Paper DS12

Challenges of Genomic Data in Clinical Trials

Lingsong Yun, Bayer Inc., Toronto, Canada
Hong Zheng, Bayer Inc., Toronto, Canada

ABSTRACT

Biomarkers are widely used in every stage of drug development, especially in targeted cancer therapies. Next-Generation Sequencing (NGS) based on tissue or liquid biopsies are commonly used to select patients, predict treatment responses, and monitor genetic alterations related to therapy resistance. NGS genomic data are complex, poorly standardized, and are not fully integrated into clinical trial development.
Key questions include: What data elements are required? How can genomic information be harmonized and annotated? What challenges are encountered when NGS data including CDx data is collected and integrated into SDTM? What challenges exist in abstracting genetic information into analysis data sets (ADS), and how can these data be effectively summarized in the analysis phase?
This article will explore the challenges and complexities surrounding the collection, structuring, standardization, integration, data readiness and analysis of genomic data in clinical trials.

Deleted: ¶

INTRODUCTION

Biomarkers are widely used in every stage of drug developments. Predictive biomarkers are especially crucial in targeted cancer therapies, guiding the selection of treatments tailored to individual molecular profiles [1, 2, 3]. The next-generation sequencing (NGS) approach has become an invaluable technique in diagnostic and assay developments for identifying biomarkers. NGS enables the comprehensive sequencing of large numbers of genes in a single test, offering the capability to detect a wide range of genomic alterations. Tissue biopsy is a standard biomarker sample type applying NGS to select patients. Circulating tumor DNA (ctDNA) based on liquid biopsy sample from blood has been a promising technology in both patient selections and disease monitors [4, 5, 6, 7].

The complexity of NGS genomic data and its vast content can often be overwhelming, particularly when integrating this data with clinical data. Significant challenges arise from the harmonization of genomic data, the need for additional genome variant annotation, and the rapidly evolving landscape of diagnostic testing, including the integration of companion diagnostics (CDx).
This article seeks to address key questions related to the standardization and utilization of NGS data:

- What data elements are captured during NGS testing, and how are these standardized?
- How can genetic change information be harmonized and annotated to ensure consistency and accuracy?
- What challenges are encountered when NGS data including CDx data is collected and integrated into SDTM?
- What do challenges exist in abstracting genetic information into analysis data sets (ADS)?

All data included in the article are simulated for illustrations only.

GENOMIC DATA

GENOMIC DATA ELEMENTS AND THE GENOMIC FINDINGS DOMAIN

The CDISC Study Data Tabulation Model Implementation Guide (SDTM Version 3.4) finalized the Genomics Finding (GF) Domain in its 2021 and 2022 releases [9]. This guide provides implementation details and controlled terminology guidance for capturing genomic data findings in clinical trials. Genomic NGS data typically comprises three levels of information: specimen samples, genes, and variant details, for each subject at each time point.

- Sample level data contains information such as vendor name, test assays, test method (e.g. Next Generation Sequencing [NGS]), and the types of specimen material tested (e.g., DNA, RNA, mRNA, ctDNA). These samples are usually collected at specific time points, such as screening, baseline, or during each scheduled visit or treatment cycle.

- Gene level data contains information such as gene symbol, chromosome identifier, and genome reference. A single sample includes usually many genes. The data from both the sample and gene levels are typically stored under different variables.

- Variant-level data provides variant details associated with specific genes. These variant details are categorized using two genomic measurement variables, GFTEST and GFTSTDTL. GFTEST captures the types of alterations, such as short variation, tumor mutation burden, single nucleotide variation (SNV), and copy number variation, while GFTSTDTL usually includes variant-specific details, such as amino acid change, predicted coding sequence change, variant impact classification, read depth, allele frequency, and more. The original result findings are stored in GFORRES. The unique variant identifier and genetic location are under the variables, GFPVRID and GFGENLOC respectively. The result category (FGRESCAT) and the sub-category of genomic finding (GFSCAT). The result category (FGRESCAT) and the sub-category of genomic finding (GFSCAT) are useful to categorize the results of original genomic finding results. Categories enable meaningful grouping (e.g., from annotations) for analysis and reporting. For example, the classifications could be "mutation detected', 'Resistance mutation', 'point mutation', 'on-target mutation', or 'off-target mutation' based on study specific conventions or standard terminologies. A unique GFGRPID (Group Identifier) should be assigned to each unique variant, ensuring proper identification and linkage across the data.

The Table 1 shows an example of sample and gene level data variables for a Single Nucleotide Variation (SNV) captured in the GF domain. The data is from a ctDNA NGS technology test by the 'abc precision assay'. Five genes, ALK, BRCA2, EGFR, NTRK3, and KRAS, with their chromosome location are presented.

Table 1: An example of sample and gene data for single nucleotide variation in Genomic Finding (not show variant level data).

STUDYID	USUBJID	GFGRPID	GFTEST	GFSYM	GFCHROM	GFGENREF	GFSPEC	GFMETHOD	SPGEVID	GFNAM
Study Identifier	Unique Subject Identifier	Group ID	Name of Genomic Measurement	Genomic Symbol	Chromosome Identifier	Genome Reference	Specimen Material Type	Method of Test	Sponsor Device Identifier	Vendor Name
ABCD-123	123101	1111	Single Nucleotide Variation	ALK	Chr2	GRCh37.p5	ctDNA	NGS	abc PRECISION ASSAY	abc SOLUTIONS
ABCD-123	123101	1112	Single Nucleotide Variation	BRCA2	Chr13	GRCh37.p5	ctDNA	NGS	abc PRECISION ASSAY	abc SOLUTIONS
ABCD-123	123101	1113	Single Nucleotide Variation	EGFR	chr7	GRCh37.p5	ctDNA	NGS	abc PRECISION ASSAY	abc SOLUTIONS
ABCD-123	123101	1114	Single Nucleotide Variation	NTRK3	Ch15	GRCh37.p5	ctDNA	NGS	abc PRECISION ASSAY	abc SOLUTIONS
ABCD-123	123101	1115	Single Nucleotide Variation	KRAS	chr12	GRCh37.p5	ctDNA	NGS	abc PRECISION ASSAY	abc SOLUTIONS
				Level 2 at gene: gene and reference			Level 1 at sample: assay/testing/vendor			

In the Table 2, The Row 1-4 shows details (GFTSTDTL) of a variant G12D (amino acid change) on the Gene KRAS. The unique variant identifier is COSV55497369. All the variant details are under the variable GFTSTDTL. GFSCAT categorizes these variant details as 'GENETIC CHANGE'.

ANNOTATION DATA

Genomic finding results can't be fully defined and classified for clinically meaningful analysis from the raw NGS sequencing data in the GF domain. For instance, defining off-target resistance mutations (off-target resistance mutations occurred outside the intended molecular target of a therapy indirectly affect the therapy's effectiveness, allowing the disease to persist or progress) requires proper variant categorization and association with specific cancers. This is achieved through an annotation procedure identifying and mapping genes, transcripts, regulatory regions, and other functional elements in a genome, as well as characterizing genetic variants for their functional and clinical implications.

The annotation procedure is extremely complex and requires linking unique variants to public databases and analyzing their significance, typically conducted by computational biologists or bioinformaticians. The Catalogue of Somatic Mutations in Cancer (COSMIC) is a widely used database to catalog genes with a proven role in cancer, and to distinguish between driver genes and passenger genes mutated incidentally without functional impact [10]. The Cancer Gene Census (CGC) is a key component of the COSMIC database for targeted cancer research and drug

development. It its particularly valuable for studying resistance mechanism linked to specific gene alterations [11, 13].

In the Table 2, in addition to the KRAS-G12D mutation details, the Row5-8 shows the annotation data (GFSCAT) categorized as gene type, variant impact classification, mutation status, and COSMIC mutation tier (GFTSTDTL).

Table 2: Examples of single nucleotide variation (variant level) with annotation data and CDx data in Genomic Finding.

Row	USUBJID	GFGRPID	GFSCAT	GFTEST	GFTSTDTL	GFORRES	GFPVRID	GFGENLOC	GFSYM	GFCHROM
Row	Unique Subject Identifier	Group ID	Subcategory for Genomic Finding	Name of Genomic Measurement	Measurement, Test, or Examination Detail	Result or Finding in Original Units	Published Variant Identifier	Genetic Location	Genomic Symbol	Chromosome Identifier
1	123101	1115	GENETIC CHANGE	Single Nucleotide Variation	AMINO ACID CHANGE	G12D	COSV55497369	25398284	KRAS	chr12
2	123101	1115	GENETIC CHANGE	Single Nucleotide Variation	PREDICTED CODING SEQUENCE CHANGE	C>T	COSV55497369	25398284	KRAS	chr12
3	123101	1115	GENETIC CHANGE	Single Nucleotide Variation	Percent Of Genetic Change	0.43	COSV55497369	25398284	KRAS	chr12
4	123101	1115	GENETIC CHANGE	Single Nucleotide Variation	HGVS Protein Mutation	NP_004976.2:p.(Gly12Asp)	COSV55497369	25398284	KRAS	chr12
5	123101	1115	ANNOTATION	Single Nucleotide Variation	Gene type	oncogene	COSV55497369	25398284	KRAS	chr12
6	123101	1115	ANNOTATION	Single Nucleotide Variation	Variant impact classification	missense	COSV55497369	25398284	KRAS	chr12
7	123101	1115	ANNOTATION	Single Nucleotide Variation	Mutation status	Mutated	COSV55497369	25398284	KRAS	chr12
8	123101	1115	ANNOTATION	Single Nucleotide Variation	Cosmic mutation tier	Red - known pathogenic	COSV55497369	25398284	KRAS	chr12
9	123101	1116	CDx	Sequence Rearrangement	REARRANGEMENT GENE 1	NTRK3			NTRK3	Chr15
10	123101	1116	CDx	Sequence Rearrangement	REARRANGEMENT GENE 2	ETV6			ETV6	chr12
11	123101	1116	CDx	Sequence Rearrangement	DETECTION	Yes			NTRK3	Chr15
12	123101	1116	CDx	Sequence Rearrangement	DETECTION	No			NTRK1	Chr1
13	123101	1116	CDx	Sequence Rearrangement	DETECTION	No			NTRK2	Chr9

COMPANION DIAGNOSTICS DATA

Oncology target therapies require identifying patients with positive biomarker who are likely to benefit from the therapy. Clinical trial protocols may specify one or more Clinical Trial Assays (CTAs) to test for biomarker positivity. These assays, usually central laboratory test, may not require full regulatory approvals. Companion diagnostics (CDx) are specialized in vitro diagnostic (IVD) assays that are critical adjuncts in precision medicine to accurately identify patients who may be more likely to respond to a targeted therapy. CDx assays requires a pre-market approval (PMA) application. The development of Companion Diagnostics (CDx) is a requirement by the FDA for targeted anti-cancer therapies [8].

Due to the typically accelerated timeline of oncology drug development, CDx development may not always align with the drug's timeline. During clinical trials, Clinical Trial Assays (CTA) are often used to identify biomarker-positive patients for enrollment. Subsequently, baseline tissue or plasma samples are retrospectively tested using the CDx assay to determine biomarker status. This retrospective testing establishes a CDx clinical validation study, also known as a CDx bridging study, to bridge clinical outcomes to CDx assay results. The development of CDx assays requires collaboration between pharmaceutical companies and device manufacturers to ensure alignment in development and validation processes.

CDx sequencing data is provided by the device company to the pharmaceutical. Similar to genomic data generated by CTAs, CDx data typically contains two key types of genetic sequencing information, biomarker Status used in the bridging study and genetic changes used in mutation analysis. In the Table 2, the Rows 9-13 show NTRK3-ETV6 fusion is detected by the CDx assay. GFSCAT categorizes as 'CDx' and GFTEST as 'Sequence Rearrangement'.

Within the three genes NTRK1-3, only NTRK3 is detected as 'Yes'.

GENOMIC DATA CHALLENGES IN STANDARDIZATION AND UTILIZATION

Although the SDTM IG 3.4 provides a framework and specifications for managing and utilizing complex genomic data within the GF domain, significant challenges remain. These challenges primarily impact the statistical analytics and biometrics as they work to

- Integrate diverse and complex genomic data from various sources.
- Standardize and define genomic data to ensure consistency and usability
- Analyze genomic data efficiently to address critical scientific questions.

DIFFERENT GENOMIC RAW DATA ACROSS VENDORS

Differences in genomic raw data across vendors present a significant challenge. Vendors use varying file formats such as FASTQ, VCF, and BAM, which complicates alignment with the SDTM GF domain structure. Additionally, misalignments arise when data are mapped to different reference genomes, such as GRCh37 versus GRCh38. Gene and variant systems further contribute to inconsistencies, with some vendors using Ensembl (by European Bioinformatics Institute) and others using RefSeq (by National Center for Biotechnology Information). The involvement of multiple vendors in studies adds complexity, as data often cannot be seamlessly integrated across vendors. Vendors also differ in their ability to handle data, ensure traceability, provide complete metadata, and meet regulatory requirements.

DATA HARMONIZATIONS

Data harmonization also face hurdles due to the lack of standardization in data collection protocols, sequencing technologies, and nomenclature. Synonyms and inconsistent formats create further challenges [13]. For example, the ALK gene may be reported as ALK or CD246 in different data sources, and a same amino acid change may be represented in different formats such as Gly12Ser (three-letter format) or G12S (one-letter format). Expertise gaps may complicate these issues, as statistical programmers often lack knowledge in computational biology, while bioinformaticians may not fully understand regulatory-compliant CDISC databases.

DATA CONTENT MAPPING CHALLENGES

Mapping genomic data to SDTM GF domain variables is a complex task due to several factors [12].

Genomic data involves various specimen types, such as tissue and plasma, and genetic types, including DNA, RNA, mRNA, ctDNA, and cfDNA. These different genetic types may originate from different specimens and be tested at varying time points. For example, tissue samples from surgeries or biopsies are often used at screening, while plasma-based ctDNA testing can occur both at screening and during treatment. This variability requires careful alignment of specimen sources, genetic types, and testing time points to ensure accurate data mapping.

NGS method has a rigorous quality control procedure. Failure in quality control such as due to insufficient DNA or RNA extraction or poor sample quality can lead to incomplete sequencing, resulting in missing genetic variant records. The Table 3 shows a patient has genetic testing at screening and Cycle 1-4. ctDNA, DNA, and RNA are abstracted at screening and only ctDNA is tested during the treatment period. The ctDNA testing failed quality control at Cycle 1-Day 15 and Cycle 3-Day1 will not have any genetic changes records for variants and details because of unsuccessful NGS sequencing. The quality control information will contribute to the sample accountability analysis at each time point.

A lack of familiarity with genetic concepts, terminologies, and abbreviations among programmers adds the additional challenges of mapping genomic data to SDTM domains.

Table 3: Example quality control data for multiple genetic types at different time points

STUDYID	USUBJID	GFGRPID	GFTEST	GFTSTDTL	GFORRES	AVISIT	GFSPEC	GFMETHOD
Study Identifier	Unique Subject Identifier	Group ID	Name of Genomic Measurement	Measurement, Test, or Examination Detail	Result or Finding in Original Units	Genomic Symbol	Specimen Material Type	Method of Test
ABCD-123	123101	0	Quality Control		PASS	Screening -28	ctDNA	NGS
ABCD-123	123101	0	Quality Control		PASS	Screening -28	DNA	NGS
ABCD-123	123101	0	Quality Control		PASS	Screening -28	RNA	NGS
ABCD-123	123101	0	Quality Control		PASS	Cycle 1, Day 1	ctDNA	NGS
ABCD-123	123101	0	Quality Control		FAIL	Cycle 1, Day 15	ctDNA	NGS
ABCD-123	123101	0	Quality Control		PASS	Cycle 2, Day 1	ctDNA	NGS
ABCD-123	123101	0	Quality Control		FAIL	Cycle 3, Day 1	ctDNA	NGS
ABCD-123	123101	0	Quality Control		PASS	Cycle 4, Day 1	ctDNA	NGS

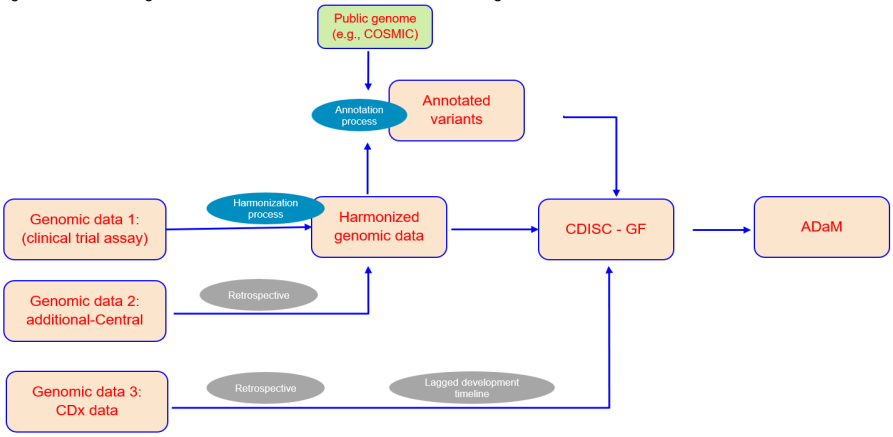
DATA INTEGRATIONS

Integrating harmonized genomic raw data, annotation data, and CDx data into one GF domain presents significant challenges due to retrospective genomic data collection, multiple vendor or partner company involvements, and different timelines between drug and CDx device developments.

The diagram (Figure 1) illustrates a common workflow for processing genomic raw data from vendors, harmonization and annotation to the SDTM-GF domain. Unlike other clinical trial data, genomic data typically includes three types of collections, each following distinct timelines:

- Clinical Trial Assays (CTA): Genetic testing conducted at enrollment follows similar procedures to other clinical trial data.
- Retrospective Testing: Genetics may be retested during or after treatment using central laboratory assays.
- CDx Development: Health authority-mandated Companion Diagnostics (CDx) development often begins much later than the associated drug trials. While CDx data collection may coincide with drug developments, the CDx submission timeline frequently lags by several months to years. CDx submissions sometime may occur after drug approval and commercialization especially when anti-cancer drugs in the situation of accelerated approval pathways.

Figure 1: Process of genomic data harmonization, annotation, and integration to GF domain.



MISSING DATA

Missing data brings significant complications for genomic data analysis and potential bias. The sources of missing

data include:

- **Ineligible Samples**: Tissue slides may fail eligibility criteria due to insufficient quantity or poor quality. In many cases, archived tissue slides are used to determine baseline eligibility. However, slides obtained long time ago may degrade over time, making them ineligible for assay testing.
- **Invalid Sequencing**: NGS testing failed quality control procedure leads to unusable sequencing data.
- **Undefined or Unannotated Variants**: Incorrect genetic identifiers can cause linkage failures with public genome databases during the annotation process, resulting in undefined or unannotated variants.
- **Data Collection Gaps**: Blood samples for ctDNA NGS sequencing is often not prospectively scheduled or collected as a predefined, longitudinal assessments in clinical trial protocols. The primary reason is high Costs of ctDNA sequencing makes routine and scheduled collection financially challenging. Another reason may be due to unclear resistance mutations mechanism and their associations with disease progression. Retrospective testing and data collection often introduce a significant number of missing samples, leading to potential bias and challenges in longitudinal analysis.
- **Unconsented samples for CDx development**: Clinical trials often enroll patients globally but varying national policies on shipping bio-samples create challenges. Some countries prohibit exporting samples, and regulatory bodies have strict rules on genomic data, alongside ethical considerations such as patient consent and privacy. In many cases, consent for additional testing on bio-samples is not obtained at enrollment. As a result, significant missing data can occur, complicating the use of data in CDx developments.

ADS DEVELOPMENT STRUCTURES

Genomic data involves massive volumes, with a single sample potentially generating thousands of records, including numerous genes and multiple variants associated with each gene. The traditional approach of structuring analysis datasets with parameters as rows and criteria as variables may not adequately address the specific needs of variant analysis.

- The extremely large data volumes and the overall data size significantly increase the time required for data abstraction in analysis phase.
- Studies often focus on a limited set of genes and their variants, making it inefficient to run reporting programs on the entire ADGF dataset.
- Variant-specific details (e.g., genes, amino acid changes, allele frequencies) are scattered across multiple records in GF domain, which complicates reporting and analysis.
- Descriptive analyses, such as frequency and proportion calculations, are typically performed at the patient or visit/treatment-cycle level, requiring a more tailored data structure.
- Standard one ADGF data for genomic change findings and its structure is insufficient and inefficient to support advanced data visualization or the development of interactive applications.

The Figure 2 provides recommendations of ADGF and downstream structural relational datasets at positive variant level (ADGF-POS-VARIANT), all visit level (ADGF-VISITS), and (ADGF-ADSL). And the Table 4 describes derivation details and the usages.

Figure 2: Three-level analysis data ADGF structure

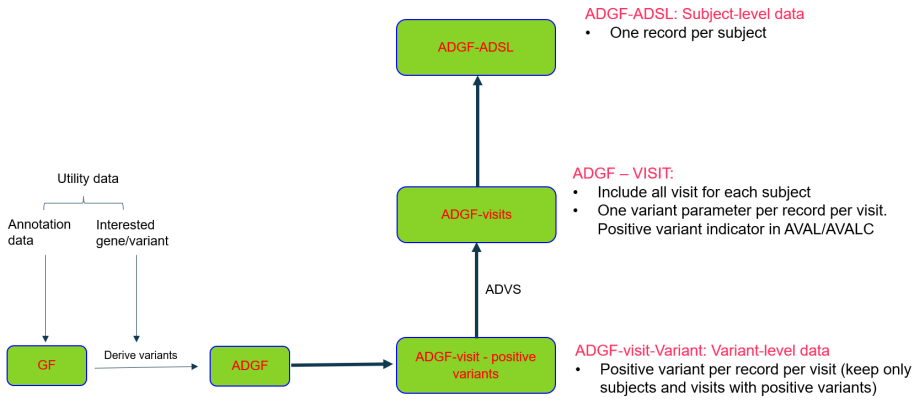


Table 4: Descriptions of three-level ADGF

Dataset	Data level	Derive	Usages
ADGF-ADSL	Subject	Variant flags	Summary at subject
ADGF-VISITS	Per variant per visit	• One variant parameter per record per visit. • VAL/AVALC indicates 'Positive' or 'Negative'	Summary at visits Longigudinal analysis
ADGF-POS-VARIANT	Positive variants at all visits	• Positive variant per record per visit • Keep subjects with positive variants in all visits	Listing of positive variants

CONCLUSION (HEADER 1)

Integrating NGS genomic data into the SDTM GF domain presents significant challenges, including managing complex data elements and terminology, handling large numbers of vendors, ensuring standardization and harmonization, accurate annotation, seamless integrations particularly in CDx development with lagged timeline, and maintaining regulatory compliance. Successfully overcoming these challenges needs strategic planning in longitudinal data collection and CDx development, as well as implementing robust data governance frameworks.

Key success factors include early engagement with vendors, establishing uniform Data Transfer Specifications (DTS) across multiple vendors, and leveraging advanced bioinformatics pipelines to streamline genomic data processing. Additionally, cross-functional collaboration between statistical programmers and bioinformaticians is critical to ensuring accurate and efficient data integration. A deep understanding of both NGS data and SDTM standards is essential to maintaining data traceability, optimizing downstream analyses, and supporting regulatory submissions.

REFERENCES (HEADER 1)

[1] Mariya Gromova. Biomarkers: Opportunities and Challenges for Drug Development in the Current Regulatory Landscape. *Biopharma Excellence* 2020
[2] Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril. 2018 Jun;109(6):952-963. doi: 10.1016/j.fertnstert.2018.05.006. PMID: 29935653; PMCID: PMC6366451.
[3] FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) resource. 2021. Accessed Dec 5, 2024. https://www.ncbi.nlm.nih.gov/books/NBK326791/
[4] Kim H, Park KU. Clinical Circulating Tumor DNA Testing for Precision Oncology. Cancer Res Treat. 2023 Apr;55(2):351-366. doi: 10.4143/crt.2022.1026. PMID: 36915242; PMCID: PMC10101787.
[5] Corcoran RB. Liquid biopsy versus tumor biopsy for clinical-trial recruitment. Nat Med. 2020;26:1815–6. doi: 10.1038/s41591-020-01169-6.
[6] Ghosh RK, Pandey T, Dey P. Liquid biopsy: a new avenue in pathology. Cytopathology. 2019;30:138–43. doi: 10.1111/cyt.12661.
[7] Marrugo-Ramirez J, Mir M, Samitier J. Blood-based cancer biomarkers in liquid biopsy: a promising non-invasive alternative to tissue biopsy.

[8] S. Badve, G. L. Kumar (eds.), Predictive Biomarkers in Oncology. Section 49: Use of Companion Diagnostics (CDx) and Predictive Biomarkers for Cancer Targeted Therapy: Clinical Applications in Precision Medicine. Springer Nature Switzerland AG 2019. https://doi.org/10.1007/978-3-319-95228-4_49
[9] Clinical Data Interchange Standards Consortium (CDISC). (2021). *Study Data Tabulation Model Implementation Guide: Version 3.4.* Austin, TX: CDISC. Retrieved from https://www.cdisc.org
[10] COSMIC: Catalogue of Somatic Mutations in Cancer. Wellcome Sanger Institute. Available at: https://cancer.sanger.ac.uk/cosmic.
[11] The Cancer Gene Census. Wellcome Sanger Institute. Available at: https://cancer.sanger.ac.uk/census
[12] Xuejie Zhou. Time to Get in the Genomics Findings (GF) Domain. PharmaSUG China 2022 - Paper SR-145
[13] Lea Zittlau. Biomarkers Unleashed: A Programming Pipeline from Chaos to Clarity in Genomic Data. Phuse EU Connect 2024 – Paper DH02

CONTACT INFORMATION

Author Name: Lingsong Yun
Company: Bayer Inc.
Address: 2920 Matheson Blvd E, Mississauga, ON L4W 5J4, Canada

Work Phone: 1 (416) 571-6495
Email: lingsong.yun@bayer.com
Website: http://www.bayer.ca