisc.org/

Kalow, Werner. 1962. Pharmacogenetics: Heredity and the Response to Drugs. Philadelphia, PA: W.B. Saunders Co.

ICH E15: Terminology in Pharmacogenomics.

Available at
http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E15/Step4/E15_Guideline.pdf.

Ventola, C. Lee. 2013. "Role of Pharmacogenomic Biomarkers In Predicting and Improving Drug Response: part 1: the clinical significance of pharmacogenetic variants." Pharmacy and Therapeutics. 38:545.

Table of Pharmacogenomic Biomarkers in Drug Labeling.

Available at
https://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

Niemi, M. 2010. “Transporter Pharmacogenetics and Statin Toxicity” Clinical pharmacology & Therapeutics. 87:130-133.

Bentley, David R. 2000. “The Human Genome Project--an overview.” Medicinal Research Reviews. 20:189-96.

Fox, Samuel Fox; Filichkin, Sergei; Mockler, Todd. 2009. “Applications of ultra-high-throughput sequencing” Plant Systems Biology. 553:79-108.

Chan, Eugene. 2005. “Advances in sequencing technology” Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 573:13-40.

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