PharmaSUG 2018 - Paper DS18

Complexity in collection of PGx data and challenges in mapping to SDTM

Rama Kudaravalli, Syneos Health, NJ, USA.

ABSTRACT

Pharmacogenomics is the study of how genes affect a person's response to drugs. It is the combination of pharmacology and genomics to develop effective, safe medications and doses that will be tailored to a person's genetic makeup. This new field is still in its infancy and its use is currently quite limited, but new approaches are emerging in clinical trials. The pharmacogenomics research is entering the era of "Big data"- a term that refers to the explosion of available information. This vast amount of data brings both opportunities and new challenges in data conversion, standardization and analysis. It is a complex process to convert and standardize the pharmacogenomics data collected from clinical laboratory, as it is difficult to translate the results into actionable prescribing decisions for the affected drugs.

This paper will introduce the challenges faced in collection of pharmacogenomics data, curation and mapping it to SDTM domains. Examples for gene mutation and gene expression studies are illustrated with SDTM mappings.

INTRODUCTION

The ultimate goal of pharmacogenomics is to improve health care based on individual genomic profiles (Evans & Relling, 1999). Together with other factors that may affect drug response - such as diet, age, diseases, lifestyle, environment and state of health - pharmacogenomics has the potential to facilitate the creation of individualized treatment plans for patients and lead to the overarching goal of personalized medicine. The term pharmacogenomics is often used interchangeably with pharmacogenetics. Although both terms relate to drug response based on genetic influences, pharmacogenetics focuses on single drug-gene interactions, while pharmacogenomics encompasses a more genome-wide association approach, incorporating genomics and epigenetics while dealing with the effects of multiple genes on drug response.

The ultimate clinical goals of PGx are to use genomics to guide therapy, that is, to avoid Adverse drug response events (ADEs), maximize drug efficacy and prescribe the right dose to patients, all of which, if achieved will reduce the burden for both patients and the health care system.

This paper will discuss the challenges in pharmacogenomics data collection, curation and mapping to SDTM domains, by focusing on genetic variation and gene expression.

PGX DATA COLLECTION AND CURATION

Precision medicine holds the key to better health. In clinical trials, PGx is faced with big hopes and high expectations by everybody involved: patients who demand effective treatment free of adverse effects; physicians in need of guidance for selecting the most appropriate drug and right dose for the patient; health care providers who have to find ways to improve care by reducing cost; regulatory agencies who need proof of concept to issue guidelines.

The key challenges faced in PGx data are classified into: size, structure, security, standardization, storage and skilled personnel.

Size: Recent technologies in Next Generation Sequencing have resulted in the production of voluminous data at cheaper price and faster rate. Over the coming years, the National Cancer Institute will sequence

a million genomes to understand the biological pathways and the genomic variation. Given that the whole genome of a tumor and a matching normal tissue sample consumes 1 TB (terabyte) of data; one million genomes will require 1 million TB, equivalent to 1000 PB (petabyte) or 1 EB (Exabyte) (Grossman RL, 2012). Thus, a separate database is needed to collect and maintain large amounts of data. Knowledge bases are also built by using many structured and unstructured databases to store and archive the genomic data.

Structure: PGx data is heterogeneous, where data is collected in structured, semi-structured and unstructured databases. It is often, fragmented, dispersed and rarely standardized. It is hard to integrate and analyze the diversified data that grows explosively. Genomic data contains images, sequences, annotations, plain text and values in different formats. Unstructured data cannot be easily queried, analyzed or standardized.

Security: As the PGx data comes from different sources there could be a concern for safety and security of the data. The data has created new challenges related to development of methods for visualizing and searching information. Organizations use different policies in collecting and transferring the data. There will be privacy issues, if the data is in public databases, as there will be personal genetic information stored. Genomic data faces issue in acquisition and cleansing of data into a standardized format to enable analysis and global sharing.

Storage: Data generation is inexpensive compared with storage and transfer of the same. It is difficult to store and transfer the genomic data compared to the traditional structured data as the unstructured genomic data cannot be easily standardized. Mostly, the data needed for analysis is extracted from a database and stored in a flat file or migrated to a different database. This transfer of data may lead to inaccuracy or loss of data.

Standardization: Genomic data tend to be fairly complex with multiple concepts and relationships that are maintained in different databases. There is no single standard for describing all the genomic data. Various research groups are collecting genomic information, by using various methods in many formats, which is a challenge in the standardization process. The nomenclature used is not intuitive and conflicts with other commonly accepted conventions from various research groups. For example, the amino acids can be represented by a three-letter code or one-letter code. SDTMIG uses three-letter code for amino acid, as per the International Union of Pure and Applied Chemistry (IUPAC) recommendations. Limited interoperability is big challenge for genomic data as it is rarely standardized.

Skilled Personnel: Explosion of sequencing data, building of new databases, and analyses tools have generated the need for data scientists, statisticians, and computer programmers. It is most important that database curators are kept up to date with the use of constantly changing technologies, tools and growing databases in genomics. The study designers have to plan ahead about the collection PGX data, with the study protocol data. Clinical and genetic expertise is also needed to determine how to fit data retroactively to standards and harmonize the terminology.

CHALLENGES IN MAPPING PGX DATA TO SDTM

The SDTMIG-PGx provides guidance on implementation of the SDTM for biospecimen collection, specimen handling and genetic data, such as genetic variation, gene expression, cytogenetics, viral genetics and proteomics. Depending on the nature of the genetic data collected, one or more SDTM implementation guides need to be used in addition to SDTMIG-PGx, to map to different domains.

Pharmacogenomics data can be mapped into three general categories:

1. Biospecimen domains: Biospecimen domains (BE, BS including RELSPEC) are not just limited to specimens obtained in the genetic study, leading to some question as to which implementation guide for the SDTM is the most appropriate place for these specifications. These domains collect information from both clinical (human) and non-clinical (Bacteria, Virus) specimens.

Biospecimen Events (BE) - Include the data related to the action taken (e.g., transportation, freezing, thawing), the action occurred (date/time) and the party accountable (e.g., site, lab) for the specimen.

Presently, it is an events domain, but it is more suitable as Activities domain, which is fourth general observation class currently proposed for SDTM.

Hypothetical Example 1

In this example the sample is collected, flash frozen and shipped to another location. Some samples are very sensitive to temperature and time spent in transit.

BEREFID should contain only the specimen IDs. For the final aliquot, it should be the child Specimen ID rather than the parent ID.

SPDEVID in 1st row identifies the container of the specimen collected. In the 3rd row, it identifies the freezer number in which the specimen is stored. In the 5th row, it identifies the shipping container.

BEPARTY is the individual or organization that is responsible for the biospecimen as a result of the activity performed in the associated BETERM variable. BEPRTYID is null in the 5th row, as there is only one BIO Lab.

Table 1 Biospecimen Events domain showing specimen collection, freezing, thawing and shipping.

ROW	STUDYID	DOMAIN	USUBJID	SPDEVID	BESEQ	BEREFID	BETERM	BEDECOD	BECAT
1	ABC-001	BE	ABC-001-0123	TU30678	1	1011.234	Collected	COLLECTED	COLLECTION
2	ABC-001	BE	ABC-001-0123		2	1011.234	Flash Frozen	FLASH FROZEN	PREPARATION
3	ABC-001	BE	ABC-001-0123	20581	3	1011.234	Stored in Freezer	STORED IN	STORING
4	ABC-001	BE	ABC-001-0123		4	1011.234	Thaw	THAW	PREPARATION
5	ABC-001	BE	ABC-001-0123	R14325	5	1011.234	Shipped	SHIPPED	TRANSPORT

ROW	BELOC	BEPARTY	BEPRTYID	VISITNUM	VISIT	BEDTC	BESTDTC	BEENDTC
1(Cont)	ENDOCERVIX	SITE	001	1	BASELINE	2013-06-01	2013-06-01T17:06	
2(Cont)		SITE	001	1	BASELINE	2013-06-01	2013-06-01T17:06	2013-06-01T19:58
3(Cont)		SITE	001	1	BASELINE	2013-06-01	2013-06-01T17:06	2013-06-02T13:27
4(Cont)		SITE	001	1	BASELINE	2013-06-01	2013-06-02T13:27	2013-06-02T13:34
5(Cont)		BIO LAB		1	BASELINE	2013-06-01	2013-06-02T14:02	2013-06-02T17:02

The Specimen type is given in a supplemental qualifier, which is like a Findings variable. The value in QVAL is taken from the SPECTYPE codelist.

Table 2 Supplemental qualifier for BE showing the specimen type.

ROW	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG
1	ABC-001	BE	ABC-001-0123	BEREFID	1011.234	BESPEC	Specimen Type	EPITHELIAL CELL	CRF

Biospecimen Findings (BS) - Include the details regarding the characteristics biospecimen and extracted samples (e.g., RNA, DNA) such as specimen volume, specimen condition etc.,. Specimen handling is important to maintain the integrity of the specimens used in genetic variation and gene expression testing.

Hypothetical Example 1

In this example the sample volume, temperature, and flash freeze material are mapped.

For genetic material, BSSPEC value is drawn from the GENSMP (C111114) codelist. Non-genetic values are drawn from SEND terminology, SPEC (C77529) codelist. BSANTREG is used to further define BSSPEC when it is desirable to identify specific region with in an organ.

Table 3 Biospecimen Findings domain showing the characteristics of specimen.

ROW	STUDYID	DOMAIN	USUBJID	BSSEQ	BSGRPID	BSREFID	BSTESTCD	BSTEST	BSCAT
1	ABC-001	BS	ABC-001-0123	1		1011.234	VOLUME	Volume	SPECIMEN MEASUREMENT
2	ABC-001	BS	ABC-001-0123	2	234FF	1011.234	FFRZTMP	Flash Freeze Temperature	SPECIMEN HANDLING
3	ABC-001	BS	ABC-001-0123	3	234FF	1011.234	FFRZMAT	Flash Freeze Material	SPECIMEN HANDLING

ROW	BSORRES	BSORRESU	BSSTRESC	BSSTRESN	BSSTRESU	BSSPEC	BSANTREG	VISITNUM	BSDTC
1(Cont)	1	cm^3	1	1	cm^3	CERVIX	MUCOSA	1	2013-06-01
2(Cont)	-80	C	-80	-80	C	CERVIX	MUCOSA	1	2013-06-01
3(Cont)	DRY ICE/ ISOPROPANOL		DRY ICE/ ISOPROPANOL			CERVIX	MUCOSA	1	2013-06-01

RELREC relates the records between BE and BS domains. The records for flash frozen event to its temperature can be linked together. The specimen volume from BS domain is tied with specimen collection in BE domain. The flash frozen event from BE domain is tied to temperature in BS domain.

Table 4 RELREC domain showing the relationship between BE and BS domains.

ROW	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
1	ABC-001	BE	ABC-001-0123	BESEQ	1		1
2	ABC-001	BS	ABC-001-0123	BSSEQ	1		1
3	ABC-001	BE	ABC-001-0123	BESEQ	2		2
4	ABC-001	BS	ABC-001-0123	BSGRPID	234FF		2

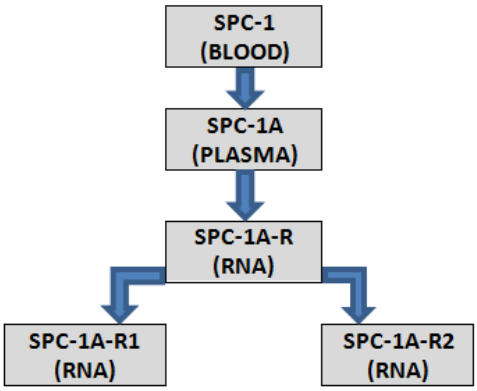
Related Specimens (RELSPEC) – RELSPEC holds the hierarchy of specimen relationships, such as specimen is re-sectioned or aliquoted, and keeps track of the relation between the aliquot and the original sample. RELSPEC domain preserves the specimen hierarchy. It is not used to relate any other datasets or domains. There are three CDISC controlled terminology codelists that can be used in the SPEC variable:

SPEC (C77529) – Specimen codelist from SEND terminology

SPECTYPE (C78734) – Specimen Type codelist from SDTM terminology

GENSMP (C111114) **–** Genetic Sample Type codelist from SDTM terminology

Figure 1 Specimen genealogy showing the relationship between the original specimen and aliquoted specimen.



Hypothetical Example 1

In this example the sample specimen lineage is mapped.

PARENT variable identifies the REFID of the parent of a specimen to track the genealogy of the specimen. The 1st row in PARENT variable is null, because it is the original sample collected.

SPEC variable has the values drawn from SPECTYPE/GENSMP codelist.

LEVEL variable has the generation number of the sample.

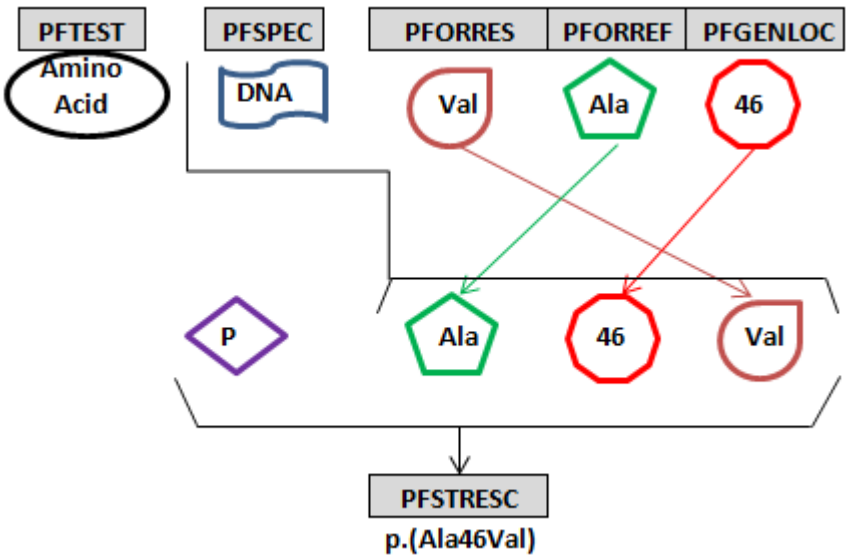
Table 5 RELSPEC domain showing the specimen hierarchy.

ROW	STUDYID	USUBJID	REFID	SPEC	PARENT	LEVEL
1	ABC-001	ABC-001-0123	SPC-1	BLOOD		1
2	ABC-001	ABC-001-0123	SPC-1A	PLASMA	SPC-1	2
3	ABC-001	ABC-001-0123	SPC-1A-R	RNA	SPC-1A	3
4	ABC-001	ABC-001-0123	SPC-1A-R1	RNA	SPC-1A-R	4
5	ABC-001	ABC-001-0123	SPC-1A-R2	RNA	SPC-1A-R	4

2. Genetic Observation domains:

Pharmacogenomics/Genetics Findings (PF) - PF domain is a findings domain. It contains the results of genetic variation and gene expression tests. Reporting a result in genetic variation is a complex concept to explain in SDTM, as it has more than one piece of information. In SDTM, each variable holds only one piece of information. As the genetic variation is not a set of multiple results but a single complex result, it has to be parsed out into different variables. For example, the position of nucleotide (PFGENLOC), the observed nucleotide (PFORRES) and the expected nucleotide (PFORREF) according to the reference sequence need to be stored in three different variables. The complete variation is represented as a whole in PFSTRESC variable, as given in the standard scientific format or nomenclature.

Figure 2 Genetic variation data represented in different variables in SDTM



Hypothetical Example 1 (Genetic Variation)

In this example both amino acid and the nucleotide variations are shown for the virus and the study subject.

PFNSPCES and PFNSTRN variables are populated for the virus's species and strain in 1st and 2nd row.

PFREFSEQ is used to map reference sequence used in identifying the genetic variation.

PFTEST and PFTESTCD should not contain gene name or symbol. These variables contain the type of genetic material, such as nucleotide, amino acid, codon or allele. PFGENRI can be used to collect the genetic region of interest.

PFREFID contains the unique identifier for the genetic sample.

PFGRPID is not mapped as PFRSNUM is sufficient to group the amino acid and nucleotide records.

PFRUNID is used to distinguish between records for the same test performed at different times.

Table 6 Genetic variations from both virus and study subject are shown.

ROW	STUDYID	DOMAIN	USUBJID	PFSEQ	PFREFID	PFTESTCD	PFTEST	PFGENRI	PFREFSEQ	PFCAT	PFNSPCES	PFNSTRN
1	ABC-001	PF	ABC-001-0123	1	XYZ-004	AA	Amino Acid	E6	NC_001526.4	GENETIC VARIATION	HPV	33
2	ABC-001	PF	ABC-001-0123	2	XYZ-004	NUC	Nucleotide	E6	NC_001526.4	GENETIC VARIATION	HPV	33
3	ABC-001	PF	ABC-001-0123	3	XYZ-009	AA	Amino Acid	P53	NM_000546.5	GENETIC VARIATION		
4	ABC-001	PF	ABC-001-0123	4	XYZ-009	NUC	Nucleotide	P53	NM_000546.5	GENETIC VARIATION		

ROW	PFORRES	PFORREF	PFGENLOC	PFSTRESC	PFRSNUM	PFNAM	PFSPEC	PFMETHOD	PFRUNID	PFBLFL	VISITNUM	PFDTC
1(Cont)	Val	Ala	46	p.(Ala46Val)	rs1110222	QLAB	DNA	Parallel Sequencing	D256894-001	Y	1	2013-06-01
2(Cont)	T	C	245	c.245C>T	rs1110222	QLAB	DNA	Parallel Sequencing	D256894-001	Y	1	2013-06-01
3(Cont)	Arg	Pro	72	p.(Pro72Arg)	rs1042522	QLAB	DNA	Parallel Sequencing	D256894-007	Y	1	2013-06-01
4(Cont)	G	C	215	c.215C>G	rs1042522	QLAB	DNA	Parallel Sequencing	D256894-007	Y	1	2013-06-01

Hypothetical Example 2 (Gene Expression)

The quantitative RT-PCR (qRT-PCR) is the most powerful, sensitive and quantitative assay used in the detection of RNA levels.

It is frequently used in the expression analysis of single or multiple genes, and expression patterns for identifying infections and diseases.

In qRT-PCR, the RNA sample is reverse transcribed to cDNA. Then the cDNA is further quantified via conventional PCR in thermal cycler.

The rate of generation of the amplified product is measured at each PCR cycle and the number of cycles required to reach a defined signal threshold is called threshold cycle (Ct).

The data generated can be analyzed by software to calculate the relative gene expression (mRNA copy number) in several samples to produce a measurement of the quantity of cDNA produced by the reverse transcriptase step of the assay.

This example shows the gene expression data obtained through quantitative reverse transcriptase polymerase chain reaction (qRT-PCR).

Table 7 Gene expression results from qRT-PCR

ROW	STUDYID	DOMAIN	USUBJID	PFSEQ	PFGRPID	PFREFID	PFTESTCD	PFTEST	PFGENRI	PFCAT	PFORRES	PFORRESU
1	ABC-001	PF	ABC-001-0123	1	1	NM_003194	RAWCT	Raw Ct Value	TBP	GENE EXPRESSION	29.684243	Cycles
2	ABC-001	PF	ABC-001-0123	2	1	NM_003194	RAWCT	Raw Ct Value	TBP	GENE EXPRESSION	29.088622	Cycles
3	ABC-001	PF	ABC-001-0123	3	1	NM_003194	RAWCT	Raw Ct Value	TBP	GENE EXPRESSION	29.499651	Cycles
4	ABC-001	PF	ABC-001-0123	4	1	NM_003194	MEANCT	Mean Ct Value	TBP	GENE EXPRESSION	29.424172	Cycles
5	ABC-001	PF	ABC-001-0123	5	1	NM_003194	COPYNUM	Copy Number	TBP	GENE EXPRESSION	1.47	Copies/ng
6	ABC-001	PF	ABC-001-0123	6	2	NM_002046	RAWCT	Raw Ct Value	GAPDH	GENE EXPRESSION	21.33377	Cycles
7	ABC-001	PF	ABC-001-0123	7	2	NM_002046	RAWCT	Raw Ct Value	GAPDH	GENE EXPRESSION	21.22124	Cycles
8	ABC-001	PF	ABC-001-0123	8	2	NM_002046	RAWCT	Raw Ct Value	GAPDH	GENE EXPRESSION	21.56392	Cycles
9	ABC-001	PF	ABC-001-0123	9	2	NM_002046	MEANCT	Mean Ct Value	GAPDH	GENE EXPRESSION	21.372978	Cycles
10	ABC-001	PF	ABC-001-0123	10	2	NM_002046	COPYNUM	Copy Number	GAPDH	GENE EXPRESSION	0.81	Copies/ng

ROW	PFSTRESC	PFSTRESN	PFSTRESU	PFNAM	PFSPEC	PFMETHOD	PFRUNID	PFLLOQ	PFREPNUM	VISITNUM	VISIT	PFDTC
1(Cont)	29.684243	29.684243	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1	0.1	1	-1	Pre-Treatment	2013-06-01T19:56
2(Cont)	29.088622	29.088622	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1	0.1	2	-1	Pre-Treatment	2013-06-01T19:56
3(Cont)	29.499651	29.499651	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1	0.1	3	-1	Pre-Treatment	2013-06-01T19:56
4(Cont)	29.424172	29.424172	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1			-1	Pre-Treatment	2013-06-01T19:56
5(Cont)	1.47	1.47	Copies/ng	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1			-1	Pre-Treatment	2013-06-01T19:56
6(Cont)	21.33377	21.33377	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1	0.1	1	-1	Pre-Treatment	2013-06-01T19:56
7(Cont)	21.22124	21.22124	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1	0.1	2	-1	Pre-Treatment	2013-06-01T19:56
8(Cont)	21.56392	21.56392	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1	0.1	3	-1	Pre-Treatment	2013-06-01T19:56
9(Cont)	21.372978	21.372978	Cycles	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1			-1	Pre-Treatment	2013-06-01T19:56
10(Cont)	0.81	0.81	Copies/ng	Vendor X	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1			-1	Pre-Treatment	2013-06-01T19:56

Pharmacogenomics/Genetics Methods (PG) - This is a supporting information domain containing additional data that may affect interpretation of PF data, such as setup process, test or sequencing parameters.

This domain is used for mapping both clinical and non-clinical results obtained from the study subject and infectious viruses or microbes.

PGREFID contains the identifier for the genetic sample.

PGTESTCD and PGTEST should not include gene names or symbols and use the CDISC controlled terminology codelists. PGGENRI is used to collect the gene of interest.

PGNSPCES and PGNSTRN are used to identify the species and strain of infectious microorganisms and viruses. Records in PGNSPECS is not populated for the study subject.

Hypothetical Example 1

In this example the test parameters and details of for a qRT-PCR run are shown.

The sequence length (PGTEST) of the gene is recorded in PGORRES, PFSTRESC and PFSTRESN. PGCAT identifies the genetic technique used.

Table 8 Test parameters and details for a qRT-PCR run performed to determine gene expression.

Row	STUDYID	DOMAIN	USUBJID	PGSEQ	PGGRPID	PGREFID	PGTESTCD	PGTEST	PGGENRI
1	ABC-001	PG	ABC-001-0123	1	1	NM_003194	SEQLNGTH	Genetic Sequence Length	TBP
2	ABC-001	PG	ABC-001-0123	2	2	NM_002046	SEQLNGTH	Genetic Sequence Length	GAPDH

Row	PGCAT	PGORRES	PGSTRESC	PGSTRESN	PGSPEC	PGMETHOD	PGRUNID	VISITNUM	PGDTC
1(Cont)	GENE EXPRESSION	1903	1903	1903	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999910_m1	-1	2013-06-01T19:56
2(Cont)	GENE EXPRESSION	1875	1875	1875	RNA	QUANTITATIVE REVERSE TRANSCRIPTASE POLYMERASE CHAIN REACTION	Hs99999905_m1	-1	2013-06-01T19:56

3. Genetic Biomarker domains:

Pharmacogenomics/Genetics Biomarker (PB) – PB is a special purpose domain that serves as a reference to associate observed genetic variations with medical conclusions (e.g., disease diagnosis, resistance of a virus to particular drug). PB does not hold the subject data, but instead the genetic variations that serve as biomarkers of interest to the study. PBSTMT variable holds the medical statement, drawn by the implications of the variations caused by the drug (PBDRUG) or the diagnosis of a medical condition (in PBDIAG) associated with the genetic biomarker. PBMRKR identifies the individual variant and its value is derived from the standard nomenclature. PBMRKRID is used to group genetic variation records which belong to a set and which form the basis for medical statement inference. When more than one PBMRKR contribute to the PBMRKRID, it is recommended that the value in PBMRKRID be formed from the short names of the genetic variations that make up the set, separated with a plus (+) symbol. This domain holds only genetic and genomic data and primarily designed to track the changing science.

Subject Biomarker (SB) – SB is a special purpose domain that holds the data about genetic biomarkers (as defined in PB) that a subject may have. It is currently a special purpose domain, but the data for which it was designed may be represented in findings domain in future. SB domain provides a linkage between the genetic findings for a subject in PF domain and clinical statement (PBSTMT) in PB domain. Sometimes only the biomarker data is collected for the subject and the genetic variation information is not collected. In such cases only SB domain is mapped and not the PF domain. The value in SBMRKRID should match the value in PBMRKRID in PB. SBNSPCES and SBNSTRN are used only to identify the species and strain of infectious microbes and viruses. This domain is currently defined for genetic and genomic data only. Data in SB domain may originate from CRFs (when only the biomarker is collected) or be derived at a lab (when complete set of data are generated).

CONCLUSION

Genomic data and clinical trial data are collected in different databases, because of the size, structure and formats. It is a challenge when it comes to integrate and standardize the genomic data as it comes in the form of sequences, images, annotations, research articles etc.,. Many organizations are developing systems or knowledge bases to retrieve and store the information. In addition to collection and curation of the genomic data, SDTM modelling is going to be a challenge as the data is obtained by using various methods and assays. Gene expression studies can be mapped differently in SDTM based on the sample used, method performed and the results obtained. For example, BE is currently an events domain, However, it is better suited as an Activities domain, which is a fourth general observation class proposed for SDTM. The class of BE domain may change in future, depending on implementation of Activities class. It is still a question if the PB domain can be mapped as special purpose or trial design domain. SB is a special purpose domain at present, but it may be represented in findings domain later. Thus, the SDTM domains and variables may change based on the need and usage in analysis of the genomic data.

Qualified and trained experts are needed in harmonizing the genomic data, as it has multiple concepts and different naming conventions. In conclusion, the challenges surrounding PGx data and its implementation in SDTM will continue to evolve as the standards strive to keep up with the advancement of the technologies and science of genomics.

REFERENCES

Evans WE and Relling MV.1999. “Pharmacogenomics: translating functional genomics into rational therapeutics”. Science, 286(5439):487–491.

Grossman RL. 2012. “Managing and Analysing 1,000,000 Genomes”. Available at http://rgrossman.com/2012/09/18/million-genomes-challeng/

CDISC: Study Data Tabulation Model Implementation Guide: Pharmacogenomics/Genetics Version 1.0 Available at: https://www.cdisc.org/system/files/members/standard/foundational/pgx

ACKNOWLEDGMENTS

I would like to thank Dana House for reviewing the paper and giving valuable input.

CONTACT INFORMATION

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

Rama Kudaravalli
Syneos Health
610-306-8974
rama.kudaravalli@syneoshealth.com