PHUSE 2026

Paper AS12

Working with Biomarker - What Every Programmer Should Know

Aboli Katdare, Cytel Statistical Software & Services Pvt. Ltd., Pune, India

ABSTRACT

Biomarker analysis plays an increasingly vital role in clinical trials by enhancing understanding of treatment response, identifying patient subgroups, and supporting go/no-go decisions. For statistical programmers, biomarker work presents unique challenges distinct from standard safety and efficacy reporting. Biomarker datasets often originate from external labs, feature complex structures, and require careful handling of missing values, detection limits, and derived variables such as high/low expression groups. This paper highlights the core knowledge statistical programmers need when working with biomarker data. Through real-world examples, we demonstrate how biomarker programming differs from traditional ADaM and TFL workflows and explore how predictive and prognostic biomarkers are evaluated using appropriate statistical methods. The paper also covers visualization techniques to explore biomarker variability and summary outputs to communicate key findings effectively. Our goal is to make biomarker programming more approachable and practical, equipping statistical programmers with tools and insights to contribute meaningfully to biomarker-driven clinical research.

INTRODUCTION

Clinical research covers a wide range of therapeutic areas, providing programmers opportunities to gain diverse exposure. A few years ago, I was assigned to work on biomarkers. At that time, I mistakenly thought of biomarkers as separate therapeutic areas. This misconception may have resulted from my limited biological background, which affected how I approached and reviewed the data. Later, I realized that many programmers, especially those without a strong background in biology, face similar challenges when first working with biomarker data. This experience motivated me to work on this paper. My goal is to clarify what biomarkers are, their importance in clinical research, and to demonstrate their uses with relevant examples, so programmers can approach this area with more insight and understanding.

SCIENCE BEHIND BIOMARKERS

The National Cancer Institute (NCI) defined a biomarker as, “A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition. Also called molecular marker and signature molecule.” Biomarkers may include molecular, histologic, radiographic, digital, or physiologic types. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker

For many years, programmers have routinely worked with biomarkers. For example, in diabetes studies, HbA1c is used to monitor blood sugar control; in cardiovascular studies, cholesterol levels or cardiac enzymes are used to assess heart health; and in infectious disease studies, viral load is used to track infection and treatment response. Although these measurable parameters are often treated as routine variables, they are fundamentally biomarkers because they reflect underlying biological processes. Similarly, routinely collected vital signs such as blood pressure and heart rate can also serve as biomarkers of physiological status. Biomarkers can be measured from various biological samples

such as blood, tissue, urine, saliva, sputum, feces, or semen, each revealing unique molecular or cellular information relevant to diagnosis, prognosis, and therapy selection.

In recent years, pharmaceutical companies have been developing a new generation of biomarkers aiming to reduce treatment failure rates and increase drug efficacy, especially for Oncology trials. In oncology, where cancer remains one of the most complex and heterogeneous diseases, biomarkers serve as key indicators that help decode tumor biology, disease progression, and treatment response. Biomarkers in cancer can include gene alterations, protein changes, or immune system modifications, providing highly relevant information about the disease. Biomarker data can be highly diverse, including genomics, proteomics, various imaging approaches, blood differentials, and more.

Clinical role of Biomarkers across the Cancer Disease Process

Figure 1.

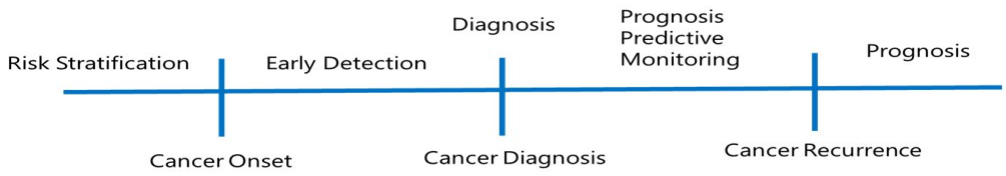


Figure 1 shows how biomarkers are used at different stages of cancer. Each type of biomarker has a specific purpose. Risk stratification biomarkers are used to identify people who may have a higher chance of developing cancer before any symptoms appear. Individuals with a strong family history or specific genetic variants may be classified as higher risk or low risk for developing cancer. Early detection biomarkers are used to identify cancer at an early stage, before it is formally diagnosed. Breast screening can detect early signs of breast cancer before symptoms appear.

In clinical trials, biomarker analyses are mainly used during and after treatment to help understand treatment response, disease progression, and long-term outcomes. This paper focuses on predictive and prognostic biomarkers.

BIOMARKER ANALYSIS

In clinical trials, biomarkers are classified into primary, secondary, or exploratory objectives. The primary and secondary objectives are documented in the Clinical Study Report (CSR). Exploratory biomarker analyses are conducted to generate new hypotheses and enhance understanding of how a drug works across different patient subgroups, focusing on various biomarker types. Biomarker analyses performed for exploratory objectives do not influence drug approval; their results are considered supportive. In the CSR, assessments should be performed for all trial participants. However, biomarker analyses may not be evaluated in all subjects, as it can be optional or complex technology does not allow for all subjects to be analyzed.

Programmer Workflow for Biomarker Analysis

As programmers, we always start any study task by looking for study documents such as the Statistical Analysis Plan (SAP), Protocol, Case Report Form (CRF), Standard Data Tabulation Model (SDTM), and Analysis Data Model (ADaM) datasets. It should be noted that the CSR presents information in a structured format, but exploratory biomarker analyses typically do not follow standard formats.

The process for biomarker analysis starts with a Separate Analysis Plan; each biomarker will have its own separate IAP. This plan outlines specific objectives for biomarkers, statistical methods, required datasets, and pathways to reference documents. Since biomarker data often differs in structure from standard clinical trial data frequently collected

from various vendors, interpretation can be challenging without adequate documentation. Therefore, it is essential to consult the Data Transfer Specification (DTS). This Excel-based document provides variable names, definitions, and contextual information that must be referenced when analyzing biomarker datasets.

All biomarker and genetic analyses are performed only after obtaining explicit consent from the patients. This consent specifically cover for example DNA analysis, sample storage, and the use of data for future exploratory research. Please note that, not all biomarker information is captured on the CRF due to its volume and complexity.

In the most recent version of the SDTM Implementation Guide (SDTMIG 3.4), several domains are available for the mapping of biomarker data, including Laboratory (LB), Cell Phenotype (CP), Biospecimen Findings (BS), Microbiology Specimen (MB), Genomics Findings (GF), Microbiology Susceptibility (MS), and Microscopic Findings (MI). However, it is important to note that there are no dedicated SDTM domains designated explicitly for each type of biomarker data.

For hypothesis generation, creating formal CDISC ADaM datasets is not necessary; instead, studies typically develop a pseudo-analysis dataset that consolidates all required information to generate statistical outputs, such as Tables, Listings, and Figures (TLFs).

This paper lists only the biomarkers to which we were exposed, as multiple biomarker categories were evaluated independently, each representing a complex biological measure, including PD-L1 expression, immune cell infiltration, genomic alterations, molecular/soluble protein, and pharmacogenetics (PGt).

Scenario 1: A Phase II Randomized, Open-Label Study of a Study Drug Versus Chemotherapy in Locally Advanced or Metastatic Urothelial Carcinoma After First-Line Platinum-Based Treatment

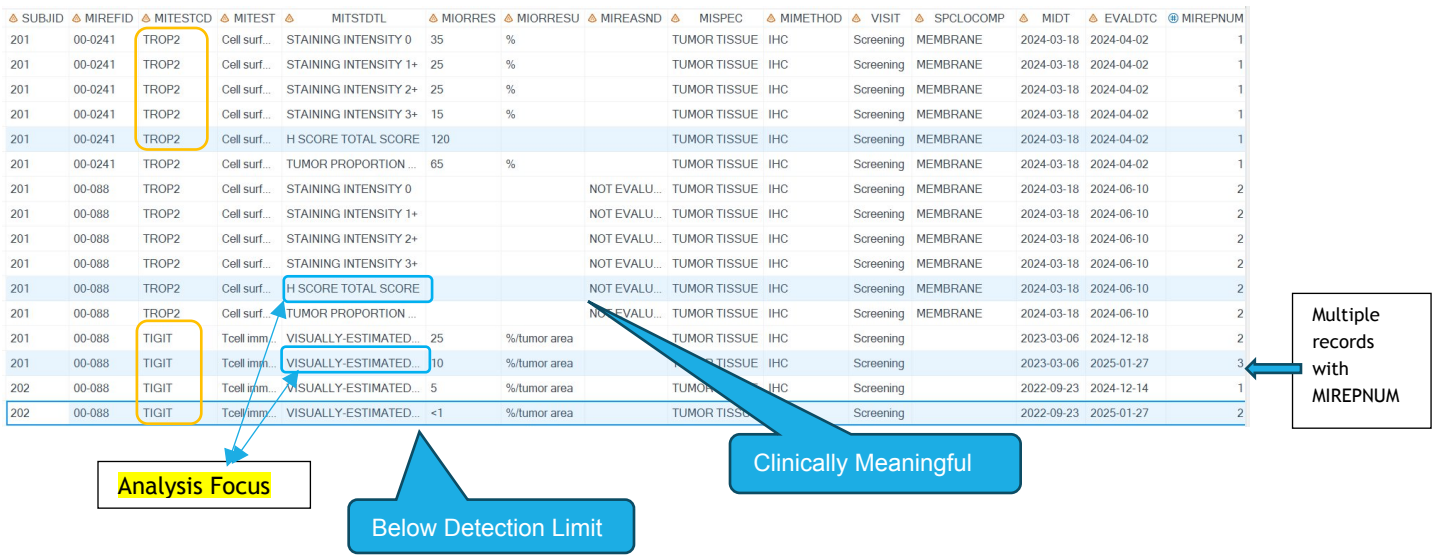
Table 1.

Biomarkers	Key Details
TROP-2(Trophoblast cell-surface antigen 2)	It is a cell-surface protein that is frequently overexpressed in tumor cells. It contributes to their growth, survival, and spread.
TIGIT (T-cell Immunoreceptor with Ig and ITIM domains)	TIGIT is an immune checkpoint receptor on immune cells that suppresses immune responses, allowing cancer to evade the immune system.
Cytokine Profile	It is a set of measurements of different cytokines in a patient's sample (usually blood) to understand how the immune system is behaving.

Biomarker *Raw Data*

Data *1*

SUBJID	MIREFID	MITESTCD	MITEST	MITSTDTL	MIORRES	MIORRESU	MIREASND	MISPEC	MIMETHOD	VISIT	SPCLOCOMP	MIDT	EVALDTC	MIREPNUM
201	00-0241	TROP2	Cell surf...	STAINING INTENSITY 0	35	%		TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-0241	TROP2	Cell surf...	STAINING INTENSITY 1+	25	%		TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-0241	TROP2	Cell surf...	STAINING INTENSITY 2+	25	%		TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-0241	TROP2	Cell surf...	STAINING INTENSITY 3+	15	%		TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-0241	TROP2	Cell surf...	H SCORE TOTAL SCORE	120			TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-0241	TROP2	Cell surf...	TUMOR PROPORTION ...	65	%		TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-04-02	1
201	00-088	TROP2	Cell surf...	STAINING INTENSITY 0			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TROP2	Cell surf...	STAINING INTENSITY 1+			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TROP2	Cell surf...	STAINING INTENSITY 2+			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TROP2	Cell surf...	STAINING INTENSITY 3+			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TROP2	Cell surf...	H SCORE TOTAL SCORE			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TROP2	Cell surf...	TUMOR PROPORTION ...			NOT EVALU...	TUMOR TISSUE	IHC	Screening	MEMBRANE	2024-03-18	2024-06-10	2
201	00-088	TIGIT	Tcell imm...	VISUALLY-ESTIMATED...	25	%/tumor area		TUMOR TISSUE	IHC	Screening		2023-03-06	2024-12-18	2
201	00-088	TIGIT	Tcell imm...	VISUALLY-ESTIMATED...	10	%/tumor area		TUMOR TISSUE	IHC	Screening		2023-03-06	2025-01-27	3
202	00-088	TIGIT	Tcell imm...	VISUALLY-ESTIMATED...	5	%/tumor area		TUMOR TISSUE	IHC	Screening		2022-09-23	2024-12-14	1
202	00-088	TIGIT	Tcell imm...	VISUALLY-ESTIMATED...	<1	%/tumor area		TUMOR TISSUE	IHC	Screening		2022-09-23	2025-01-27	2



Data 2

SUBJID	DOMAIN	LBTESTCD	LBTEST	LBCAT	LBSTRESC	LBSTRESU	LBSPEC	LBBLFL	LBTOXGR	LBLLQ	VISIT	LBDTC	LBSTRESN	LBNRLO
201	LB	IFNG	Interferon Gamma	SOLUBLE FACTOR	3.68	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	3.68	.
201	LB	INTLK10	Interleukin 10	SOLUBLE FACTOR	1.01	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	1.01	.
201	LB	INTLK12	Interleukin 12	SOLUBLE FACTOR	<2.44	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK13	Interleukin 13	SOLUBLE FACTOR	<8.42	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK1B	Interleukin 1 Beta	SOLUBLE FACTOR	<1.29	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK2	Interleukin 2	SOLUBLE FACTOR	<1.78	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK4	Interleukin 4	SOLUBLE FACTOR	<0.44	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK6	Interleukin 6	SOLUBLE FACTOR	<1.27	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	.	.
201	LB	INTLK8	Interleukin 8	SOLUBLE FACTOR	19.00	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	19	.
201	LB	TNF	Tumor Necrosis Factor	SOLUBLE FACTOR	2.37	ng/L	SERUM	Y			WEEK 1 DAY 1	2023-08-16...	2.37	.

Data 1 presents the immunohistochemistry (IHC) based expression of the biomarkers TIGIT and TROP-2, assessed on baseline tumor tissue samples and generated by an external pathology vendor. IHC and pathology-derived biomarker data are typically represented in the SDTM Microscopic Findings (MI) domain.

Clinically, TROP-2 expression is summarized using the H-score total score, which is derived from the percentage of tumor cells at each staining intensity level (0, 1+, 2+, 3+). These staining intensity percentages typically sum to 100%. The H-score is calculated as a weighted composite score (e.g., 35×0 + 25×1 + 25×2 + 15×3 = 120); therefore, it is unitless and MIORRESU is not populated. TIGIT expression is reported as the visually estimated combined positive score (vCPS/CPS), reflecting the percentage of positive immune cells.

For analysis, only records representing the clinically relevant summary measures (MISTDTL = H-score for TROP-2 and MISTDTL = vCPS for TIGIT) were considered. Records with missing numeric results, results below the assay's lower limit of detection, and multiple assay runs were reviewed in consultation with the study statistician and subject matter experts (SMEs).

In contrast to standard practice where missing results are excluded, SMEs recommended retaining the most recent pathology assessment when a numeric value was not reported but the evaluation was considered clinically interpretable (e.g., due to visible tumor features or staining interpretation constraints). Accordingly, for each subject, the most recent record based on assessment date (MIDT/EVALDTC) was selected for analysis.

For values reported as below the lower limit of detection (e.g., <1), the numeric component of the reported value was used for analysis rather than treating the result as missing or zero. This approach was chosen to preserve low-level biological signal and was implemented based on SME guidance.

Data 2 presents cytokine profile data, consisting of quantitative measurements of multiple soluble immune markers that reflect different aspects of immune activity. Cytokines including IFNG, IL10, IL12, IL13, IL1B, IL2, IL4, IL6, IL8, and TNF were measured in serum samples. Each record represents a single cytokine measurement for a subject at a specific visit.

Cytokine data are collected longitudinally, with the same markers assessed at multiple timepoints to evaluate changes in immune response over the course of treatment. In SDTM, these data are typically represented in the Laboratory Test Results (LB) domain, where repeated measurements across visits and timepoints are expected.

It is critical to confirm that the visit definitions specified in the Separate Analysis Plan align with the visits recorded in the dataset. During data review, discrepancies were identified in which visits were initially mapped based on the general laboratory visit structure rather than the biomarker-specific visit schedule. These discrepancies were subsequently resolved in collaboration with the biomarker data management team.

Results reported as below the assay's lower limit of detection were handled according to predefined analysis rules to ensure consistency and preservation of low-level biological signal.

Scenario 2. A Phase III trial comparing a PD-L1 inhibitor (Study Drug) with chemotherapy in PD-L1–positive advanced non-small cell lung cancer (NSCLC).

Table 2.

Biomarkers	Key Details
TMB (Tumor Mutational Burden)	Measures the number of mutations present in the tumor's DNA. These mutations can create neoantigens, which help the immune system recognize cancer cells. Non-synonymous mutations are counted because they change proteins and can influence the immune response.
Genomic SNP (Single Nucleotide Polymorphism)	It is a one-letter change in DNA at a specific position in the genome. Germline DNA is inherited from your parents and is present in all cells of your body.

Biomarker Raw Data

Data 3

Analysis Focus

USUBJID	PFTESTCD	PFCAT	PFSCAT	PFRUNID	VISIT	PFSTRESC	PFSTDTL	PFDTc	PFSEQ	VISITDY	PFDY
201	INDEL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	4.88	Somatic_variants_per_Mb	2016-12-20	2	-28	-42
201	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	589	Non_synonymous_Somatic_Variants	2016-12-20	3	-28	-42
201	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	8.66	Non_synonymous_Somatic_Variants_per_Mb	2016-12-20	4	-28	-42
201	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	4785	Somatic_variants	2016-12-20	5	-28	-42
201	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	70.37	Somatic_variants_per_Mb	2016-12-20	6	-28	-42
201	TOTAL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	5117	Somatic_variants	2016-12-20	7	-28	-42
201	TOTAL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	75.25	Somatic_variants_per_Mb	2016-12-20	8	-28	-42
202	INDEL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	5.22	Somatic_variants_per_Mb	2018-05-23	2	-28	-50
202	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	674	Non_synonymous_Somatic_Variants	2018-05-23	3	-28	-50
202	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	9.91	Non_synonymous_Somatic_Variants_per_Mb	2018-05-23	4	-28	-50
202	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	4913	Somatic_variants	2018-05-23	5	-28	-50
202	SNV	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	72.25	Somatic_variants_per_Mb	2018-05-23	6	-28	-50
202	TOTAL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	5268	Somatic_variants	2018-05-23	7	-28	-50
202	TOTAL	GENETIC VARIATION	TUMOR MUTATIONAL BURDEN	tumor-only	SCREENING	77.47	Somatic_variants_per_Mb	2018-05-23	8	-28	-50

Data 4

Analysis Focus

SUBJECT_ID	TREATMENT_ARM	SPEC_TYPE	VARIANT_TYPE	GENE_SYMBOL	MEASUREMENT	RESULT	UNIT	REASON_ND	reference_id
201	CHEMO	DNA	SNP	IL6	Allele_Zygosity	Homozygous			rd1345
201	CHEMO	DNA	SNP	IL6	Base_Call	G/G			rd1345
201	CHEMO	DNA	SNP	FCGR2A	Allele_Zygosity	Homozygous			rd7738
201	CHEMO	DNA	SNP	FCGR2A	Base_Call	A/A			rd7738
201	CHEMO	DNA	SNP	FCGR3A	Allele_Zygosity	Heterozygous			rd3837
201	CHEMO	DNA	SNP	FCGR3A	Base_Call	C/A			rd3837
202	INVEST	DNA	SNP	IL6	Allele_Zygosity	Homozygous			rd1345
202	INVEST	DNA	SNP	IL6	Base_Call	G/G			rd1345
202	INVEST	DNA	SNP	FCGR2A	Allele_Zygosity	Heterozygous			rd7738
202	INVEST	DNA	SNP	FCGR2A	Base_Call	G/A			rd7738
202	INVEST	DNA	SNP	FCGR3A	Allele_Zygosity	Homozygous			rd3837
202	INVEST	DNA	SNP	FCGR3A	Base_Call	C/C			rd3837

Data 3 presents Tumor Mutational Burden (TMB) data derived from next-generation sequencing and received from the bioinformatics team, as these data were not captured in the Case Report Form (CRF). In this study, TMB data were initially mapped to the PF domain. According to SDTM version 3.4 and later, genomic biomarkers such as TMB are more appropriately represented in the Genomic Findings (GF) domain.

The study analysis focuses specifically on the PFSTDTL descriptor non_synonymous_somatic_variants_mb, which represents the number of non-synonymous somatic mutations per megabase (mut/Mb), and this variant changes the

protein structure. This value is generated through specialized bioinformatics pipelines and therefore is not recalculated or imputed during clinical programming.

Data 4 This presents Single Nucleotide Polymorphism (SNP) data. Genomic testing captures numerous SNPs. A single SNP is a genetic variation or mutation, and hundreds or thousands are usually generated per biosample and supplied by an external vendor. Hence, it is considered non-crf data.

All three genes (IL6, FCGR2A, and FCGR3A) provide different insights into the patient's immune system. However, in this trial, our focus is on FCGR2A and FCGR3A because these genes are linked to how effectively immune cells can kill tumor cells through antibody-dependent cell-mediated cytotoxicity (ADCC) or we are interested in the SNPs which are biologically relevant to the drug mechanism.

From a programming standpoint, the SNP field Base_Call already indicates the patient's genotype at each gene.
Base_Call: A/A means both alleles are the same, so the Allele_Zygosity is Homozygous.
Base_Call: C/A indicates two different alleles, so the Allele_Zygosity is Heterozygous.

Table 3

Column Name	Label (" ") / Description	Data Type (max. length)*	Examples / Options	Mandatory
STUDY_ID	"Study ID" / As specified in Shipping Manifest	Char (20)	e.g FSP_ZZZ	☒
SUBJECT_ID	"Subject ID" / As specified in Shipping Manifest	Char (10)	e.g. 201	☒
VISIT	"Visit Name" / As specified in Shipping Manifest	Char (50)	e.g. W1D1	☒
COLL_DATE	"Date of Specimen Collection" / As specified in Shipping Manifest	Char (15)	e.g. 21SEP2016	☒
COLLECTSMP_ID	"Collected Sample Identifier" / Accession number as specified in the electronic Case Report Form	Char (30)	e.g. S5550903	☒
ASSAYSMP_ID	"Assayed Sample Identifier" / Assay sample ID (label by vendor).	Char (30)	e.g. 670740	☒
SPEC_TYPE	"Specimen Type" / This field captures the type of specimen assayed.	Char (50)	DNA	☒
VENDOR	"Vendor Name"	Char (100)	XX	☒
METHOD	"Test Method"	Char (200)	REAL-TIME POLYMERASE CHAIN REACTION ASSAY	☒
VARIANT_TYPE	"Variant Type"	Char (8)	SNP	☒
REFERENCE_DB	"Reference Database" / This field captures the name and version (if applicable) of the external database from which the reference will be provided.	Char (50)	NCBI dbSNP Build 153	☒
REFERENCE_ID	"Reference Identifier"	Char (50)	e.g. rs396991 See Table 3	☒
GENE_SYMBOL	"Official Gene Symbol"	Char (25)	e.g. FCGR3A	☐

Table 3 presents Data Transfer specifications/agreement (DTS/DTA) file. The external vendor supplies this file in a specific format. This document aims to guide the organization on the structure and format of Single Nucleotide Polymorphisms (SNP) related data transferred to the pharmaceutical companies. External vendors supporting study specific data need to perform specialized testing.

Processed Biomarker Data

Data 5

	USUBJID	BORIERC	SNPFL	SEX	SMOKGR	FASFL	TRT01P	PDL1ST1C	TMBGR1	FCGR3A	FCGR2A	OS	PFSIE	TMB	CENOS	CENPFSIE
1	201	PR	Y	M	Ever Smoker	Y	INVEST	High (80-100%)	>8.95-<=10.80	Homozygous A/A	Homozygous A/A	22.5	12.6	8.97	0	0
2	202	SD	Y	M	Ever Smoker	Y	INVEST	Low (1-<50%)	>7.05-<=8.95	Heterozygous C/A	Homozygous A/A	11.6	2.8	8.02	0	0
3	203	PD	Y	M	Ever Smoker	Y	INVEST	Low (1-<50%)	<=7.05	Homozygous C/C	Homozygous A/A	15.2	1.2	6.01	0	0
4	204	PR	Y	M	Ever Smoker	Y	INVEST	Low (1-<50%)	>10.80	Homozygous C/C	Heterozygous A/G	14.1	14.1	14.52	0	0
5	205	SD	Y	M	Ever Smoker	Y	INVEST	Low (1-<50%)	>8.95-<=10.80	Homozygous A/A	Homozygous A/A	26.8	5.6	9.27	1	1
6	206	SD	Y	M	Ever Smoker	Y	CHEMO	Low (1-<50%)	<=7.05	Homozygous A/A	Homozygous A/G	8.5	3.8	4.94	1	1
7	207	PD	Y	F	Ever Smoker	Y	INVEST	Moderate (50-<80%)	>7.05-<=8.95	Homozygous A/A	Homozygous G/G	24.1	1.5	7.64	0	0
8	208	PR	Y	M	Ever Smoker	Y	INVEST	High (80-100%)	>10.80	Heterozygous C/A	Homozygous A/A	9.7	4.6	11.82	0	0
9	209	NE	Y	F	Ever Smoker	Y	CHEMO	Low (1-<50%)	>7.05-<=8.95	Homozygous A/A	Homozygous A/A	15.2	1.2	7.07	1	1

Find quantiles for TMB and assign
"<= {TMBQ1}"
"> {TMBQ1} - <= {TMBMED}"
"> {TMBMED} - <= {TMBQ3}"
"> {TMBQ3}"

Data 5 This pseudo-ADaM–like dataset is created for biomarker analysis and serves as an exploratory biomarker for hypothesis generation. Because the data do not follow standard CDISC ADaM structures, they are fit for purpose and support statistical analysis.
To ensure consistency across vendors, most harmonization is handled by the bioinformatics and data standards teams, using either client-specific conventions or CDISC standards during SDTM mapping.
During biomarker dataset preparation, programmers play a key role in identifying data issues, including missing or incomplete values, invalid categories, duplicate or multiple records, incorrect sample sources, and visit or evaluation dates. Any discrepancies are reviewed with data management, Subject Matter Experts (SMEs), and statisticians to ensure data quality and scientific interpretability.

This table outlines key programming checks and data-handling rules applied to cytokine, TMB, and SNP biomarker data to ensure scientific validity and consistency before analysis.

Programming Checks	Cytokines	TMB	SNP
Data type	Continuous, longitudinal	Continuous, summary measure	Categorical (genotype)
Units	Must be biologically valid and consistent per DTS	Must be reported as mutations per megabase (mut/Mb)	Not applicable
Value range	Valid range defined in DTS	Valid range defined in DTS	Not numeric; no ranges
Below/Above limits	Handle as per SME/statistician guidance	Handle as per SME/statistician guidance	Not applicable
Missing data	Handle as per SME/statistician guidance	Keep missing; no derivation or imputation	Do not impute (fixed germline state)
Categorization	Not categorized	Categorized (e.g., quartiles, High/Low)	Fixed genotype categories
Visit structure	Visits must match IAP, not	Usually baseline-only	Usually baseline-only

	standard lab visits		
Data consistency checks	Unit consistency and visit alignment	Ensure mut/Mb (not raw counts)	Standard genotype notation (e.g., A/C = C/A)
Controlled terminology	As defined in DTS	As defined in DTS	Must follow CDISC or sponsor-controlled terms
Multiple records	Handle as per SME/statistician guidance	Not applicable	Not applicable
Sample Source	Serum/plasma as defined	Tumor (or blood if specified)	Blood or tumor, per biomarker definition

STATISTICAL ANALYSIS

This section summarizes the analyses performed on biomarker data. These analyses support the expert teams in identifying biomarkers that may contribute to hypothesis generation and in understanding whether the Study drug shows differential effects across patient subgroups.

Key analysis included:

- Comparison of baseline characteristics between the biomarker-evaluable population and the full analysis set.
- Descriptive or Graphical summaries of the biomarker-evaluable population.
- Analyses of efficacy endpoints, including progression-free survival (PFS) and overall survival (OS), using Kaplan–Meier methods.
- Statistical modeling approaches to explore associations between biomarker status and clinical outcomes

Case 1. Distribution comparison of the Biomarker-evaluable Population versus the Full Analysis Set

As an initial step, baseline distributions were compared between the biomarker-evaluable populations and the Full Analysis Set (FAS). **Table a** compares baseline characteristics between the SNP biomarker population and the FAS, while **Table b** compares those between the TMB biomarker-evaluable population and the FAS. The FAS included 1,190 subjects, of whom 545 had SNP biomarker data available and 853 had Tumor Mutational Burden (TMB) data available.

Baseline demographic and disease characteristics, including sex and age categories (<65 and ≥65 years), were summarized and compared across populations. This methodological step was performed to confirm that the biomarker-evaluable subsets were broadly representative of the overall study population and to support appropriate interpretation of subsequent exploratory biomarker and efficacy analyses. Across both tables, a similarly higher proportion of male subjects than of female subjects, and a similarly greater representation of patients aged <65 years than of those aged ≥65 years, were observed.

Table a. **SNP Biomarker Population vs Full Analysis Set Overall**

	SNP Biomarker Population (N = 545)	FAS Set (N = 1190)
Analysis Age		
n	545	1190
Missing	0	0
Mean (SD)	61.8(9.42)	62.4(9.67)
Min;Max	29.0;87.0	24.0;88.0
Q1;Q3	55.0;69.0	56.0;70.0
Median	62.0	63.0
Pooled Analysis Age Group 1 (N)		
<65	296 (54.3%)	618 (51.9%)
>=65	249 (45.7%)	572 (48.1%)
Sex		
Male	401 (73.6%)	876 (73.6%)
Female	144 (26.4%)	314 (26.4%)

Table b. **TMB Biomarker Population and Full Analysis Set Overall**

	TMB Biomarker Population (N = 853)	FAS Set (N = 1190)
Analysis Age		
n	853	1190
Missing	0	0
Mean (SD)	63.4 (9.11)	63.9 (9.48)
Min; Max	27.0; 87.0	24.0; 88.0
Q1; Q3	57.0;70.0	57.0;71.0
Median	63.0	64.0
Pooled Analysis Age Group 1, n (%)		
<65	468 (54.9)	631 (53.0)
>=65	385 (45.1)	559 (47.0)
Sex, n (%)		
Male	636 (74.6)	896 (75.3)
Female	217 (25.4)	294 (24.7)

Higher proportion of Male compared to Female

Case 2. Graphical summary of Biomarker-evaluation Population

Graphical visualizations highlight data distribution, variability, and trends that tables alone may not clarify. Figures such as bar diagrams and box plots provide a visual understanding of the dataset without indicating biological or statistical conclusions; they simply show the distribution of the biomarker population.

Figure a. **Graphical Summary of FCGR3A by Treatment Arm- SNP Biomarker Population**

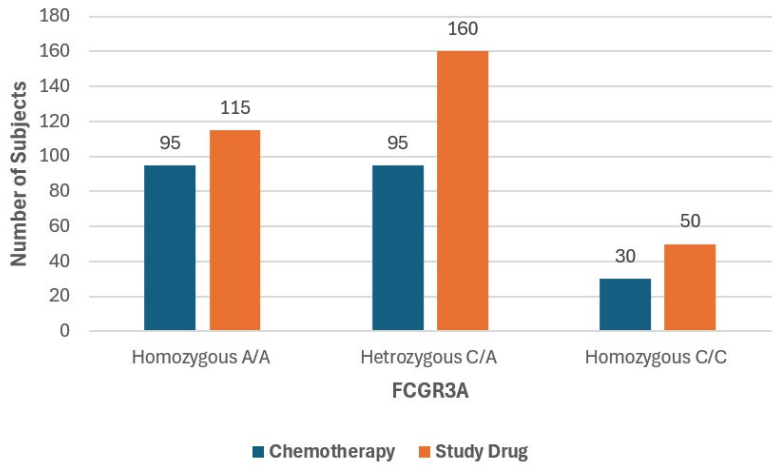


Figure a shows the distribution of FCGR3A SNP genotypes (A/A, C/A, and C/C) by treatment arm. The x-axis represents genotype categories, and the bars show the number of subjects in the Chemotherapy and Study groups. All genotype categories are present in both treatment arms.

When interpreted in the context of subgroup proportions rather than total patient counts, the relative distribution of genotypes appears broadly similar between the Chemotherapy and Study arms, with C/A being the most common genotype in both groups. This figure primarily shows subgroup sample sizes, which are important to consider before statistical modeling, as small subgroup sizes may result in unstable or imprecise estimates.

Figure b. **Boxplot of TMB by Treatment Arm- TMB Biomarker Population**

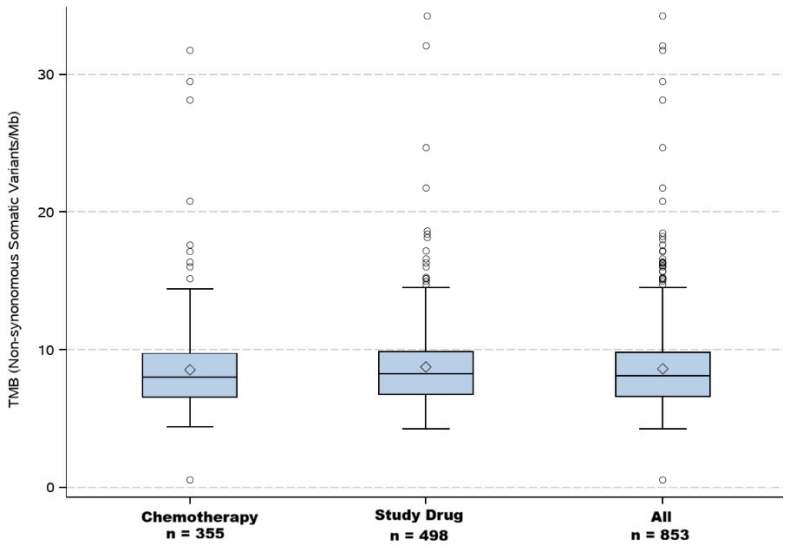


Figure B shows the distribution of Tumor Mutational Burden (TMB, non-synonymous somatic variants per Mb) by treatment arm (Chemotherapy, Study) and overall. The box plot summarizes the median, interquartile range, and spread of TMB values, with individual points indicating potential outliers.

This visualization is descriptive and intended to assess baseline comparability and variability of TMB across treatment arms, rather than to infer treatment effects.

Case 3. Descriptive Exploratory Analysis of Efficacy Endpoints

Table a. Summary of Progression Free Survival (IERC) by Treatment Group: SNP Biomarker Population - FCGR3A gene: Homozygous A/A

	Chemotherapy N=95	Study Drug N=115
Number of events (PD or Deaths), n (%)	62 (65.3)	85 (73.9)
PFS (IERC)		
Median (months)	6.5	7.8
Min, Max	0.2, 38.9	0.3, 55.6
95% CI	[5.4, 7.8]	[3.0, 10.9]
Number of subjects at risk/failed/PFS (IERC) rates (%) up to [95% confidence interval]		
6 months	42 / 36 / 56.2 [45.7,65.4]	54 / 45 / 60.8 [51.3,69.2]
12 months	18 / 54 / 30.4 [21.7,40.0]	28 / 63 / 38.7 [30.1,47.1]
18 months	11 / 60 / 22.1 [14.5,31.0]	18 /74 / 28.9 [21.2,36.9]
24 months	9 / 62 / 18.9 [11.8,27.4]	14 / 80 / 24.6 [17.7, 32.1]
30 months	7 / 62 / 16.0 [9.5, 24.3]	11 / 82 / 21.7 [15.1, 29.2]
36 months	5 / 62 / 14.1 [7.6, 22.7]	9 / 84 / 19.3 [13.2, 26.4]
42 months	3 / 62 / 11.6 [5.6, 19.9]	7 / 85 / 16.8 [11.1, 23.7]
48 months	2 / 62 / nd	5 / 85 / 14.6 [9.0, 21.5]
54 months	1 / 62 / nd	3 / 85 / 12.8 [7.5, 19.8]
60 months	0 / 62 / nd	2 / 85 / nd
66 months	0 / 62 / nd	0 / 85 / nd

Table a summarizes Progression-Free Survival (PFS, assessed by IERC) for the FCGR3A Homozygous A/A SNP subgroup, stratified by treatment group.

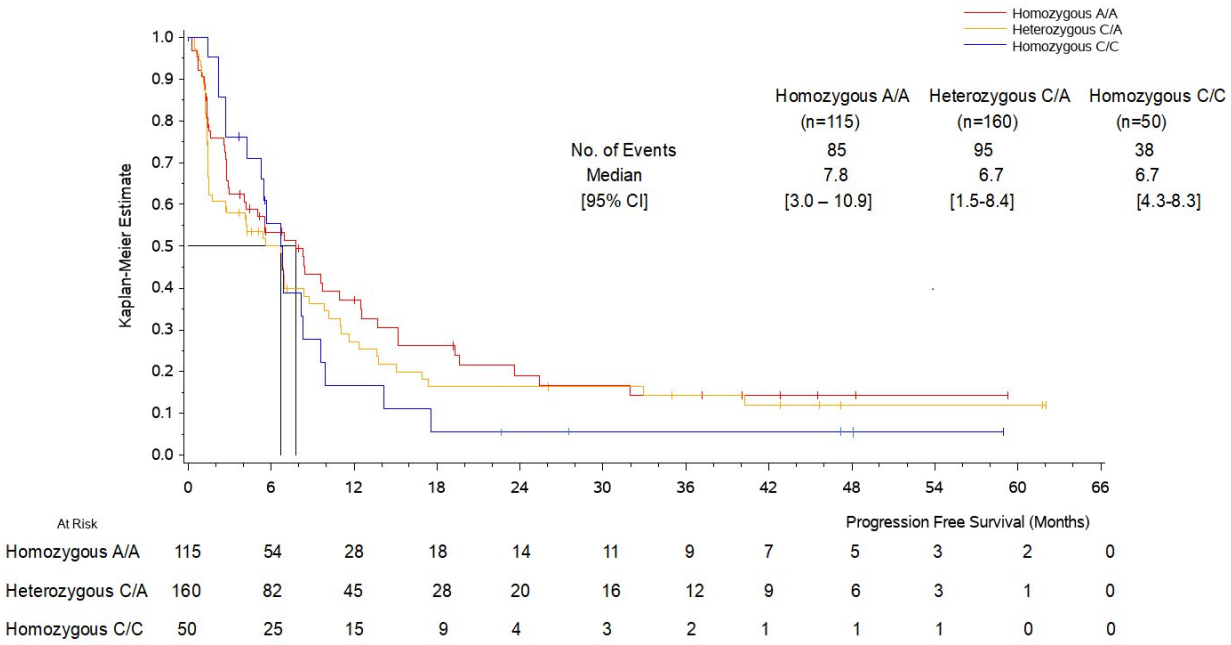
PFS was analyzed using the Kaplan–Meier method implemented via PROC LIFETEST, based on time-to-event data (PFS time) and an event indicator (progression or death). KM estimated PFS rates at prespecified timepoints (e.g., 6, 12, and 24 months).

Event counts and survival probabilities are cumulative over time. Censored subjects are considered in the calculation of survival probability up to the time they are censored. A log-rank test was used to assess differences in time-to-event between the chemotherapy and Study arms.

In this genotype subgroup, the Study arm had a longer median PFS (7.8 months) than the chemotherapy arm (6.5 months). The maximum observed PFS time (38.9 or 55.6 months) reflects the longest follow-up and does not represent the median.

After 6 months, the number of subjects at risk decreases due to events and censoring, which results in wider confidence intervals and some PFS estimates reported as “nd.” The same analytical approach and summary structure were applied to the other FCGR3A genotype subgroups (Heterozygous C/A and Homozygous C/C).

Figure a. **Kaplan-Meier estimates of Progression Free Survival (IERC) by FCGR3A gene- SNP Biomarker Population- Treatment: Study Drug**



At Risk												
Homozygous A/A	115	54	28	18	14	11	9	7	5	3	2	0
Heterozygous C/A	160	82	45	28	20	16	12	9	6	3	1	0
Homozygous C/C	50	25	15	9	4	3	2	1	1	1	0	0

Figure a displays progression-free survival (PFS), defined as the time a subject remains progression-free or alive. Time (months) is shown on the x-axis and the estimated event-free probability on the y-axis. The plot is generated using PROC LIFETEST, using the same time-to-event and censoring data as in the PFS summary table.

From a programming and QC perspective, the median PFS corresponds to the time at which each curve crosses the 50% survival probability. In the Study arm, the Homozygous A/A genotype crosses the 50% line at approximately 7–8 months, consistent with the median PFS reported in the summary table.

Although the A/A curve shows a steeper early decline, it remains above the C/A and C/C curves around the median time window, resulting in a longer median PFS. The number-at-risk table shows a decrease over time due to events and censoring. Overall, the KM plot visually supports and validates the PFS summary results across FCGR3A genotypes in the Study arm.

Case 4: Statistical Modeling of Biomarker Associations with PFS

In Case 3, survival analysis was used to assess the effect of treatment on survival and to assess whether the FCGR3A SNP biomarker is associated with survival or modifies the treatment effect. A Cox proportional hazards model (PROC PHREG) was applied to estimate the effects of treatment and genotype on survival. The analysis evaluated whether FCGR3A is prognostic, meaning it is associated with survival regardless of treatment, or predictive, meaning the treatment effect differs across genotype subgroups.

Table a **Cox Regression of PFS(IERC) SNP Biomarker Population – Model with treatment and FCGR3A Gene**

Model 1 evaluates the main effects of treatment and genotype on PFS. Both **treatment and genotype** are included as covariates.

Model 2 includes **treatment, genotype, and their interaction term** (treatment × genotype). The interaction term tests whether the effect of treatment differs by genotype.

Sometimes, other factors (age, smoking status) influence survival rates. To observe the effect of an additional variable along with treatment, an interaction term is used

	Effect			Hazard Ratio
	Factor	p-value	Comparison	(HR) [95% CI]
Model1: Treatment + FCGR3A	FCGR3A	0.184	C/A vs A/A	1.22 [0.91–1.63]
			C/C vs A/A	1.15 [0.74–1.78]
	Treatment Group	0.231	Investi vs Chemo	0.88 [0.71–1.11]
Model2: Treatment + FCGR3A + Treatment * FCGR3A	FCGR3A	0.612	—	—
	Treatment Group	0.744	—	—
	Treatment Group by FCGR3A	0.802	Investi vs Chemo (A/A)	0.93 [0.67–1.30]
			Investi vs Chemo (C/A)	0.81 [0.56–1.16]
			Investi vs Chemo (C/C)	1.09 [0.59–2.00]

A hazard ratio (HR) greater than 1 indicates a higher risk of the event in the Study treatment group compared with chemotherapy, whereas an HR less than 1 indicates a lower risk; an HR equal to 1 indicates no difference between treatments.

Table a summarizes results from two Cox models assessing progression-free survival (PFS).

Model 1 examines the overall effects of treatment and FCGR3A genotype. The hazard ratio (HR) for Study treatment versus chemotherapy was 0.88, indicating that patients on Study treatment had a numerically lower risk of progression (12%) compared with those on chemotherapy. The HRs for FCGR3A genotype were close to 1, indicating that patients with different genotypes had similar PFS outcomes in this dataset. Overall, these exploratory results suggest that FCGR3A genotype alone does not have a strong prognostic impact on PFS.

Model 2 evaluates whether the treatment effect varies by FCGR3A genotype. The treatment HRs were 0.93 (A/A), 0.81 (C/A), and 1.09 (C/C), and all confidence intervals included 1. These similar estimates across genotype groups indicate that the treatment effect is generally consistent across genotypes, with no clear evidence of a genotype-specific treatment benefit. In this exploratory analysis, FCGR3A genotype does not appear to meaningfully influence how patients respond to treatment.

This is an exploratory, hypothesis-generating analysis, interpretations are based on the size and consistency of the estimated effects rather than on p-values.

Note: Subjects with missing results will not appear in summaries, boxplots, survival models, or association tests. However, they are included in listings and subject-level summaries.

Real-world Example

Why collecting biomarker data matters: Biomarkers can change how trial results are understood. This is a real-world example of a drug approved following biomarker analysis.

Therapeutic Area / Drug	Biomarker	Without Biomarker Stratification	With Biomarker Stratification
Breast Cancer – Trastuzumab (Herceptin)	HER2 overexpression	In all breast cancer patients, the effect was weak or insignificant.	In HER2+ patients (~20% population), the benefit was strong. A drug approved specifically for HER2-high patients.

CONCLUSION

Biomarker programming requires far more than technical coding skills. It also requires understanding the biological context, analysis objectives, data sources, vendor differences, and the expectations of statisticians and SMEs. Dataset structures vary widely across vendors, so assumptions based on previous studies should be avoided. Programmers must invest time in understanding the scientific meaning of each biomarker and reviewing the Separate Analysis Plan before coding. Experience with one biomarker does not guarantee expertise with others, as the same marker can be interpreted differently across studies. Overall, success in biomarker analysis depends on technical skill, scientific awareness, and proactive cross-functional communication.

REFERENCES

1. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker
2. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ncbi.nlm.nih.gov/books/NBK326791
3. https://library.cdisc.org/

ACKNOWLEDGEMENT

Sincerely, I would like to thank Ashik Chowdhury, Aslihan Gerhold-Ay, Andreas Kloetgen, and Cytel colleagues for their valuable feedback on the drafts and for patiently answering all my queries.

CONTACT INFORMATION

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

aboli.katdare@cytel.com

Aboli Tilve | LinkedIn

Cytel Statistical Software & Services Pvt. Ltd., Pune, India