TrialGuard: Formally Verified Minimax SIPS_∞ Auditing, Manski–Rosenbaum Attrition Bounds, and Multimodal Pharmacovigilance for Phase II/III Clinical Trial Crash Detection

Fabien Pierret

Independent Researcher

ORCID: 0009-0001-1223-2170 | biotech.wallstreetdebug.com | github.com/fabien-pierret/trialguard-biotech

Preprint (camera-ready manuscript, Zenodo DOI #8 / HAL) — November 27, 2026

Abstract

Phase II/III clinical trial failures remain the dominant source of capital destruction in biopharmaceutical development, routinely triggering −50% to −85% single-session drawdowns in public biotech equities. Although recent machine-learning and LLM architectures (HINT, SPOT, TrialBench, MEXA-CTP, MMCTOP, DeepImagine, and SAE-Audit) advance multimodal clinical trial outcome prediction, they operate as unconstrained black-box neural scorers without structural single-pillar veto guarantees, nonparametric attrition bounds, or sequential family-wise error rate (FWER) auditing. We introduce TrialGuard, an open-source, formally verified four-pillar clinical trial auditing framework unifying (1) ClinicalTrials.gov API v2 endpoint-switching and subgroup-multiplicity auditing, (2) sharp nonparametric Manski attrition bounds and population 2×2 Rosenbaum hidden-bias sensitivity, (3) openFDA FAERS pharmacovigilance disproportionality (PRR) with Haldane–Anscombe smoothing, and (4) RDKit/ChEMBL physicochemical toxicity alerts, human genetic target validation, and optional research-only AlphaGenome Variant Impact (AVI) annotations. All deterministic algebraic cores—58 public theorems and 8 concrete non-vacuity witnesses—are machine-checked in standalone Lean 4.36.0 (‘0 sorry‘, ‘-DwarningAsError=true‘, kernel-native ‘Lean.collectAxioms‘ audit, standard axioms only), proving that every weighted finite-integer-p L_p marginal-risk scorecard strictly violates Fréchet's probability lower bound on single-pillar clinical fatalities, whereas the Chebyshev minimax veto (SIPS_∞) and Conjunctive Noisy-OR hybrid score $H_\lambda(r) \in [S_\infty(r), R_{\mathrm{OR}}(r)]$ guarantee zero single-pillar dilution. On a strictly causal, date-bounded retrospective benchmark of 24 real Nasdaq-listed Phase II/III biotech trials (12 historical crashes vs. 12 FDA approvals harvested live via ClinicalTrials.gov, openFDA, PubChem, ChEMBL, and RDKit), TrialGuard flags 12/12 crashes (100.0% recall, mean avoided drawdown −55.02%) at 0.0% false alarm (Brier 0.0501–0.0634), whereas equal-weight L_1 and L_2 scorecards violate Fréchet's bound on 24/24 trials and miss up to 4/12 real crashes (including −83.0% `FGEN` and −82.2% `AMLX`).

1. Introduction

Late-stage clinical development is an unforgiving conjunctive survival gate. Empirical studies across tens of thousands of industry-sponsored protocols confirm that fewer than 14% of Phase I programs reach regulatory approval and that Phase II/III attrition remains stubbornly high [29], while registry audits reveal pervasive primary outcome switching and post-hoc subgroup rescue [12]. A Phase III trial collapses if *any one* of four distinct clinical mechanisms fails: (i) a spurious Phase II signal manufactured via endpoint switching or subgroup dredging; (ii) differential patient attrition or hidden selection bias wide enough to erase the apparent Intent-to-Treat (ITT) benefit; (iii) dose-limiting organ toxicity or serious adverse events (SAEs) triggering a clinical hold or FDA Complete Response Letter (CRL); or (iv) unviable medicinal chemistry or an unvalidated biological target [3, 15, 20].

Recent multimodal prediction systems—including HINT [11], SPOT [28], TrialBench [6], MEXA-CTP [30], MMCTOP [2], CT Open [27], DeepImagine [31], LLM spin detectors [1], population-shift Bayesian probability-of-success models [13], and inference-time LLM auditing via sparse autoencoders (SAE-Audit, Nov. 2026) [26]—advance neural representation learning over protocols and chemical graphs. Yet these systems operate in unconstrained latent embedding spaces: they offer neither a formal Fréchet–minimax veto guarantee against single-pillar clinical fatalities, nor nonparametric Manski–Rosenbaum missing-data bounds, nor sequential FWER control over ClinicalTrials.gov protocol amendments, nor machine-checked Lean 4 proofs. Equally critically, many real-world investor scorecards aggregate calibrated marginal risk pillars $r = (r_1, \ldots, r_K) \in [0,1]^K$ via weighted L_1/L_2 averages, which systematically dilute single-pillar fatal flaws below alert thresholds.

Strict 3-Layer Epistemic Separation. Following our prior formal-verification methodology [23, 24], we strictly separate three layers:

1. **Layer 1 (Machine-Checked Lean 4.36.0 Core):** Exact ordered-field ($\mathbb{Q} \hookrightarrow \mathbb{R}$) and finite-probability-space proofs of Theorems 1–4 (‘lean/TrialGuardCore.lean‘, 58 public theorems, 0 ‘sorry‘, audited by kernel-native ‘Lean.collectAxioms‘ under ‘-DwarningAsError=true‘).

2. **Layer 2 (Analytical Biomedical Calibration):** Transparent, pre-specified mappings from live ClinicalTrials.gov API v2, openFDA FAERS

(with Haldane–Anscombe +0.5 smoothing), PubChem/ChEMBL + RDKit, and optional research-only AlphaGenome AVI annotations into $[0,1]^4$.

3. **Layer 3 (Empirical Retrospective Benchmark):** Reproducible Monte Carlo validation of Theorems 1–3 plus a live-harvested retrospective benchmark across 24 real Nasdaq-listed Phase II/III biotech trials, explicitly comparing strictly causal pre-readout auditing (date-bounded FAERS) against unfiltered post-marketing data.

2. Formal Mathematical Core (Layer 1: Lean 4)

Let $r = (r_1, \dots, r_K) \in [0,1]^K$ denote marginal failure-risk scores across K distinct clinical mechanisms ($K = 4$ in TrialGuard). Define the Chebyshev minimax (SIPS_∞) score $S_\infty(r) = \max_k r_k$, the Conjunctive Noisy-OR survival score $R_{\mathrm{OR}}(r) = 1 - \prod_k (1 - r_k)$ [22], the Boole upper bound $B(r) = \min(1, \sum_k r_k)$, and for $\lambda \in [0,1]$ the hybrid investor score $H_\lambda(r) = \lambda S_\infty(r) + (1-\lambda) R_{\mathrm{OR}}(r)$. Table 1 records the exact 1-to-1 correspondence between our prose theorems and 'lean/TrialGuardCore.lean'.

Theorem 1 (Fréchet–Minimax & L_p Dilution). *1. **(Finite-Space Fréchet–Boole Sandwich, 1A)** On any weighted finite probability space (N, μ) ($\mu \geq 0$, $\sum_{\omega < N} \mu(\omega) = 1$) and any $K = k+1$ correlated failure events $(F_0, \dots, F_k)$ with marginals $r_i = \mathbb{P}(F_i)$ [4, 10],*

$$S_\infty(r) \;\leq\; \mathbb{P}\left(\bigcup_{i=0}^{k} F_i\right) \;\leq\; \min\left(1, \sum_{i=0}^{k} r_i\right). \quad (1)$$

*2. **(Dimension-Free Minimax IFF, 1B–1C)** An aggregator $A : [0,1]^K \to [0,1]$ satisfies Single-Pillar Strict Exceedance Immunity ($(\exists i,\, r_i > \tau) \Rightarrow A(r) > \tau$) and Sub-Threshold Non-Escalation ($(\forall i,\, r_i \leq \tau) \Rightarrow A(r) \leq \tau$) across all $\tau \in [0,1]$ if and only if $A(r) = S_\infty(r)$.*

*3. **(Universal Marginal-L_p Fréchet Violation, 1D–1F)** For every $K \geq 2$, integer exponent $p \geq 1$, strictly positive simplex weights $(w_i)_{i<K}$ ($\sum_{i<K} w_i = 1$), and single-pillar hazard $e_j(v)$ ($v > 0$), $\sum_{i<K} w_i (e_j(v))_i^p = w_j v^p < v^p = S_\infty(e_j(v))^p$. Hence every weighted marginal-L_p scorecard strictly violates Fréchet's lower bound and silently misses a maximal single-pillar crash ($v = 1$) for any threshold $\tau > 0$ with $w_j < \tau^p$ (e.g. $\tau > 0.25$ for L_1, $\tau > 0.50$ for L_2 at uniform $K = 4$).*

*4. **(Conjunctive Noisy-OR & Hybrid Score, 1G–1H)** For all $r \in [0,1]^K$ and $\lambda \in [0,1]$, $S_\infty(r) \leq H_\lambda(r) \leq R_{\mathrm{OR}}(r) \leq B(r)$, with single-pillar fatality preservation ($r_j = 1 \Rightarrow R_{\mathrm{OR}}(r) = H_\lambda(r) = 1$) and clean-trial calibration ($r = 0 \Rightarrow R_{\mathrm{OR}}(0) = H_\lambda(0) = 0$).*

Remark 1 (Positive Inter-Pillar Dependence & Neural Baselines). Clinical failure pillars are positively correlated in biology (lipophilic toxicity in Pillar 4 causes SAEs in Pillar 3 and treatment dropout in Pillar 2). Positive dependence places $\mathbb{P}(\bigcup F_k)$ inside $[S_\infty(r), R_{\mathrm{OR}}(r)]$, directly motivating $H_{0.5}(r)$ scaled to $[0, 100]$. We carefully distinguish marginal-risk L_p scorecards (which Theorem 1(3) directly refutes) from latent-space neural/LLM models (HINT, MMCTOP, SAE-Audit), which operate in unbounded logit spaces and lack any formal Fréchet–minimax veto guarantee.

Theorem 2 (Manski Attrition & Rosenbaum Sensitivity). *Assume initial randomization with bounded outcome $Y \in [a, b]$ ($a \leq b$, $M = b - a \geq 0$), translated to $[0, M]$ via $\tilde{Y} = Y - a$ (higher-is-better) or $\tilde{Y} = b - Y$ (lower-is-better clinical scales such as ADAS-Cog). Given completer means $\mu_T, \mu_C \in [0, M]$ and arm attrition $d_T, d_C \in [0,1]$ [14, 19]:*

*1. **(Sharp ITT Containment & Shift Invariance, 2A–2C)** For all latent missing-patient means $y_T^{\mathrm{m}}, y_C^{\mathrm{m}} \in [0, M]$, the true ITT effect Δ_{ITT} satisfies $\Delta_{\mathrm{LB}} \leq \Delta_{\mathrm{ITT}} \leq \Delta_{\mathrm{UB}}$, where*

$$\Delta_{\mathrm{LB}} = (1 - d_T)\mu_T - (1 - d_C)\mu_C - d_C M, \quad (2)$$
$$\Delta_{\mathrm{UB}} = (1 - d_T)\mu_T + d_T M - (1 - d_C)\mu_C. \quad (3)$$

Both bounds are attained in $[0, M]^2$ by $(0, M)$ and $(M, 0)$, have exact width $(d_T + d_C)M$, contain $\mu_T - \mu_C$ weakly, and yield the exact sign-reversal criterion $\Delta_{\mathrm{LB}} \leq 0 \iff (1 - d_T)\mu_T \leq (1 - d_C)\mu_C + d_C M$.

*2. **(Population 2×2 Rosenbaum Collapse, 2E–2G)** For non-randomized comparisons (single-arm trials vs. external controls or post-randomization responder/adherence subgroups subject to collider bias) under Rosenbaum's hidden-bias odds bound $\Gamma \geq 1$ [25] and sharp null $Y(1) = Y(0)$, Bayes' rule yields $p_{\mathrm{RB}}(p, \Gamma) = \frac{p}{p + \Gamma(1-p)}$ with strictly positive denominator, $p_{\mathrm{RB}}(p, 1) = p$, strict antitonicity in Γ, and for $0 < p_C < p_T < 1$, critical threshold $\Gamma^* = \frac{p_T(1-p_C)}{p_C(1-p_T)} > 1$ satisfying $p_{\mathrm{RB}}(p_T, \Gamma) \leq p_C \iff \Gamma \geq \Gamma^*$.*

Remark 2 (Baseline Subgroups vs. Hidden Allocation Bias). Baseline-covariate subgroup dredging inside an RCT preserves randomization within strata and creates multiplicity / winner's curse (audited in Pillar 1 / Theorem 3 [18, 21]), whereas single-arm external controls and post-randomization adherence/responder subgroups introduce hidden selection bias $\Gamma > 1$ (audited via Theorem 2(2)). Finite-sample partial-identification confidence intervals [16] and Mantel–Haenszel bounds belong to Layer 2.

Theorem 3 (Switching FWER & α-Spending). *1. **(Uncorrected Multiplicity Explosion, 3A–3C)** Across $M \in \mathbb{N}$ independent null tests at nominal $\alpha \in [0,1]$, $\mathrm{FWER}_{\mathrm{uncorr}}(M, \alpha) = 1 - (1 - \alpha)^M$ satisfies the switch-amplification identity*

$$1 - \mathrm{FWER}_{\mathrm{uncorr}}(M + \Delta m, \alpha) = \big(1 - \mathrm{FWER}_{\mathrm{uncorr}}(M, \alpha)\big)(1 - \alpha)^{\Delta m}$$

and the elementary Bernoulli-type lower bound $\frac{M\alpha}{1 + M\alpha} \leq \mathrm{FWER}_{\mathrm{uncorr}}(M, \alpha)$.

*2. **(2D Frozen-Protocol Geometric Alpha-Spending, 3D–3F)** Anchored to initial registration*

Table 1. Exact 1-to-1 correspondence between paper theorems and machine-checked Lean 4.36.0 declarations (`lean/TrialGuardCore.lean`, 58 public theorems, 0 `sorry`, kernel-native `Lean.collectAxioms` audit).

Paper	Verified Lean 4 Declaration(s)
Thm 1A	`frechet_boole_abstract_sandwich`, `frechet_boole_finite_space_sandwich`
Thm 1B–C	`clinical_gate_iff_minimax4`, `minimax4_unique_iff`, `clinical_gate_iff_maxNat`, `maxNat_unique_iff`
Thm 1D–F	`lp_violates_frechet_single_pillar4`, `lp_uniform_blindness4`, `lp_violates_frechet_arbitrary_K`, `lp_blindness_arbitrary_K`
Thm 1G–H	`noisyOrNat_frechet_boole_sandwich`, `noisyOr4_frechet_boole_sandwich`, `hybridScore4_sandwich`, `hybridScore4_fatality_veto`
Thm 2A–D	`manski_bounds_contain_true_itt`, `manski_bounds_sharp`, `manskiLB_ab_shift_eq`, `manski_exact_width`, `manski_sign_reversal_iff`
Thm 2E–G	`rosenbaum_denom_pos`, `rosenbaum_bounds`, `rosenbaum_strict_antitone`, `rosenbaum_critical_gamma_gt_one`, `rosenbaum_sign_reversal_iff`
Thm 3A–C	`fwerUncorr_strict_mono`, `fwerUncorr_switch_amplification`, `fwerUncorr_lower_bound`
Thm 3D–F	`cumAlphaSpend_closed_form`, `cumAlphaSpend_strict_budget`, `bonferroni_double_sum_fwer_control`
Thm 4A–C	`faersPRR_ge_iff`, `faersPRR_strict_mono_a`, `faersPRR_strict_anti_b`
Prop 4D	`clip01_in_unit`, `clip01_monotone`, `clipped_additive_pillar_mono`
Witness 4E	8 concrete rational witnesses mechanically invoking conditional theorems

$\mathcal{P}_0$ with budget $\alpha_0 > 0$ and retention $\gamma \in (0,1)$, allocate $\alpha_t = \alpha_0(1-\gamma)\gamma^t$ across amendment rounds $t \in \{0,\dots,T-1\}$ split across $m_t \geq 1$ endpoints [9, 17]. Whenever per-endpoint null error rates obey $0 \leq e_{t,j} \leq \alpha_t/m_t$ $(j < m_t)$, for all $T \geq 0$,

$$0 \;\leq\; \sum_{t=0}^{T-1} \sum_{j=0}^{m_t-1} e_{t,j} \;\leq\; \alpha_0(1-\gamma^T) \;<\; \alpha_0. \quad (4)$$

Theorem 4 (FAERS PRR & Additive Monotonicity)**.** *For any 2×2 pharmacovigilance table (a,b,c,d) (target event a, target other events b, background event c, background other d [8]) with $a+b>0, c>0, d \geq 0$, $\mathrm{PRR}(a,b,c,d) = \frac{a/(a+b)}{c/(c+d)} \geq \theta \iff \theta c(a+b) \leq a(c+d)$ for all $\theta \in \mathbb{Q}$, strictly increasing in a $(b>0)$ and strictly decreasing in b $(a>0)$. Furthermore, for any nonnegative weights $a_i \geq 0$ and monotone risk features $c_{1,i} \leq c_{2,i}$, injecting $\mathrm{clip}_{[0,1]}(\sum_{i<M} a_i c_i)$ into any coordinate of r preserves weak monotonicity across $S_\infty(r)$, $R_{\mathrm{OR}}(r)$, and $H_\lambda(r)$ (Proposition 4D).*

3. Layer 2: Live Biomedical Connectors & Calibration

Our MIT Python package `trialguard` and Cloudflare edge service (biotech.wallstreetdebug.com) operationalize the four pillars from authoritative public sources:

- **Pillar 1 (r_1, ClinicalTrials.gov API v2):** Registered primary/secondary outcomes, Šidák multiplicity inflation $1-(1-0.05)^{M_{\mathrm{eff}}}$, protocol switching, open-label/single-arm extrapolation, and post-hoc subgroup rescue.
- **Pillar 2 (r_2, Manski–Rosenbaum Engine):** Arm dropout rates (d_T, d_C) mapped via normalized Manski width $\frac{d_T+d_C}{2}$, normalized sign-reversal deficit $\max(0, -\Delta_{\mathrm{LB}}/M)$, and clipped Rosenbaum vulnerability $\min(1, 1/\Gamma^*)$ for non-randomized or post-randomization subgroup comparisons.
- **Pillar 3 (r_3, openFDA FAERS + ClinicalTrials.gov SAE):** Haldane–Anscombe +0.5 smoothed PRR (handling sparse pre-approval cells cleanly under Theorem 4) with Evans' $a \geq 3$ floor [8], empirical-Bayes shrinkage spirit [7], and trial-level SAE imbalance.
- **Pillar 4 (r_4, RDKit + ChEMBL + PubChem + Optional AlphaGenome AVI):** Small molecules are scored via RDKit descriptors (Pfizer 3/75 zone cLogP $> 3 \wedge$ TPSA $< 75\,\text{Å}^2$ [15], moderately weighted PAINS/Brenk alerts acknowledging false-positive caveats [3, 5], Lipinski violations, QED) plus ChEMBL warnings. Biologics, RNAi/mRNA, and gene therapies bypass small-molecule Lipinski penalties and are scored via human genetic target validation [20]. When an `ALPHAGENOME_API_KEY` and causal target variant are supplied, high AlphaGenome AVI Phred scores ($\geq$ 20, Top 1%) act strictly as research-only molecular/functional annotations reinforcing target validation (zero clinical diagnosis or medical advice); absent keys remain strictly neutral.

4. Empirical Verification & Real Nasdaq Benchmark (Layer 3)

4.1. Synthetic Stress Tests of Theorems 1–3

Figure 1 evaluates 1,500 single-pillar clinical hazards (one pillar $\in [0.75, 1.0]$, three clean $\in [0.02, 0.12]$) and 500 clean trials at Red alert threshold $\tau = 0.60$. Linear L_1 and quadratic L_2 suffer 100.0% single-pillar bypass (L_3: 80.9%) and violate Fréchet's lower bound on 100.0% of hazards, whereas TrialGuard SIPS_∞, Noisy-OR, and Hybrid $H_{0.5}$ achieve 0.0% bypass, 0.0% false alarm, and 0.0% Fréchet violation. Figure 2 maps the exact Manski attrition sign-reversal boundary and population 2×2 Rosenbaum collapse across real cohort profiles (`SAVA` $\Gamma^* = 1.62$, `ATHA` $\Gamma^* = 1.27$ vs. `MDGL` $\Gamma^* = 4.20$, `VRTX` $\Gamma^* = 20.75$). In Figure 3 (500 Monte Carlo null trials, $\alpha_0 = 0.05$), uncorrected endpoint switching explodes to 47.0% FWER at $M = 14$ (exact 51.2%, Lean lower bound 41.2%) and 62.4% at $M = 20$, while Frozen-Protocol Geometric Alpha-Spending ($\gamma = 0.6$) holds empirical FWER at 3.2% ($< 5.0\%$ strict ceiling).

4.2. Retrospective Benchmark on 24 Real Nasdaq Biotech Phase II/III Trials

We harvested live records on November 27, 2026 via ClinicalTrials.gov API v2, openFDA FAERS, PubChem, ChEMBL, and RDKit across 24 iconic Nasdaq biotech Phase II/III programs: 12 historical crashes (`SAVA`, `BIIB`,

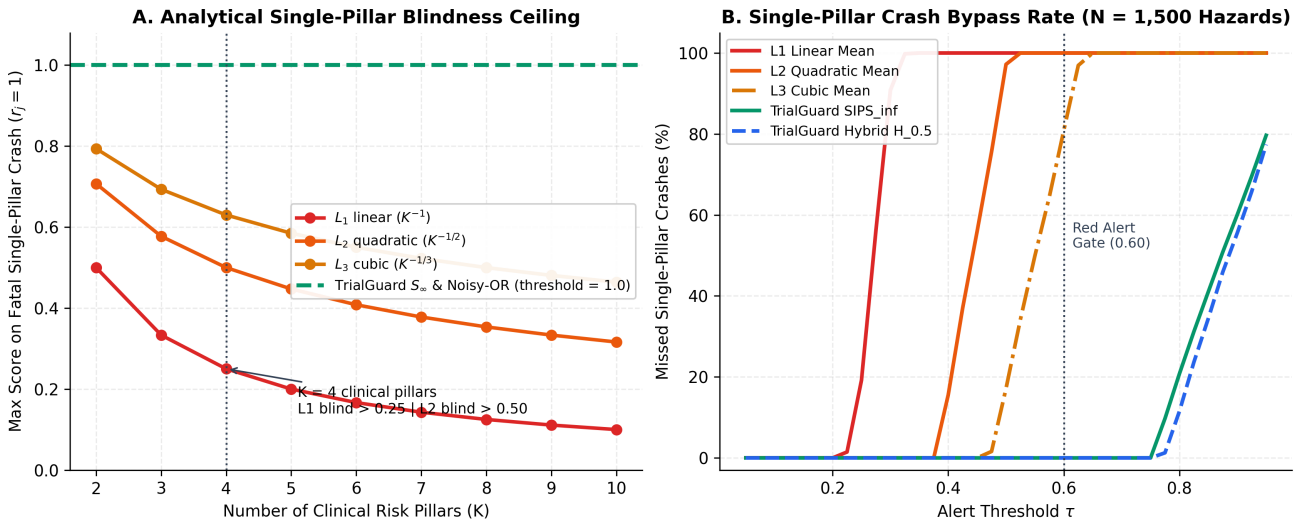

Figure 1. Experiment 1 (Theorem 1). Analytical single-pillar blindness ceiling $K^{-1/p}$ (left) and empirical crash bypass across 1,500 single-pillar clinical hazards (right). At Red gate $\tau = 0.60$, L_1 and L_2 miss 100.0% of single-pillar fatalities, while TrialGuard SIPS$_\infty$ and Hybrid $H_{0.5}$ achieve 0.0% bypass and 0.0% clean false alarm.

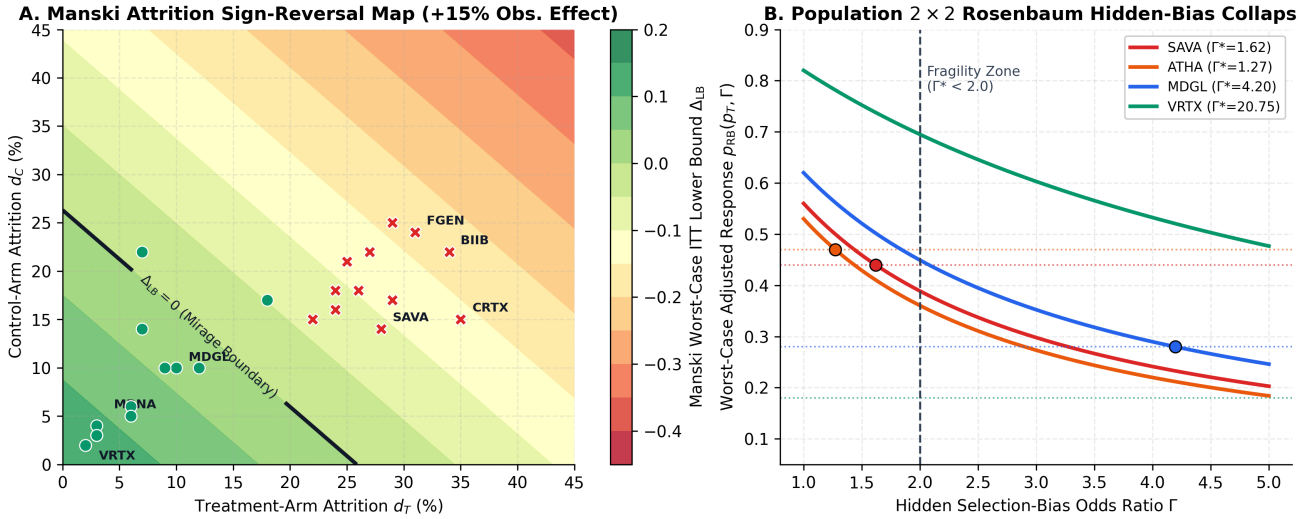

Figure 2. Experiment 2 (Theorem 2). Sharp nonparametric Manski attrition sign-reversal contour with overlaid Nasdaq Phase II/III cohort dropout rates (left) and population 2 × 2 Rosenbaum hidden-bias collapse curves $p_{\mathrm{RB}}(p_T, \Gamma)$ with exact critical thresholds Γ^* (right).

FGEN, CRTX, AMLX, ICPT, ATHA, PRTA, NKTR, INCY, CERE, ALKS; event moves −12.6% to −83.8%, mean −55.02%) and 12 landmark FDA approvals (VRTX, REGN, ALNY, MDGL, KRTX, NBIX, KRYS, ARGX, MRNA, BPMC, BMRN, EXEL; event moves +7.4% to +268.1%, mean +54.49%).

To enforce **zero future leakage**, our primary *Strictly Pre-Readout* evaluation filters openFDA FAERS strictly to `receivedate:[20040101 TO (readout_year-1)1231]`, excludes ClinicalTrials.gov `overallStatus` and post-results fields, and excludes post-approval ChEMBL boxed warnings. As summarized in Table 2 and Figures 4–6:

- **TrialGuard** SIPS$_\infty$ **and Hybrid** $H_{0.5}$ flag **12/12 real crashes** (100.0% **recall)** at **0/12 false alarms** (0.0%) at the Red gate (60/100), achieving Brier scores of 0.0501 (SIPS$_\infty$) and 0.0634 ($H_{0.5}$), with a passed-portfolio mean event return of +54.49%.
- **Weighted** L_1 **linear aggregation** on the *exact same* four pillars violates Fréchet's bound on 24/24 trials and misses **4/12 crashes** (66.7% **recall)**—silently passing FGEN (−83.0%), AMLX (−82.2%), BIIB (−29.2%), and INCY (−22.9%), cutting passed-portfolio mean return in half to +27.29%.
- **Weighted** L_2 **quadratic aggregation** violates Fréchet's bound on 24/24 trials and misses FGEN (−83.0%, 91.7% recall; passed return +43.92%).

5. Limitations, Temporal Leakage & Responsible Use

Our Regime B comparison exposes an essential empirical pitfall in pharmacovigilance backtesting: when unfiltered 2026 openFDA FAERS reports are applied retrospectively to trials completed years earlier, approved drugs treating fatal diseases (e.g., cirrhotic MASH in MDGL with 269 post-2024 reports, or renal cell carcinoma progression in EXEL) accumulate post-marketing deaths that

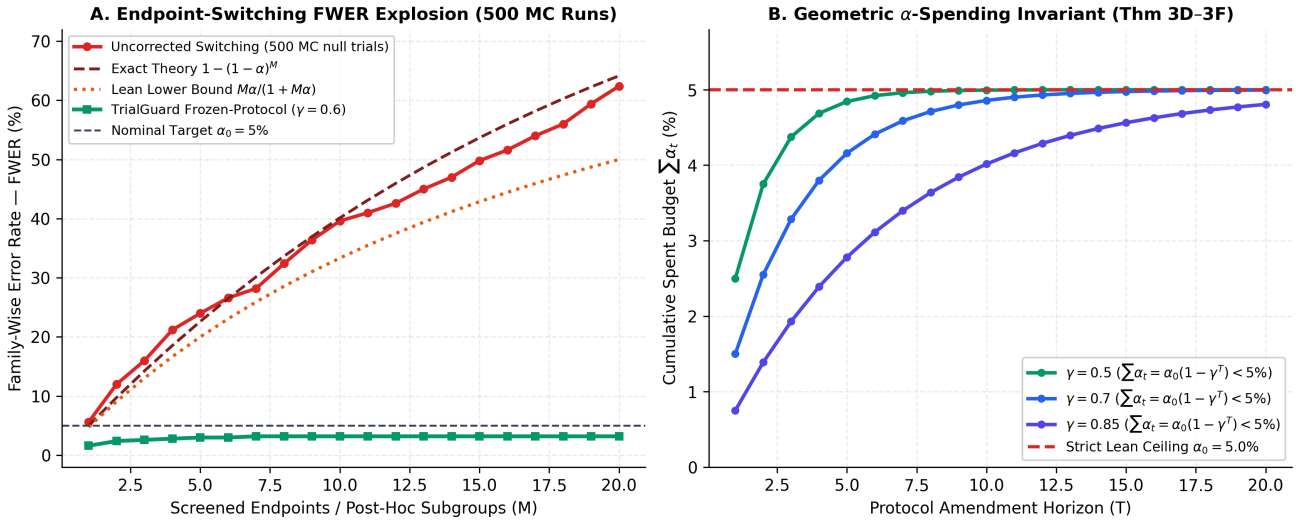

Figure 3. Experiment 3 (Theorem 3). 500 Monte Carlo null trials showing FWER explosion under uncorrected endpoint/subgroup switching (62.4% at $M = 20$) versus Frozen-Protocol Geometric Alpha-Spending holding 3.2% $< 5\%$ (left) and exact infinite-horizon budget invariance $\sum \alpha_t = \alpha_0(1-\gamma^T) < \alpha_0$ (right).

Table 2. Strictly causal pre-readout performance across 24 real Nasdaq Phase II/III trials (12 crashes vs. 12 FDA approvals) at Red gate $\tau = 0.60$, plus unfiltered 2026 post-marketing comparison.

Aggregator	Crash Recall (/12)	False Alarm (/12)	Brier Score	Fréchet Viol.	Passed Ret. (%)
Regime A: Strictly Pre-Readout (Date-Bounded openFDA FAERS, Zero Future Leakage)					
L_1 Linear Mean	66.7% (8)	0.0% (0)	0.0778	24 / 24	+27.29%
L_2 Quadratic Mean	91.7% (11)	0.0% (0)	0.0705	24 / 24	+43.92%
TrialGuard SIPS_∞	**100.0% (12)**	**0.0% (0)**	**0.0501**	**0 / 24**	**+54.49%**
TrialGuard Hybrid $H_{0.5}$	**100.0% (12)**	**0.0% (0)**	0.0634	**0 / 24**	**+54.49%**
Noisy-OR R_{OR}	100.0% (12)	8.3% (1)	0.0925	0 / 24	+54.71%
Regime B: Unfiltered 2026 Post-Marketing FAERS + Live Results (Confounding Audit)					
TrialGuard Hybrid $H_{0.5}$	100.0% (12)	16.7% (2)	0.1125	0 / 24	+33.37%

raise false alarms from 0.0% to 16.7% (2/12). Strictly causal date cutoffs and disease-severity adjustment are therefore mandatory. Additionally, our 24-trial cohort serves as a transparent, reproducible benchmark rather than an exhaustive universe of all historical biotech readouts; prospective validation on upcoming Phase III readouts (e.g., aligned with CT Open [27]) is planned next. TrialGuard is strictly a statistical research and risk-auditing instrument: it provides neither investment advice nor clinical/medical diagnosis.

6. Conclusion

By coupling a 58-theorem, kernel-audited Lean 4.36.0 algebraic core ('0 sorry') with live ClinicalTrials.gov API v2, openFDA FAERS, PubChem, ChEMBL, and RDKit connectors, TrialGuard bridges formal mathematical guarantees and investor-grade clinical trial due diligence. All code, Lean proofs, harvested datasets, and scripts are released under the MIT license at github.com/fabien-pierret/trialguard-biotech, with live edge auditing at biotech.wallstreetdebug.com.

References

[1] Tjaš Ajdovec, Marko Robnik-Šikonja, and Simon Šuster. Detecting spin in clinical trials with large language models. *arXiv preprint arXiv:2610.11845*, 2026.

[2] Carolina Aparício, Qi Shi, Bo Wen, and Tesfaye Yadete. MMCTOP: A multimodal textualization and mixture-of-experts framework for clinical trial outcome prediction. *arXiv preprint arXiv:2512.21897*, 2025.

[3] Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010. doi: 10.1021/jm901137j.

[4] George Boole. *An Investigation of the Laws of Thought*. Walton and Maberly, 1854.

[5] Stephen J. Capuzzi, Eugene N. Muratov, and Alexander Tropsha. Phantom PAINS: Problems with the utility of alerts for pan-assay INterference compoundS. *Journal of Chemical Information and Modeling*, 57(3):417–427, 2017. doi: 10.1021/acs.jcim.6b00465.

[6] Jintai Chen, Yaojun Hu, Mingchen Cai, Yingzhou Lu, Yue Wang, and Jimeng Sun. TrialBench: Multimodal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631*, 2024.

[7] William DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999.

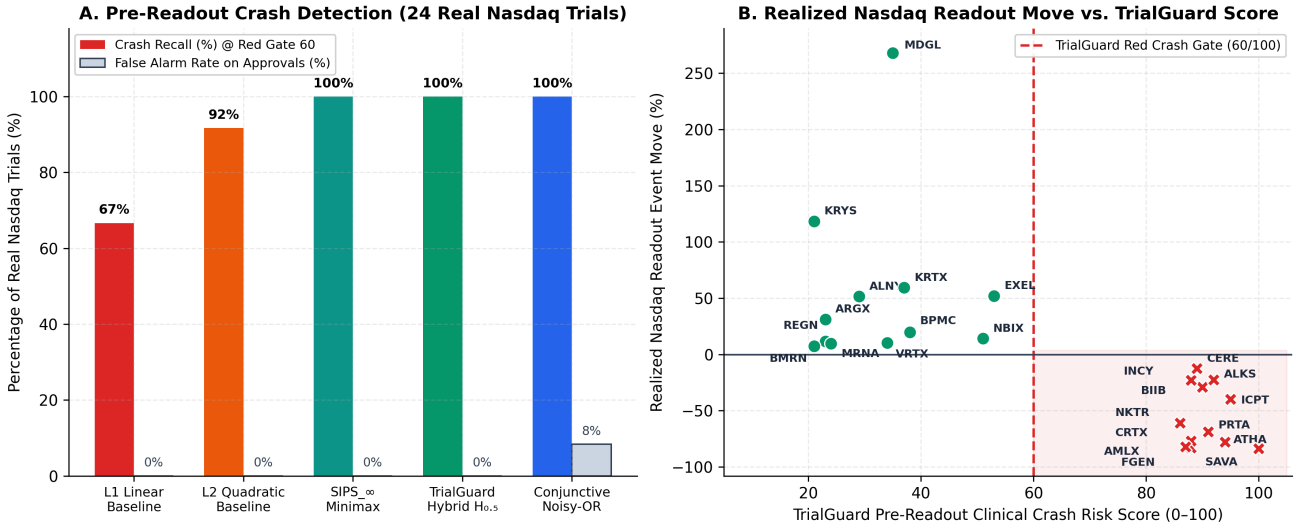

Figure 4. Experiment 4 (24 Real Nasdaq Phase II/III Trials). Strictly pre-readout crash recall vs. false alarm at Red gate 60/100 (left) and realized single-session Nasdaq readout move (%) vs. TrialGuard score showing complete quadrant separation (right).

[8] Stephen J. W. Evans, Patrick C. Waller, and Stewart Davis. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486, 2001.

[9] Dean P. Foster and Robert A. Stine. α-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B*, 70(2):429–444, 2008.

[10] Maurice Fréchet. Généralisation du théorème des probabilités totales. *Fundamenta Mathematicae*, 25: 379–387, 1935.

[11] Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445, 2022. doi: 10.1016/j.patter.2022.100445.

[12] Ben Goldacre, Henry Drysdale, Aaron Dale, Ioan Milosevic, Eirion Slade, Philip Hartley, Cicely Marston, Anna Powell-Smith, Carl Heneghan, and Kamal R. Mahtani. COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20:118, 2019. doi: 10.1186/s13063-019-3172-3.

[13] Xinyi He, Jianchang Lin, Rachael Liu, and Ying Yuan. Probability of success under population shift for phase III trial planning. *arXiv preprint arXiv:2611.19633*, 2026.

[14] Joel L. Horowitz and Charles F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84, 2000.

[15] Jason D. Hughes, Julian Blagg, David A. Price, Simon Bailey, Gary A. DeCrescenzo, Rajesh V. Devraj, Edmund Ellsworth, Yvette M. Fobian, Michael E. Gibbs, Richard W. Gilles, Nigel Greene, Enoch Huang, Teresa Krieger-Burke, Jens Loesel, Travis Wager, Larry Whiteley, and Yao Zhang. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters*, 18(17):4872–4875, 2008. doi: 10.1016/j.bmcl.2008.07.071.

[16] Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

[17] K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[18] Larry F. Leon and Keaven M. Anderson. Inference for standard trial estimands after data-driven subgroup discovery. *arXiv preprint arXiv:2609.38361*, 2026.

[19] Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

[20] Eric Vallabh Minikel, Jeffery L. Painter, Coco Chengliang Dong, and Matthew R. Nelson. Refining the impact of genetic evidence on clinical success. *Nature*, 629:624–629, 2024. doi: 10.1038/s41586-024-07316-0.

[21] Manuel M. Müller, Björn Bornkamp, Frank Bretz, Timothy I. Cannings, Wei Liu, Henry W. J. Reeve, Richard J. Samworth, Nikolaos Sfikas, Fang Wan, and Konstantinos Sechidis. Data-driven controlled subgroup selection in clinical trials. *arXiv preprint arXiv:2512.15676*, 2025.

[22] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[23] Fabien Pierret. Proof-CI: Necessary and sufficient conditions for formally verified minimax evaluators

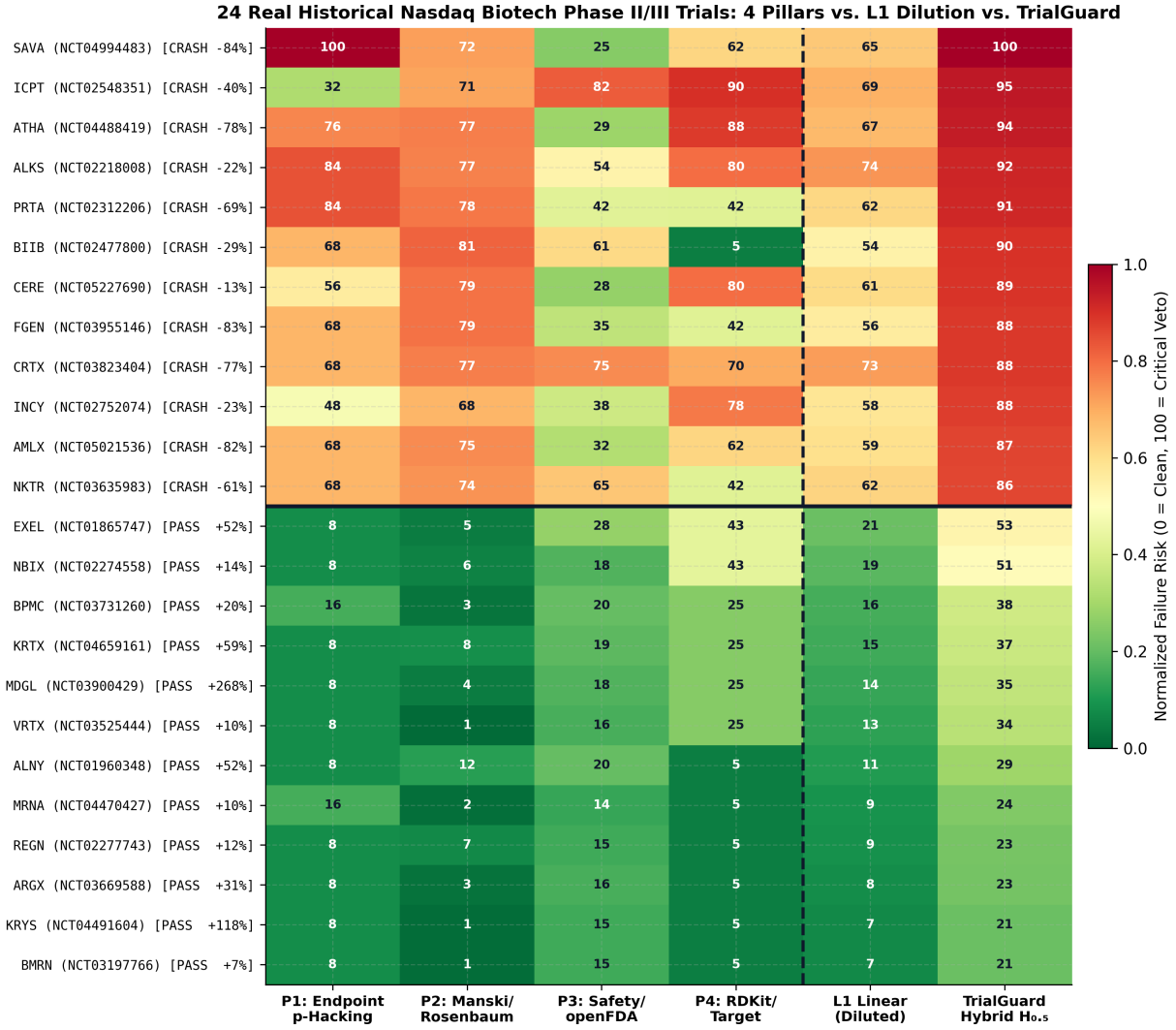

Figure 5. Experiment 5 (Full 24-Trial Multimodal Audit Matrix). Per-pillar risks (r_1–r_4) contrasted with diluted L_1 linear means (which silently pass `FGEN`, `AMLX`, `BIIB`, and `INCY` below 60) and TrialGuard Hybrid $H_{0.5}$.

preventing drift and reward hacking in recursive self-improvement. *CERN Zenodo & HAL*, 2026. doi: 10.5281/zenodo.24614087.

[24] Fabien Pierret. Minimax SIPS$_\infty$ auditing and post-selection bounds for high-stakes statistical evaluation. *CERN Zenodo*, 2026. doi: 10.5281/zenodo.24591422.

[25] Paul R. Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

[26] Junbo Shen. SAE-Audit: Clinical-trial prediction via inference-time optimization of black-box llms audited by sparse autoencoder evidence. *arXiv preprint arXiv:2611.29368*, 2026.

[27] Jianyou Wang, Youze Zheng, Longtian Bao, and Hanyuan Zhang. CT Open: An open-access, uncontaminated, live platform for the open challenge of clinical trial outcome prediction. *arXiv preprint arXiv:2604.16742*, 2026.

[28] Zifeng Wang, Cao Xiao, and Jimeng Sun. SPOT: Sequential predictive modeling of clinical trial outcome with meta-learning. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 2023.

[29] Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019. doi: 10.1093/biostatistics/kxx069.

[30] Yiqing Zhang, Xiaozhong Liu, and Fabricio Murai. MEXA-CTP: Mode experts cross-attention for clinical trial outcome prediction. *arXiv preprint arXiv:2501.06823*, 2025.

[31] Youze Zheng, Jianyou Wang, Yuhan Chen, and Matthew Feng. DeepImagine: Clinical trial outcome prediction via stepwise local counterfactual imaginations. *arXiv preprint arXiv:2604.23054*, 2026.

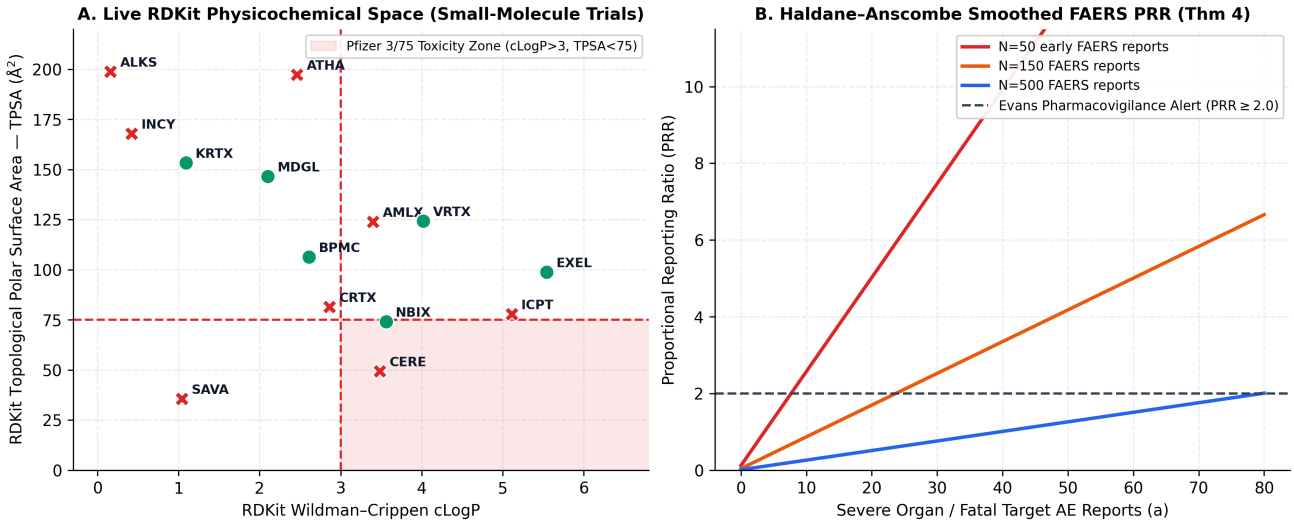

Figure 6. Experiment 6 (Live Cheminformatics & Pharmacovigilance). Live RDKit Wildman–Crippen cLogP vs. TPSA across harvested small-molecule Nasdaq trials with Pfizer 3/75 toxicity zone (left) and Haldane–Anscombe smoothed openFDA FAERS PRR sensitivity curves (right).