GeneBench-Pro Case Study: SV-driven TXR1 Tumor-board Utility Estimation

GeneBench-Pro

June 26, 2026

1 Overview

Molecular tumor boards increasingly combine sequencing, expression, treatment-history, and toxicity evidence when deciding whether a patient should receive targeted therapy. This case study turns that workflow into a synthetic registry problem in which the clinically actionable subgroup is not directly labeled. The visible registry contains 2400 advanced solid-tumor cases, 2210 time-zero analysis-set cases, and seven released gzip-compressed tables. The requested response has four fields: a binary recommendation for a TXR1 inhibitor, the week-16 clinical-benefit risk difference in percentage points, the 8-week treatment-limiting toxicity/discontinuation risk under TXR1i, and the net clinical utility. The main failure modes are raw TXR1 expression, raw alternate-fraction or split-read support without purity/copy-number single-copy CCF correction, long-read artifact/proximity shortcuts, ASE or phase support used alone, unweighted target contrasts instead of IPTW/IPCW target-trial estimation, and hard DLR1 PGx calls that ignore call confidence.

For a reader outside oncology genomics, the practical setting is a molecular tumor board: a clinical review group that combines genomic findings, treatment history, and patient risk to choose targeted therapy. TXR1 is a fictional target gene, TXR1i is a fictional inhibitor of that target, and the registry records patients who might receive it. The drug should help only when the tumor has TXR1 *target-mediated activation*: a structural DNA change has put TXR1 into an activated state that drives tumor behavior. The difficulty is that the files never reveal that state directly, so the analysis must reconstruct it from imperfect sequencing and expression summaries before estimating whether treated target patients did better than comparable untreated target patients.

There are two kinds of reasoning in the task. The first is molecular: decide which tumors truly have an actionable TXR1 event. The second is causal: because this is an observational registry rather than a randomized trial, treated and untreated patients are not automatically comparable. The correct answer therefore combines biomarker interpretation with weighting methods that emulate a target trial.

Reader orientation and acronym guide

Term	Plain-language meaning in this report
TXR1 / TXR1i	A synthetic cancer target gene and its inhibitor. The report treats them as fictional, so no external TXR1 biology is needed.
Molecular tumor board	A clinical review setting where genomic findings, treatment history, and patient risk are combined to choose targeted therapy.
Structural variant (SV)	A rearrangement of DNA. Here, the important pattern is a promoter-facing rearrangement near TXR1 that could turn the gene on.
`P2T` topology	"Promoter to target": the released code for a junction oriented as though a promoter/control element points toward TXR1. It is necessary but not sufficient.
Long-read sequencing	Sequencing reads long DNA molecules, which helps resolve rearrangements but still suffers from mapping, repeat, and support artifacts.
Purity and copy number	Tumor samples contain both tumor and normal cells, and tumor cells can carry extra DNA copies. Raw read fractions must be corrected for both.
CCF	Cancer cell fraction: the fraction of tumor cells carrying the structural event. In this synthetic single-copy SV setting, the purity/copy proxy equals CCF; in amplified real tumors, alteration multiplicity and major copy number would also matter.
Expression residual	The part of TXR1 expression left after subtracting lineage, RNA quality, library size, and batch effects. It asks whether TXR1 is high for the right reason.
ASE and phase	Allele-specific expression and read phasing. Together they ask whether the rearranged DNA copy is also the expressed TXR1 copy.
DLR1 PGx	A synthetic pharmacogenomic toxicity marker. It represents inherited drug-handling risk rather than tumor sensitivity.
Target trial	A way to phrase an observational analysis as if it were the randomized trial we wish had been run.
IPTW / IPCW	Inverse-probability weights. IPTW balances treated versus control patients; IPCW corrects for informative missing week-16 assessment.

The correct solution uses three named stages. Stage 1, target reconstruction, infers TXR1 target-mediated activation from long-read topology, mapping QC, copy-state and purity correction, expression residualization, allele-specific expression, and phase support. Stage 2, target-trial benefit estimation, estimates the treated-versus-control contrast inside that inferred target population while handling treatment-history confounding and informative week-16 assessment. Stage 3, PGx-standardized toxicity and net utility, standardizes TXR1i toxicity back to the same target population while accounting for DLR1 pharmacogenomic risk. The grading contract and realized answer values are listed in the answer-field table after the released files. The target set recovered from the released data contains 354 cases, not the > 1000 raw-expression shortcut set or the roughly 400–700 support/proximity shortcut sets.

2 Released Prompt and Files

Prompt

```
A molecular tumor board registry contains trial-eligible advanced solid-tumor cases considered
for a TXR1-directed inhibitor. Estimate, for tumors with SV-driven TXR1 target-mediated
activation at time zero, the marginal effect of TXR1i versus non-TXR1 systemic therapy on
week-16 clinical benefit as if all patients had an assessable week-16 visit. Also estimate the
8-week treatment-limiting toxicity/discontinuation risk under TXR1i in the same target
population. Report net clinical utility = benefit risk difference (percentage points) - 0.35 *
toxicity risk (percentage points), and choose therapy_class_code 1 if TXR1i has positive net
utility and 0 otherwise.

Use percentage-point units for all non-code quantities. Positive benefit means TXR1i improves
week-16 clinical benefit relative to non-TXR1 systemic therapy.

These data came from a real experiment; you will be graded not just on numerical correctness but
the quality of analytical reasoning you exhibit; do not attempt to take any shortcuts.

Return your final answer as exactly one JSON object.
Do not wrap the JSON in markdown.
Do not add prose before or after the JSON.
Do not omit any keys shown in the example.
Return the JSON object in your final answer:
 {
  "answer": {
    "therapy_class_code": <int>,
    "benefit_rd_pp": <float>,
    "toxicity_dropout_risk_pp": <float>,
    "net_clinical_utility_pp": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
clinical_registry.tsv.gz	.tsv.gz	Patient key, analysis-set flag, baseline covariates, therapy, week-16 assessment, observed benefit, and 8-week toxicity/discontinuation.
tumor_assay_metrics.tsv.gz	.tsv.gz	Tumor purity, ploidy, whole-genome-duplication flag, RNA quality, expression batch, and library size.
longread_txr1_region.tsv.gz	.tsv.gz	Long-read QC, breakpoint topology, split-molecule support, alternate molecule fraction, local copy number, phase consistency, and promoter distance.

`expression_summary.tsv.gz`	.tsv.gz	TXR1 expression, lineage-marker expression, housekeeping expression, immune score, breakpoint-linked ASE fraction, and phase-linked reads.
`germline_pgx.tsv.gz`	.tsv.gz	DLR1 diplotype calls, call confidence, and ancestry group.
`dlr1_allele_table.tsv.gz`	.tsv.gz	Released DLR1 star-allele activity values.
`data_dictionary.tsv.gz`	.tsv.gz	Neutral column descriptions for all released files.

3 Answer Fields and Tolerances

The output schema maps directly to the estimand:

$$\begin{aligned} \texttt{benefit_rd_pp} &= \widehat{\Delta}, \\ \texttt{toxicity_dropout_risk_pp} &= \widehat{\tau}, \\ \texttt{net_clinical_utility_pp} &= \widehat{U}, \\ \texttt{therapy_class_code} &= \mathbf{1}\{\widehat{U} > 0\}. \end{aligned} \tag{1}$$

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`therapy_class_code`	1	Exact integer match; valid range $\{0, 1\}$	Treatment decision code from the sign of net clinical utility.
`benefit_rd_pp`	42.9143	Absolute error ≤ 0.5 pp	Week-16 clinical-benefit risk difference for TXR1i versus non-TXR1 systemic therapy in the recovered target population.
`toxicity_dropout_risk_pp`	35.9689	Absolute error ≤ 1.0 pp	Eight-week treatment-limiting toxicity/discontinuation risk under TXR1i in the same target population.
`net_clinical_utility_pp`	30.3252	Absolute error ≤ 0.4 pp	Benefit risk difference minus 0.35 times toxicity risk, all in percentage-point units.

4 Structure Diagram

The problem is deliberately built as a layered registry rather than a single biomarker threshold. Molecular evidence determines the target population; the target population then defines the causal estimand. Treatment, assessment, and toxicity are downstream of clinical history and PGx, so a correct molecular call is necessary but not sufficient.

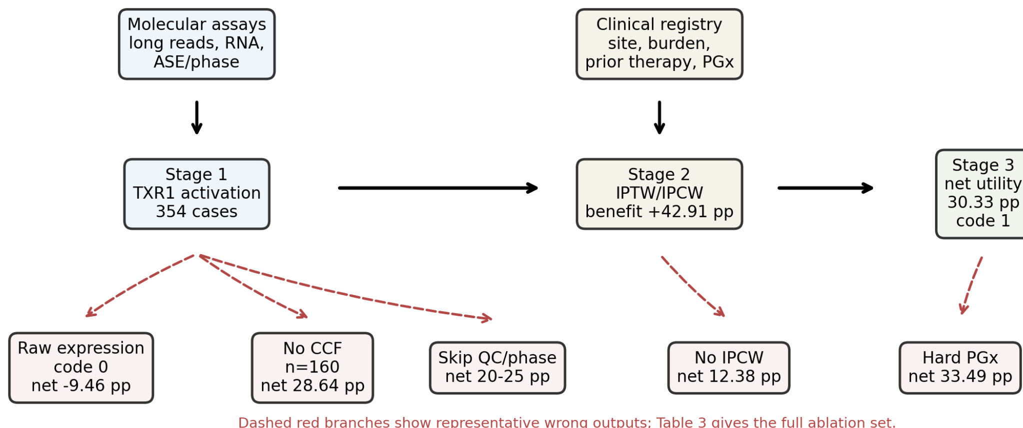

Figure 1: Cascaded problem structure. Black solid arrows show the intended analysis path from released molecular assays and clinical registry fields to the three graded outputs. Blue-tinted boxes are molecular target-reconstruction steps, tan boxes are target-trial treatment/assessment steps, and the green-tinted box is the final net-utility decision. Dashed red arrows and pink boxes show representative incorrect branches with their wrong outputs; the unified decision-point and ablation walkthrough table gives the complete ablation set. Long reads, purity/copy state, expression, allele-specific expression (ASE), and phasing feed the TXR1 target-activation call. Baseline clinical covariates and DLR1 pharmacogenomics (PGx) feed treatment, week-16 assessment, clinical benefit, and toxicity. IPTW/IPCW denotes inverse-probability treatment and censoring weighting; pp denotes percentage points.

5 Variables and Assumptions

The notation below is meant as a map between the plain-language story and the code. A non-specialist can read i as "one patient", A_i as "which treatment the patient received", Y_i as "whether clinical benefit was seen at week 16", and S_i as "whether the released evidence says this tumor is truly TXR1 target activated." Construction-only latent variables are listed to explain the simulated registry, but they are not present in the released analysis files.

Symbol or field	Type	Meaning and distributional role
i	index	Patient identifier, released as `patient_id`.
R_i	binary	Analysis-set indicator; $R_i = 1$ for ECOG ≤ 2 and no more than four prior therapy lines. All target and causal estimands condition on $R_i = 1$.
S_i	binary	Recoverable released-data target-activation call. The latent target state is unobserved in the released files; S_i is what an analyst can infer.
A_i	binary	Treatment assignment, with $A_i = 1$ for TXR1i and $A_i = 0$ for non-TXR1 systemic therapy.
Y_i	binary/blank	Week-16 clinical benefit if assessed; blank if `assessed16` $= 0$.

Symbol or field	Type	Meaning and distributional role
O_i	binary	Week-16 assessment indicator.
T_i	binary	Treatment-limiting toxicity/discontinuation by week 8.
G_i	binary	DLR1 slow activity after recomputing released diplotype activity and excluding low-confidence slow calls.
L_i	vector	Baseline clinical, site, period, lineage, disease-history, expression-residual, and PGx covariates used for nuisance models.
E_i	continuous	TXR1 expression residual after regressing `txr1_log_tpm` on lineage marker, RNA quality, library size, and batch in the analysis set.
$\widehat{C}_i$	continuous	Estimated structural-variant cancer cell fraction after correcting alternate molecule fraction for purity and local copy number.
Q_i	binary	Long-read QC gate: sufficient mapq, homology, segmental-duplication status, molecule depth, and read N50.
D_i	continuous	Breakpoint distance from the TXR1 promoter interval in kilobases.
B_i	categorical	Breakpoint topology; only `P2T` is target-facing.
H_i	continuous/vector	ASE and phase support: breakpoint-linked ASE fraction, phase-linked reads, and phase consistency.
Z_i	binary	Construction latent true target-mediated structural state, drawn by Eq. 6; it is used to build the synthetic registry but is not released.
$D_i^{\cdot}$ decoys	binary	Construction decoy indicators from Eq. 7, covering phase-dissociated, subclonal, distal, other-topology, and low-complexity/artifact signals.
site/period/lineage	categorical	Site, calendar period, and tumor lineage categories drawn by Eq. 2.
batch_i	categorical	Expression batch with levels B1/B2/B3; B3 appears in Eq. 7 as the low-complexity enrichment batch.
X_i^{clin}	mixed	Baseline clinical covariates: age, ECOG, tumor burden, prior therapy lines, and resistance status by Eq. 3.
purity/copy/altfrac	continuous	Released fields `tumor_purity`, `local_total_copy`, and `alt_molecule_fraction` used in the long-read single-copy CCF model.
ASE, reads, phase	mixed	Released handoff fields `target_ase_alt_fraction`, `phase_linked_reads`, and `phase_consistency`.
$p_{A,i}, p_{O,i}, p_{Y,i}, p_{T,i}$	probabilities	Treatment, assessment, benefit, and toxicity probabilities in the DGP or fitted estimator.
Y_i^1, Y_i^0	potential outcomes	Week-16 clinical benefit under TXR1i or non-TXR1 systemic therapy if observed.
T_i^1	potential outcome	8-week toxicity/discontinuation under TXR1i.

Assumptions for the estimand are the standard target-trial conditions applied to the synthetic registry: consistency of observed outcomes with assigned treatment, conditional exchangeability of treatment and assessment after the released covariates used in the nuisance models, positivity after probability clipping, and reliance only on released covariates. The component claims are intentionally cited separately: molecular-tumor-board actionability integrates molecular and clinical context [1]; long-read structural-variant interpretation depends on QC, breakpoint resolution, and phasing evidence [2–4]; purity/copy-state CCF reasoning is standard for somatic alteration interpretation [5];

target-trial emulation and inverse probability weighting support the observational causal contrast [6–9]; and pharmacogenomic implementation literature supports translating germline calls into toxicity-relevant activity before treatment decisions [10,11].

6 Data-Generating Process

This section describes the reference generator. The purpose of the model is to make visible shortcuts plausible while keeping the intended solution recoverable from the released files.

The equations use three recurring conventions. First, $\text{logit}^{-1}(x)$ converts a linear score into a probability between 0 and 1. Second, $\mathbf{1}\{\cdot\}$ equals 1 when a condition is true and 0 otherwise. Third, "decoy" means a released pattern that resembles the true target signal but should not count as target activation after all evidence is combined. The numbers in this section are not external biology; they are the synthetic knobs that create a realistic-looking registry with known failure modes.

Baseline registry

The generator assigns site, period, sex, and lineage class by categorical draws:

$$\begin{aligned} \Pr(\texttt{site} = S1, S2, S3, S4) &= (0.28, 0.27, 0.24, 0.21), \\ \Pr(\texttt{period} = P1, P2, P3) &= (0.33, 0.36, 0.31), \\ \Pr(\texttt{lineage} = A, B, C) &= (0.42, 0.34, 0.24). \end{aligned} \tag{2}$$

The generator also encodes lineage numerically as $0, 0.45, 1.15$. It draws age from a clipped normal distribution with mean 62 and SD 10, ECOG with probabilities $0.26, 0.46, 0.22, 0.06$, tumor burden from a clipped gamma variate plus an ECOG penalty, and prior lines from a clipped Poisson distribution. Prior resistance follows

$$\Pr(\texttt{prior_resistance} = 1) = \text{logit}^{-1}(-1.0 + 0.55\,\texttt{prior_lines} + 0.22\,\texttt{tumor_burden}). \tag{3}$$

The analysis set is

$$R_i = \mathbf{1}\{\texttt{ecog}_i \leq 2, \texttt{prior_lines}_i \leq 4\}. \tag{4}$$

The realized registry has 2210 analysis-set patients.

DLR1 PGx

PGx means pharmacogenomics: inherited variation that changes drug handling or toxicity risk. In this synthetic registry, DLR1 is not a tumor gene; it is a patient germline marker that makes TXR1i toxicity more likely when activity is slow. This is why DLR1 belongs in the treatment, assessment, and toxicity models, not in the molecular target-activation definition.

The generator draws two DLR1 alleles from $\{*1, *2, *4, *5\}$ with probabilities $(0.70, 0.18, 0.08, 0.04)$. Activity values are $1.0, 0.5, 0.0, 0.0$, and slow status is total activity ≤ 1.0. Low-confidence diplotypes are enriched in ancestry group G3 and site S4:

$$\Pr(\texttt{low_conf_pgx} = 1) = \text{logit}^{-1}(-3.15 + 1.15\,\mathbf{1}\{G3\} + 0.55\,\mathbf{1}\{S4\})R_i. \tag{5}$$

Low-confidence calls are perturbed toward false slow or false reference calls, and call confidence is visibly lower. The released allele table lets an analyst recompute activity, while the reference estimator treats low-confidence slow calls conservatively: a diplotype with activity score ≤ 1.0 is counted as DLR1 slow only when $\texttt{dlr1_call_confidence} \geq 0.94$.

Assay and latent structural classes

The generator draws tumor purity from a clipped beta-scaled variate, then draws ploidy, WGD, RNA quality, expression batch, and library size. The latent true target probability is

$$\Pr(Z_i = \text{target}) = \text{logit}^{-1}(-2.05 + 0.55\,\mathbf{1}\{B\} + 0.25\,\mathbf{1}\{C\} + 0.15\,\texttt{prior_resistance})R_i. \quad (6)$$

The generator then draws non-target decoys sequentially among remaining analysis-set cases:

$$\begin{aligned}
\Pr(D_i^{phase} = 1) &= \text{logit}^{-1}(-3.05 + 0.75\,\mathbf{1}\{C\} + 0.20\,\texttt{prior_resistance})R_i, \\
\Pr(D_i^{sub} = 1) &= \text{logit}^{-1}(-2.45 + 0.85\,\mathbf{1}\{C\} + 0.30\,\texttt{wgd_flag})R_i, \\
\Pr(D_i^{dist} = 1) &= \text{logit}^{-1}(-2.85 + 0.55\,\mathbf{1}\{B\} + 0.35\,\texttt{prior_resistance})R_i, \\
\Pr(D_i^{orient} = 1) &= \text{logit}^{-1}(-2.40 + 0.75\,\mathbf{1}\{C\})R_i, \\
\Pr(D_i^{art} = 1) &= \text{logit}^{-1}(-1.95 + 0.65\,\mathbf{1}\{S3\} + 0.75\,\mathbf{1}\{B3\})R_i.
\end{aligned} \quad (7)$$

The realized construction-level counts are 387 target-mediated, 110 phase-dissociated P2T, 173 subclonal P2T, 114 distal P2T, 137 other-topology, and 252 low-complexity signal cases.

Long-read evidence

Long-read sequencing observes long DNA fragments, which is useful for finding structural variants because a single read can span a rearranged junction. The same feature also creates failure modes: repeats, homologous sequence, low mapping quality, and short or shallow read support can make a false junction look convincing. The released file therefore includes both the apparent junction signal and the QC evidence needed to decide whether to trust it.

HiFi and ONT platforms are mixed 62/38. Molecule depth, read N50, mapq, homology, segmental duplication, and low-complexity score are drawn so that low-complexity signals have lower mapping quality, higher homology, segmental-duplication overlap, higher low-complexity score, and shorter reads. Target, phase-dissociated, subclonal, distal, and low-complexity states all can present as `P2T`; orientation decoys are `INV` or `D2T`.

For non-null structural states, local copy number and true CCF are state specific. The released alternate molecule fraction is generated from purity, copy number, and CCF under implicit alteration multiplicity one:

$$\text{altfrac}_i = \frac{\text{purity}_i\, C_i}{\text{purity}_i\,\text{copy}_i + 2(1 - \text{purity}_i)} + \epsilon_i, \qquad \epsilon_i \sim N(0, 0.018^2), \quad (8)$$

with null cases drawn near zero. A general somatic SV model would multiply C_i by the number of altered allelic copies, bounded by local major copy number; that multiplicity is fixed to one in this synthetic registry, so the inverse purity/copy correction recovers the CCF proxy used by the reference estimator. Subclones and artifacts can still have enough split molecules to mislead simple count thresholds.

Expression, ASE, and phase

Expression asks how much TXR1 RNA is present, but high RNA alone does not prove the structural variant caused it. ASE and phase add the missing link. ASE asks whether one allele contributes

more TXR1 RNA than the other; phase asks whether the expressed allele can be connected to the breakpoint-supporting reads. The correct molecular call requires all three ideas: adjusted expression, allele-specific expression, and read-backed phase consistency.

Lineage marker expression is driven by lineage and tumor burden. TXR1 expression combines lineage baseline, RNA quality, library size, batch, and a structural-state signal:

$$\begin{aligned} \mathrm{TXR1}_i =& 2.52 + 0.83\,\mathrm{lineage}_i + 0.10\,\mathrm{RNAQ}_i + 0.006\,\mathrm{library}_i \\ &+ \mathrm{batch}_i + \mathrm{target_signal}_i + \eta_i, \end{aligned} \tag{9}$$

where $\eta_i \sim N(0, 0.28^2)$. The target signal is 2.05 for true target, 1.90 for phase-dissociated, 1.12 for subclonal, 1.95 for distal, 1.55 for low-complexity, and 1.90 for orientation decoy. ASE and phase are then state specific: true targets have ASE around 0.82, phase consistency around 0.94, and high phase-linked read counts; phase-dissociated P2T cases have weak ASE/phase despite P2T topology; subclonal, distal, and orientation decoys can have deceptively strong partial support.

Treatment, benefit, toxicity, and assessment

The outcome part of the simulation is written to resemble a registry rather than a randomized trial. Clinicians tend to give TXR1i to patients who look more likely to benefit, and they avoid it in some patients with high DLR1 toxicity risk. Later, toxicity makes week-16 assessment less likely. These two selection mechanisms are why the final answer cannot be a raw treated-minus-control comparison among the observed week-16 outcomes.

The treatment model makes assignment depend on true target status, apparent raw expression, lineage, disease history, site, period, and PGx avoidance:

$$\begin{aligned} \operatorname{logit} \Pr(A_i = 1) =& -1.55 + 0.78\,Z_i + 0.72\,\mathbf{1}\{\mathrm{TXR1}_i > 7.85\} + 0.42\,\mathbf{1}\{C\} \\ &+ 0.82\,\mathrm{resistance}_i + 0.33\,\mathrm{priorlines}_i + 0.22\,\mathrm{burden}_i \\ &- 1.80\,G_i + 0.35\,\mathbf{1}\{S3\} - 0.25\,\mathbf{1}\{S1\} + 0.20\,\mathbf{1}\{P3\}. \end{aligned} \tag{10}$$

TXR1i benefit is intentionally target mediated:

$$\begin{aligned} \operatorname{logit} \Pr(Y_i = 1) =& -0.66 + 2.35\,A_i Z_i + 0.10 Z_i - 1.45\,A_i(D_i^{phase} + D_i^{sub} + D_i^{dist} + D_i^{orient}) \\ &- 0.82\,A_i \mathbf{1}\{\mathrm{lineage}_i = C, Z_i = 0\} - 1.25\,A_i D_i^{art} \\ &- 0.46\,\mathrm{resistance}_i \\ &- 0.22\,\mathrm{priorlines}_i - 0.28\,\mathrm{burden}_i - 0.19\,\mathrm{ECOG}_i \\ &+ 0.15\,\mathrm{immune}_i + 0.10\,\mathbf{1}\{\mathrm{period}_i = P3\}. \end{aligned} \tag{11}$$

Toxicity is PGx driven rather than tumor-sensitivity driven:

$$\begin{aligned} \operatorname{logit} \Pr(T_i = 1) =& -2.75 + 0.70 A_i + 3.00 A_i G_i + 0.25(\mathrm{age}_i - 62)/10 \\ &+ 0.28\,\mathrm{ECOG}_i + 0.17\,\mathrm{priorlines}_i + 0.25\,\mathrm{burden}_i/3 - 0.20\,\mathbf{1}\{S2\}. \end{aligned} \tag{12}$$

Assessment is informative:

$$\operatorname{logit} \Pr(O_i = 1) = 2.45 - 3.10 T_i - 0.55 A_i G_i - 0.25\,\mathrm{ECOG}_i - 0.12\,\mathrm{burden}_i + 0.15\,\mathbf{1}\{S2\}. \tag{13}$$

Thus dropping unassessed cases or coding them as no benefit distorts the target-trial contrast.

7 Analyst Walkthrough

Stage 1, step 1: visible registry and raw expression

The first impression from the released files is that TXR1i is commonly used and raw TXR1 expression separates many cases. There are 1306 TXR1i-treated patients and 1653 analysis-set patients above the visible raw-expression shoulder at 7.85. This is a tempting tumor-board shortcut: high target expression often drives drug selection, and the data dictionary says TXR1 expression is available.

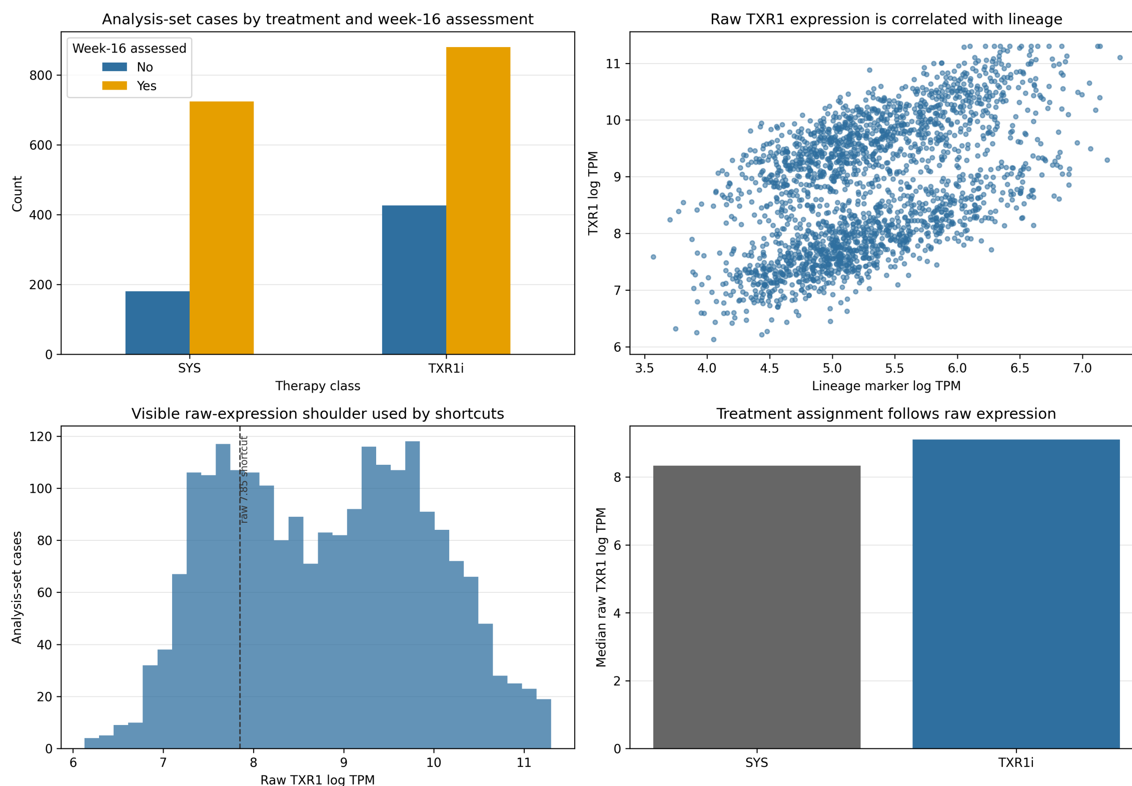

Figure 2: Registry and expression overview. In the upper-left panel, blue and orange bars are unassessed and assessed week-16 outcomes, respectively, within non-TXR1 systemic therapy (SYS) and TXR1 inhibitor (TXR1i) groups. The upper-right panel plots one analysis-set patient per point and shows that raw TXR1 expression tracks the lineage marker. The lower-left histogram counts analysis-set patients by raw TXR1 expression; the vertical dashed line marks the tempting but incorrect raw-expression shortcut at 7.85 log TPM. The lower-right bars compare median raw TXR1 expression by treatment class, with gray for SYS and blue for TXR1i. Raw TXR1 expression is associated with treatment and lineage, so it is not by itself a target-mediated activation call.

```
X = [1, lineage_marker_log_tpm, rna_quality,
     library_size_million, batch_B2, batch_B3]
beta = least_squares(X[analysis_set], txr1_log_tpm[analysis_set])
txr1_expr_resid = txr1_log_tpm - X @ beta
```

After this residualization, the residual-lineage correlation is effectively zero. The analysis must therefore move from raw expression to a structural target call: expression is one support layer, not the definition of the target population.

Stage 1, step 2: purity, copy number, and single-copy CCF

You might next treat alternate molecule fraction or split-molecule support as clonal evidence. In plain terms, the alternate molecule fraction is "what share of sequenced molecules show the rearranged junction." It sounds like the right quantity, but it is not yet the fraction of tumor cells carrying the event. That wrong path produces concrete failures before any correction: `no_purity_correction` selects only 160 cases and misses the answer by 4.318 tolerance units, while `ignore_subclonality` selects 152 cases and fails by 17.823 tolerance units. The released long-read file exposes `alt_molecule_fraction`, `split_molecule_count`, purity, and local copy number, so this mistake looks defensible at first glance. The problem is that low-purity clonal events can have small raw fractions because normal cells dilute the sample, while amplified or high-depth subclonal regions can look well supported by split-molecule counts and, in general somatic SV settings, altered-copy multiplicity can inflate raw alternate fractions.

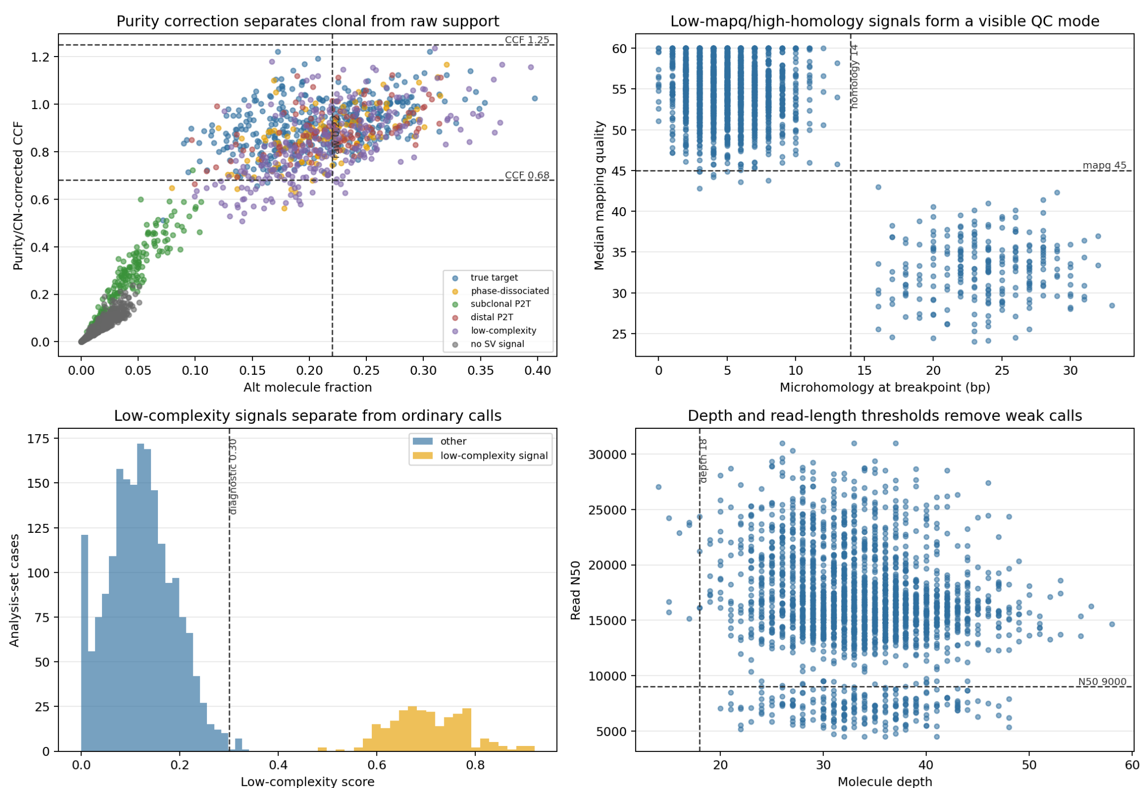

Figure 3: Purity/copy correction and long-read QC. Dashed gray lines mark thresholds; colored points in the CCF panel follow the legend for molecular signal class (true target, phase-dissociated, subclonal P2T, distal P2T, low-complexity, or no SV signal). The upper-left panel marks the 0.68–1.25 single-copy cancer-cell-fraction (CCF) proxy window and the raw 0.22 alternate-molecule-fraction shortcut. The upper-right panel marks median mapping quality (mapq) 45 and breakpoint microhomology 14 bp. The lower-left histogram uses blue for other calls and orange for low-complexity signals and marks low-complexity score 0.30 as a diagnostic separation. The lower-right panel marks molecule depth 18 and read N50 9000 bases. These thresholds define the long-read QC gate used before promoter-distance and expression/phase handoff checks.

The single-copy CCF proxy converts the raw molecule fraction into an estimated cancer-cell fraction for the synthetic alteration model. The numerator starts with the observed alternate fraction; the bracketed term adjusts for tumor purity and local copy number; the division by purity removes the

dilution from normal cells. Algebraically, this is the inverse of the simulation's molecule-fraction equation:

$$\widehat{C}_i = \frac{\texttt{alt_molecule_fraction}_i\{\texttt{tumor_purity}_i\texttt{local_total_copy}_i + 2(1 - \texttt{tumor_purity}_i)\}}{\texttt{tumor_purity}_i}.$$

Because all simulated SVs have alteration multiplicity one, $\widehat{C}_i$ equals the CCF used by the reference estimator. More generally, the same purity/copy expression estimates CCF multiplied by alteration multiplicity; in amplified regions, a biologically general clonal upper window would be interpreted relative to local major copy number rather than as a universal CCF upper bound. The synthetic single-copy clonal window, chosen from the implemented DGP and shown in the figure, is $0.68 \leq \widehat{C}_i \leq 1.25$.

```
denom = tumor_purity * local_total_copy + 2 * (1 - tumor_purity)
sv_ccf_est = alt_molecule_fraction * denom / tumor_purity
clonal = (sv_ccf_est >= 0.68) & (sv_ccf_est <= 1.25)
```

Sequentially, the released-data analysis starts with 2210 analysis-set cases, 1036 P2T cases, 756 P2T cases after long-read QC, and 584 cases after the single-copy clonal CCF window. This is still not the target set; it only says the structural event is target-facing, high quality, and clonal.

Stage 1, step 3: long-read artifact QC and promoter proximity

After this purity/copy correction, another plausible wrong path is to accept any convincing P2T junction. The mapping-artifact ablation keeps 419 cases and returns 37.6138 benefit pp, 36.6149 toxicity pp, and 24.7986 net pp, failing by 13.816 tolerance units. The distance-agnostic ablation keeps 456 cases and fails by 27.431 tolerance units. The structural-variant evidence therefore needs artifact and proximity checks. Long-read SV discovery and clinical SV interpretation both make this plausible: mapping quality, homology, segmental duplication, depth, read length, and breakpoint geometry determine whether a junction can be trusted [2–4]. The problem includes recurrent low-complexity signals and distal P2T-like junctions that share expression and support features with true target cases.

The recoverable QC gate is:

$$Q_i = \mathbf{1}\{\texttt{mapq_median} \geq 45,\ \texttt{homology_bp} \leq 14,\ \texttt{segmental_dup_overlap} = 0,\ \texttt{molecule_depth} \geq 18,\ \texttt{read_n50} \geq 9000\}.$$

The low-complexity score is displayed as a diagnostic mode in the figure; the implemented final gate captures these artifacts through mapq, homology, segmental-duplication, depth, and read-length checks. The promoter-proximity gate $D_i \leq 110$ kb is a synthetic threshold chosen from the DGP, not a literature-derived universal cutoff.

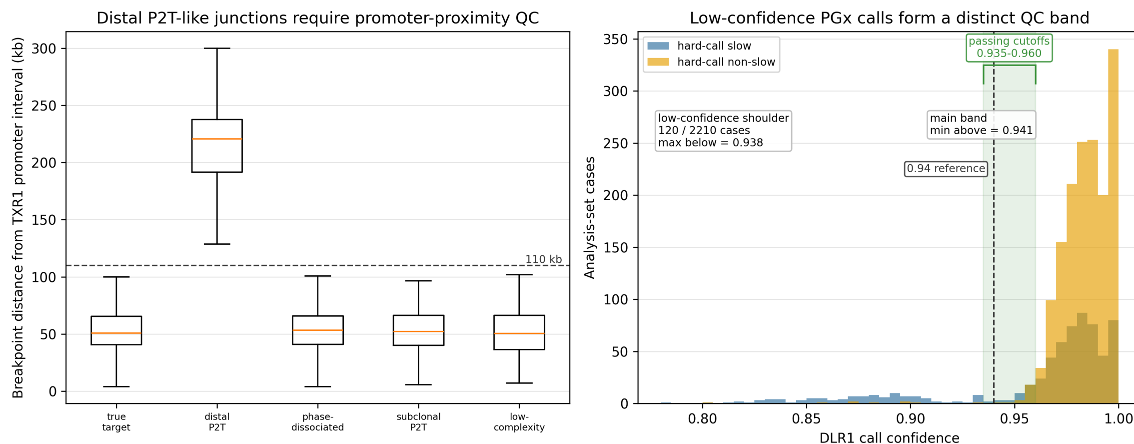

Figure 4: Promoter-distance and PGx-confidence diagnostics. In the left panel, each boxplot summarizes one molecular signal class; the orange line is the median, the box is the interquartile range, and whiskers show the plotted non-outlier range. The dashed gray horizontal line marks the 110 kb promoter-proximity threshold, separating distal P2T-like junctions from true target and other proximal signal classes. In the right panel, blue and orange histograms are hard-call DLR1-slow and hard-call non-slow pharmacogenomic (PGx) calls. The dashed gray vertical line is the 0.94 DLR1 call-confidence reference cutoff; the pale green shaded region and green bracket mark the 0.935–0.960 cutoff range that preserves answer-equivalent benefit, toxicity, and net-utility values. Text boxes label the low-confidence shoulder below 0.94 and the main high-confidence band above it.

Stage 1, step 4: residual expression, ASE, and phase handoff

Once a clonal, high-quality, promoter-proximal P2T junction is found, a shortcut still fails. The phase-expression ablation keeps 440 cases and returns 33.2016 benefit pp and 20.4921 net pp, failing by 24.583 tolerance units. The strict-threshold regression shortcut keeps 403 cases and looks close enough to be dangerous, but the released answer contract rejects it because it returns 40.4126 benefit pp, 38.4179 toxicity pp, and 26.9663 net pp, failing by 8.397 tolerance units. The missing step is to ask whether the breakpoint is linked to the expressed TXR1 allele. The synthetic handoff gate is:

$$E_i \geq 0.65, \qquad \texttt{target_ase_alt_fraction}_i \geq 0.62,$$
$$\texttt{phase_linked_reads}_i \geq 8, \qquad \texttt{phase_consistency}_i \geq 0.78.$$

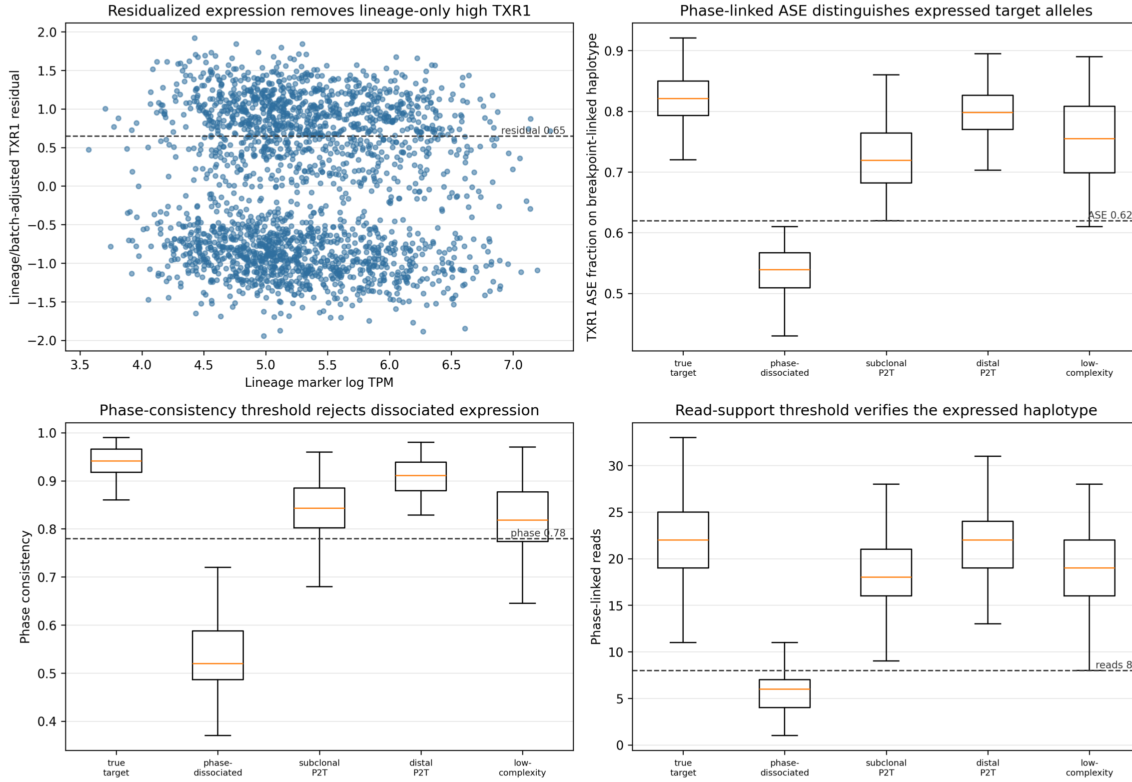

Figure 5: Correct molecular handoff. Dashed gray lines mark the synthetic handoff thresholds used in the target call: TXR1 expression residual 0.65, breakpoint-linked allele-specific-expression (ASE) fraction 0.62, phase consistency 0.78, and phase-linked reads 8. The upper-left panel plots one analysis-set patient per point after residualizing TXR1 expression for lineage, RNA quality, library size, and batch. The other panels are boxplots by molecular signal class; orange lines are medians, boxes are interquartile ranges, and whiskers show the plotted non-outlier ranges. Residual expression, ASE, phase consistency, and phase-linked read support must all pass because phase-dissociated, subclonal, distal, and low-complexity signals can satisfy only a subset of the checks.

This step is not decorative. Phase-dissociated, distal, subclonal, and orientation-like signals satisfy different subsets of these checks, so the expression residual, ASE, phase-linked read count, and phase-consistency thresholds must be applied together. The threshold values are synthetic design thresholds, while the need for phasing and read-supported SV interpretation is the real-world practice supported by long-read SV literature [2–4].

```
target = valid_breakpoint & clonal & (topology == "P2T")
target &= (breakpoint_distance_kb <= 110) & (txr1_expr_resid >= 0.65)
target &= (target_ase_alt_fraction >= 0.62)
target &= (phase_linked_reads >= 8) & (phase_consistency >= 0.78)
target &= (analysis_set == 1)
```

The final released-data target contains 354 cases. Relative to the event definition used to construct the released cohort, it retains 354 recoverable positives and excludes 33 events that fall below recoverability thresholds such as depth or CCF. This distinction is intentional: the public target is the molecularly recoverable treatment-selection population, not every simulated event in the underlying data-generating process.

Stage 2 and Stage 3: treatment, assessment, and PGx standardization

After the target call, the problem becomes causal, and the easy clinical shortcut fails first. "Causal" here means the report is not asking whether treated patients happened to do better; it asks what benefit difference would be expected if the same target population could be assigned to TXR1i versus non-TXR1 therapy under a trial-like comparison. `no_propensity` returns 39.6029 benefit pp, 31.1284 toxicity pp, and 28.7079 net pp; `no_ipcw` returns 24.9704 / 35.9689 / 12.3813; `hard_pgx_calls` returns 45.1437 / 33.3100 / 33.4852; and the crude target shortcut returns 39.2486 / 36.9123 / 26.3293. The target set has 257 TXR1i-treated patients, 97 non-TXR1 systemic controls, 256 week-16 assessed patients, and 88 toxicity/discontinuation events. These wrong outputs look plausible because they use the correct molecular target, but treatment still depends on disease history, raw expression, site, period, and DLR1 avoidance, and week-16 assessment is informative because toxicity reduces assessment probability.

Inverse probability treatment weighting (IPTW) means upweighting patients who received a treatment that was unlikely given their baseline covariates, so the treated and control groups resemble the same target population. A standardized mean difference (SMD) is a scale-free covariate imbalance diagnostic; values near zero after weighting indicate better balance [7,8].

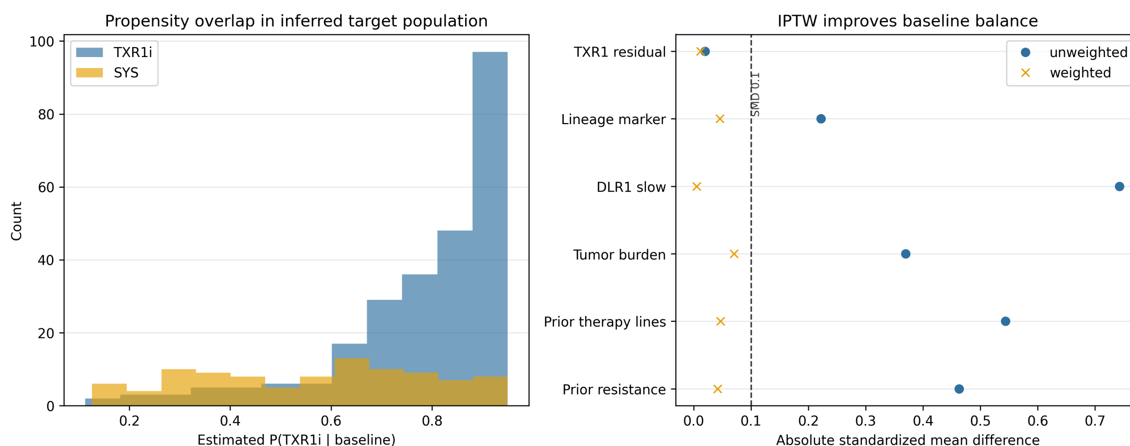

Figure 6: Treatment propensity and target-trial diagnostics. The left histogram shows estimated treatment propensity $P(\text{TXR1i} \mid \text{baseline})$ in the inferred target population, with blue for TXR1 inhibitor (TXR1i) and orange for non-TXR1 systemic therapy (SYS). The right panel plots absolute standardized mean differences (SMDs), a scale-free covariate imbalance measure; blue circles are unweighted imbalances and orange crosses are after inverse probability treatment weighting (IPTW). The dashed gray vertical line marks the conventional SMD 0.1 balance threshold. The plot shows that IPTW balances the major clinical, pharmacogenomic, and assay covariates used for the target-trial contrast [7,8].

Before weighting, absolute standardized mean differences in the target set are 0.463 for prior resistance, 0.544 for prior lines, 0.370 for tumor burden, 0.743 for DLR1 slow, 0.222 for lineage-marker expression, and 0.020 for TXR1 residual. After treatment weighting they are 0.042, 0.047, 0.070, 0.005, 0.046, and 0.012 respectively. The estimated treatment probabilities range from 0.1127 to 0.9500 with median 0.7805. Assessment probabilities range from 0.1000 to 0.9535 with median 0.7797.

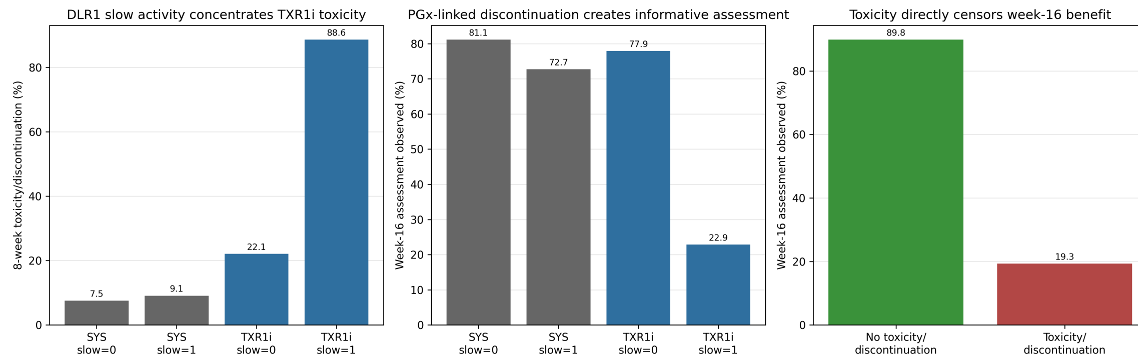

Figure 7: PGx toxicity and assessment. Gray bars are non-TXR1 systemic therapy (SYS), blue bars are TXR1 inhibitor (TXR1i), green indicates no 8-week toxicity/discontinuation, and red indicates toxicity/discontinuation. Bar labels give observed percentages. In the left panel, 'slow=1' means DLR1 slow pharmacogenomic (PGx) activity and shows the strong TXR1i-specific toxicity increase. The middle panel shows week-16 assessment rates by treatment and DLR1-slow status. The right panel collapses across treatment to show that toxicity/discontinuation directly reduces observed week-16 assessment, motivating inverse probability of censoring weighting (IPCW) for the clinical-benefit endpoint.

The target population is 22.3% DLR1 slow. Among treated target patients, toxicity is 22.1% for non-slow patients and 88.6% for slow patients. Week-16 assessment rates are 89.8% without toxicity and 19.3% with toxicity, as shown in the right panel of the PGx toxicity figure. The DLR1 confidence panel in Figure 4 shows why hard calls are not enough: a small low-confidence shoulder affects toxicity standardization. The reference cutoff is locally robust, with cutoffs from 0.935 through 0.960 yielding the same therapy class and answer-equivalent benefit, toxicity, and net-utility values, but more distant cutoffs change the realized numeric target. The cutoff is therefore a visible-data operational convention, not a hidden biological constant.

Inverse probability of censoring weighting (IPCW) uses an assessment-probability model to upweight assessed patients who resemble patients likely to be missing, so the observed week-16 benefits represent the full target trial [7,9]. The model uses baseline variables, treatment, DLR1 slow status, DLR1 low-confidence status, and the TXR1i-by-DLR1-slow interaction; it does not condition on observed week-8 toxicity, because toxicity is a post-treatment endpoint and lies on the pathway that makes assessment informative. In other words, IPCW prevents the analysis from silently learning only from the healthier patients who remained assessable, while avoiding adjustment for the post-treatment toxicity variable whose risk is separately reported. Therefore a model that ignores PGx confidence, treatment weighting, or assessment censoring can look scientific while returning the wrong estimand.

Stage 3: final net-utility accounting

The normalized weighted benefit risks are:

$$\widehat{r}_1 = 0.583826, \qquad \widehat{r}_0 = 0.154683.$$

Thus

$$\widehat{\Delta} = 100(\widehat{r}_1 - \widehat{r}_0) = 42.9143\,\text{pp}.$$

The standardized TXR1i toxicity risk is

$$\widehat{\tau} = 35.9689\,\text{pp}.$$

The net utility is
$$\widehat{U} = 42.9143 - 0.35(35.9689) = 30.3252\,\text{pp},$$
so the therapy class is 1.

8 Estimand

An estimand is the exact quantity the analysis is trying to learn. Here it is deliberately phrased as a trial question: among patients whose released data show TXR1 target activation, how much would week-16 benefit change if that same target population received TXR1i rather than non-TXR1 systemic therapy, and what toxicity risk would that same target population face under TXR1i? The notation Y_i^1 and Y_i^0 means the week-16 outcome the same patient would have under treatment 1 or treatment 0; only one is observed, so the estimator has to use comparable patients and weights.

The target population is the recoverable TXR1 target-mediated activation population in the released data:
$$\mathcal{S} = \{i : R_i = 1, S_i = 1\}. \tag{14}$$

Here S_i is not a latent label. It is the deterministic target call produced from the released assays. This distinction matters because the problem asks for what can be inferred from the provided files.

The clinical-benefit estimand is the marginal week-16 target-trial risk difference if all target patients were assessed:
$$\Delta = 100\left(E[Y_i^1 \mid i \in \mathcal{S}] - E[Y_i^0 \mid i \in \mathcal{S}]\right). \tag{15}$$

The toxicity estimand is the 8-week TXR1i toxicity/discontinuation risk in the same target population:
$$\tau = 100\,E[T_i^1 \mid i \in \mathcal{S}]. \tag{16}$$

The decision utility is
$$U = \Delta - 0.35\,\tau, \qquad \texttt{therapy_class_code} = \mathbf{1}\{U > 0\}. \tag{17}$$

9 Estimator

Preliminary harmonization and Stage 1: target reconstruction

The estimator first joins the released tables by `patient_id`. As preliminary harmonization, it recomputes DLR1 activity from the allele table for later treatment, assessment, and toxicity models; DLR1 is not used in the molecular target-activation call. Let $a(*1) = 1.0$, $a(*2) = 0.5$, and $a(*4) = a(*5) = 0$. The hard slow indicator is $I\{a(h_1) + a(h_2) \leq 1.0\}$, and the analysis slow indicator is
$$G_i = I\{a(h_{i1}) + a(h_{i2}) \leq 1.0,\ \texttt{dlr1_call_confidence}_i \geq 0.94\}.$$

The companion low-confidence indicator $I\{\texttt{dlr1_call_confidence} < 0.94\}$ remains in the treatment and assessment nuisance models. Stage 1 then calculates an expression residual. Plainly, the residual asks whether TXR1 is high after subtracting lineage, RNA quality, library size, and batch effects. Least squares is used here as a deterministic nuisance fit for this synthetic adjustment, not as a

biological claim:

$$\begin{aligned}\texttt{txr1_log_tpm}_i = \beta_0 &+ \beta_1\texttt{lineage_marker_log_tpm}_i + \beta_2\texttt{rna_quality}_i \\ &+ \beta_3\texttt{library_size_million}_i + \beta_4\mathbf{1}\{\texttt{expression_batch}_i = \texttt{B2}\} \\ &+ \beta_5\mathbf{1}\{\texttt{expression_batch}_i = \texttt{B3}\} + E_i.\end{aligned} \tag{18}$$

The model is fit by least squares in the analysis set.

It then reconstructs the single-copy CCF proxy:

$$\widehat{C}_i = \frac{\texttt{alt_molecule_fraction}_i\{\texttt{tumor_purity}_i\texttt{local_total_copy}_i + 2(1 - \texttt{tumor_purity}_i)\}}{\texttt{tumor_purity}_i}. \tag{19}$$

The target indicator is:

$$\begin{aligned}S_i =\mathbf{1}\{&R_i = 1,\ Q_i = 1,\ \texttt{breakpoint_topology}_i = \texttt{P2T},\ D_i \le 110,\ 0.68 \le \widehat{C}_i \le 1.25, \\ &E_i \ge 0.65,\ \texttt{target_ase_alt_fraction}_i \ge 0.62, \\ &\texttt{phase_linked_reads}_i \ge 8,\ \texttt{phase_consistency}_i \ge 0.78\}.\end{aligned} \tag{20}$$

Stage 2: treatment and assessment weights

Inside $S_i = 1$, the estimator needs a smooth probability for receiving TXR1i given baseline covariates. It uses logistic regression, a model for binary outcomes, with a very weak L2 penalty ($C = 10^6$) only to make the fit deterministic and numerically stable. Features are age, ECOG, tumor burden, prior lines, prior resistance, lineage-marker expression, TXR1 residual, DLR1 slow, DLR1 low-confidence status, site, period, and lineage class. Predicted probabilities are clipped to $[0.05, 0.95]$, a synthetic stability choice that preserves positivity for the inverse weighting estimator [6–8]:

$$\widehat{p}_{A,i} = \text{clip}\{\widehat{\Pr}(A_i = 1 \mid L_i), 0.05, 0.95\}. \tag{21}$$

The inverse treatment weight upweights patients whose observed treatment was unlikely under the fitted propensity model [7,8]:

$$w_i^A = \frac{A_i}{\widehat{p}_{A,i}} + \frac{1 - A_i}{1 - \widehat{p}_{A,i}}. \tag{22}$$

The assessment model uses the same baseline features plus treatment and the TXR1i-by-DLR1-slow interaction. Observed week-8 toxicity is deliberately excluded from this model because it is a post-treatment toxicity endpoint, not a baseline covariate; the assumption is that, conditional on baseline target-population covariates, treatment, DLR1 slow status, and their interaction, the assessed patients can represent the full week-16 target-trial endpoint through inverse probability of censoring weighting. Probabilities are clipped to $[0.10, 0.98]$, again as a synthetic stability choice:

$$\widehat{p}_{O,i} = \text{clip}\{\widehat{\Pr}(O_i = 1 \mid A_i, L_i, A_iG_i), 0.10, 0.98\}, \qquad w_i^O = 1/\widehat{p}_{O,i}. \tag{23}$$

The maximum combined benefit weight is 18.258; the individual clipping caps are 20 for IPTW and 10 for IPCW.

Stage 3: marginal risks

Only assessed outcomes contribute to the benefit numerator, but the weights reweight them back to the target trial [6–9]:

$$\widehat{r}_1 = \frac{\sum_{i\in\mathcal{S}} A_i O_i Y_i w_i^A w_i^O}{\sum_{i\in\mathcal{S}} A_i O_i w_i^A w_i^O}, \qquad \widehat{r}_0 = \frac{\sum_{i\in\mathcal{S}} (1-A_i) O_i Y_i w_i^A w_i^O}{\sum_{i\in\mathcal{S}} (1-A_i) O_i w_i^A w_i^O}. \tag{24}$$

Toxicity under TXR1i is standardized from treated patients back to the target population using the treatment weights [7,8]:

$$\widehat{\tau} = 100 \frac{\sum_{i\in\mathcal{S}} A_i T_i w_i^A}{\sum_{i\in\mathcal{S}} A_i w_i^A}. \tag{25}$$

Finally, $\widehat{\Delta} = 100(\widehat{r}_1 - \widehat{r}_0)$ and $\widehat{U} = \widehat{\Delta} - 0.35\widehat{\tau}$.

10 Decision-Point and Ablation Walkthrough

The ablation suite includes the correct estimator, one target-equivalent positive-control estimator, single-failure-mode outputs, the strict-threshold shortcut, and compound shortcuts. The table reports the actual public-output values; a row passes only if the therapy class is exact and benefit, toxicity, and net utility are within their configured tolerances. Quantitative output is ordered as code; benefit pp / toxicity pp / net utility pp; target n. Short display labels are used here for readability.

Decision point	Display analysis	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	Correct estimator	code 1; 42.9143 / 35.9689 / 30.3252; n=354	yes	none	Reference target reconstruction, target-trial benefit, IPCW, and PGx-standardized toxicity.
Target-equivalent check	Outcome-regression check	code 1; 42.5788 / 35.9689 / 29.9897; n=354	yes	none	Acceptable outcome-regression check after the same target, treatment, assessment, and toxicity decisions.
Stage 1 target reconstruction	Raw-expression target	code 0; 3.4654 / 36.9233 / -9.4577; n=1653	no	Raw expression	Treats lineage- and assay-inflated TXR1 expression as the causal SV-driven target.
Stage 1 target reconstruction	No CCF correction	code 1; 40.7554 / 34.6198 / 28.6385; n=160	no	CCF inference	Uses raw alternate molecule fraction instead of purity/copy-aware CCF.
Stage 1 target reconstruction	Keep mapping artifacts	code 1; 37.6138 / 36.6149 / 24.7986; n=419	no	Long-read QC	Allows low-complexity or mapping-artifact P2T-like calls into the molecular target.
Stage 1 target reconstruction	No subclonality filter	code 1; 50.8949 / 38.4018 / 37.4543; n=152	no	CCF inference	Treats subclonal amplified support as equivalent to clonal target support.
Stage 1 target reconstruction	No distance gate	code 1; 32.0529 / 36.2861 / 19.3528; n=456	no	Promoter proximity	Includes distal P2T-like events that are not the SV-driven promoter-proximal activation event.
Stage 1 molecular handoff	No ASE/phase handoff	code 1; 33.2016 / 36.3129 / 20.4921; n=440	no	ASE/phase handoff	Lets phase-dissociated or expression-only signals pass as causal TXR1 activation.
Stage 2 treatment weighting	No treatment weighting	code 1; 39.6029 / 31.1284 / 28.7079; n=354	no	IPTW	Ignores treatment-selection imbalance after the molecular target is reconstructed.
Stage 2 outcome ascertainment	No assessment weighting	code 1; 24.9704 / 35.9689 / 12.3813; n=354	no	IPCW	Treats week-16 benefit as complete even though toxicity makes assessment informative.
Stage 3 toxicity standardization	Hard PGx calls	code 1; 45.1437 / 33.3100 / 33.4852; n=354	no	PGx uncertainty	Collapses low-confidence DLR1 activity calls even though PGx confidence is non-random.
Stage 2–3 partial shortcut	Crude target contrast	code 1; 39.2486 / 36.9123 / 26.3293; n=354	no	Weighting/accounting	Uses the right target but a crude benefit contrast and incomplete toxicity standardization.
Stage 1 molecular handoff	Strict threshold shortcut	code 1; 40.4126 / 38.4179 / 26.9663; n=403	no	Handoff thresholds	Uses a stricter-looking threshold set that admits the wrong molecular target composition.
Stage 2 outcome ascertainment	Exclude toxicity dropouts	code 1; 50.2853 / 0.0000 / 50.2853; n=354	no	Selection on observed benefit	Excludes toxicity-driven missing outcomes and thereby removes the toxicity endpoint itself.
Compound shortcut	Raw target + unweighted	code 0; 2.0948 / 31.4338 / -8.9070; n=1653	no	Stages 1–3	Combines raw-expression targeting with unweighted treatment and toxicity shortcuts.
Global shortcut	Overall registry response	code 0; 2.1597 / 30.6279 / -8.5600; n=2210	no	Target definition	Answers the overall registry-response question rather than the SV-driven target question.
Expression shortcut	High expression only	code 0; 4.1196 / 31.8966 / -7.0442; n=1394	no	Target definition	Uses high expression alone, which captures lineage and assay structure rather than causal SV activation.

Decision point	Display analysis	Quantitative output	Pass?	Failure point	Why the approach is wrong
Treatment-frequency shortcut	Most-treated class	code 0; 2.1597 / 30.6279 / -8.5600; n=2210	no	Target definition	Selects the most-treated group rather than reconstructing the molecular actionability class.

Table 3: Unified decision-point and ablation walkthrough for the TXR1 SV-driven target-trial problem.

11 References

1. Tsimberidou AM, Kahle M, Vo HH, Baysal M, Johnson A, Meric-Bernstam F. Molecular tumour boards: current and future considerations for precision oncology. *Nature Reviews Clinical Oncology*. 2023;20:843–863. DOI: 10.1038/s41571-023-00824-4. https://www.nature.com/articles/s41571-023-00824-4

2. Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nature Methods*. 2023;20:1143–1158. DOI: 10.1038/s41592-023-01932-w. https://www.nature.com/articles/s41592-023-01932-w

3. Liu Z, Roberts R, Mercer TR, Xu J, Sedlazeck FJ, Tong W. Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biology*. 2022;23:68. DOI: 10.1186/s13059-022-02636-8. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02636-8

4. Smolka M, Paulin LF, Grochowski CM, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*. 2024;42:1571–1580. DOI: 10.1038/s41587-023-02024-y. https://www.nature.com/articles/s41587-023-02024-y

5. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*. 2012;30:413–421. DOI: 10.1038/nbt.2203. https://www.nature.com/articles/nbt.2203

6. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. 2022;328:2446–2447. DOI: 10.1001/jama.2022.21383. https://jamanetwork.com/journals/jama/fullarticle/2799678

7. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*. 2022;15:14–20. DOI: 10.1093/ckj/sfab158. https://academic.oup.com/ckj/article/15/1/14/6358134

8. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*. 2015;34:3661–3679. DOI: 10.1002/sim.6607. https://pmc.ncbi.nlm.nih.gov/articles/PMC4626409/

9. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2013;22:278–295. DOI: 10.1177/0962280210395740. https://journals.sagepub.com/doi/10.1177/0962280210395740

10. Wu A, Anderson H, Hughesman C, Young S, Lohrisch C, Ross CJD, Carleton BC. Implementation of pharmacogenetic testing in oncology: DPYD-guided dosing to prevent fluoropyrimidine toxicity in British Columbia. *Frontiers in Pharmacology*. 2023;14:1257745. DOI: 10.3389/fphar.2023.1257745. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1257745/full

11. Relling MV, Evans WE. Pharmacogenomics in the clinic. *Nature*. 2015;526:343–350. DOI: 10.1038/nature15817. https://www.nature.com/articles/nature15817