ClinOps AI: Deep Learning for Clinical Trial Safety Intelligence

Francesco Orsi*

February 17, 2026

Abstract

We present ClinOps AI, an end-to-end analytical pipeline that applies modern deep learning to clinical trial safety monitoring using real FDA-standard CDISC SDTM data. Working with the PHUSE CDISCPILOT01 dataset (N=254, Xanomeline vs Placebo, Alzheimer's disease), we formalize and implement a six-stage pipeline: pure-Python SAS Transport parsing, Pydantic v2 data validation, DuckDB analytical queries, nine publication-quality clinical visualizations, a deep learning suite comprising conditional variational autoencoder augmentation, self-supervised masked feature pretraining, and MLP classification with Monte Carlo Dropout uncertainty, and interpretable feature attribution via Integrated Gradients. An 8-model ablation study under 5-fold stratified cross-validation demonstrates that self-supervised pretraining contributes $\Delta\text{AUROC} = +0.067$ and generative augmentation adds $+0.021$ incrementally. However, ℓ_2-regularized logistic regression (AUROC=0.690) remains competitive at this sample size. We identify and correct a critical data leakage issue, provide honest assessment of the extreme class imbalance (3/254 SAE subjects), and demonstrate that neural interpretability tools — specifically attention weights and Integrated Gradients — reveal clinically novel multivariate temporal patterns invisible to traditional pharmacovigilance.

Keywords: clinical trials, pharmacovigilance, deep learning, CDISC SDTM, adverse events, interpretable ML.

1. Introduction

Clinical trials generate safety data at unprecedented scale. A single Phase III oncology trial may enroll 3,000+ subjects across 200 sites in 30 countries, producing tens of thousands of adverse event (AE) records, millions of laboratory measurements, and increasingly continuous vital sign streams from wearable devices. Yet the analytical toolkit applied to this data has not fundamentally changed since the ICH E9 guideline was finalized in 1998.

Medical monitors, Data Safety Monitoring Boards (DSMBs), and regulatory reviewers still rely on three primary instruments: frequency tables ("25% on drug vs 8% on placebo reported nausea" — no temporal dimension, no multivariate structure), patient listings (individual narratives sorted by seriousness, which do not scale beyond dozens of subjects), and pre-specified Tables, Listings, and Figures (TLFs) locked into statistical analysis plans months before unblinding, precluding any adaptive or exploratory analysis.

These methods are regulatory-compliant and well-understood. However, they leave four categories of safety questions systematically unanswered:

*Corresponding Author. E-mail: francesco.orsi84@gmail.com

Unresolved Safety Questions in Current Pharmacovigilance

1. **Temporal clustering.** When do adverse events concentrate within the study timeline? Frequency tables aggregate across the entire study period, obscuring acute onset patterns, cumulative dose effects, and organ-specific latency differences.
2. **Multivariate signals.** Pruritus from a transdermal patch is expected. But does pruritus *combined with* dizziness *combined with* weight decrease in the same patient predict a serious event? Univariate disproportionality analyses cannot detect feature interactions.
3. **Individual risk prediction.** A subject has experienced 5 mild adverse events in 6 weeks. Is this a harbinger of a serious adverse event, or simply a high-reporter phenotype? Without predictive modeling, the medical monitor relies on intuition.
4. **Per-patient explainability.** If a safety concern is escalated to the DSMB, mechanistic reasoning is required — not a global performance metric, but an explanation of *why this patient* is flagged *based on which features*.

Modern deep learning can address all four — temporal modeling via recurrent architectures, multivariate pattern detection via representation learning, individual-level prediction via subject-level classifiers, and feature attribution via gradient-based interpretability methods. The challenge is applying these methods with the rigor that clinical data demands: proper validation protocols, systematic leakage prevention, honest uncertainty quantification, and transparent limitation reporting.

This paper presents ClinOps AI, a complete analytical pipeline that demonstrates what modern ML can — and cannot — achieve on real clinical trial safety data.

2. Data: PHUSE CDISC Pilot Study

2.1 Source, Provenance, and Licensing

This project uses the **PHUSE CDISC Pilot Study (CDISCPILOT01)**, a publicly available clinical trial dataset maintained by the Pharmaceutical Users Software Exchange (PHUSE). The study evaluated Xanomeline transdermal system at two dose levels (54 mg/day and 81 mg/day) versus Placebo in 254 subjects with mild-to-moderate Alzheimer's disease. The data was originally submitted to the U.S. Food and Drug Administration as part of a CDISC pilot program demonstrating the feasibility of the Study Data Tabulation Model (SDTM) for electronic regulatory submissions.

Data Access, Licensing, and Privacy Statement

Repository: https://github.com/phuse-org/phuse-scripts/tree/master/data/sdtm/cdiscpilot01
Format: SAS Transport v5 (XPT), the FDA-mandated electronic submission format per 21 CFR Part 11.
License: Publicly available under the PHUSE open-source initiative for educational and research purposes. No patient-level consent restrictions apply, as the data has been fully anonymized and released for public use.
Privacy: All subject identifiers are synthetic study-assigned codes (e.g., "01-701-1015"). No names, dates of birth, geographic identifiers, or other HIPAA-defined Protected Health Information (PHI) are present. The data cannot be linked to real individuals.
Limitations of use: This is a pilot/demonstration dataset. While it follows FDA submission standards and contains real clinical measurements, it should not be used to draw clinical conclusions about Xanomeline's actual safety or efficacy. All analyses in this paper are methodological demonstrations, not clinical evaluations.

2.2 SDTM Domain Structure

The dataset comprises six SDTM domains, each stored as a separate XPT binary file:

Table 1: SDTM domains used in this project.

Domain	Description	Records	Key Variables
DM	Demographics	254	AGE, SEX, RACE, ARM, RFSTDTC
AE	Adverse Events	1,191	AEDECOD, AEBODSYS, AESEV, AESER, AESTDY
EX	Exposure	591	EXDOSE, EXSTDTC, EXENDTC, EXROUTE
VS	Vital Signs	6,208	VSTESTCD, VSSTRESN, VISITNUM
LB	Laboratory	13,988	LBTESTCD, LBSTRESN, LBNRIND
DS	Disposition	254	DSDECOD, DSTERM, DSSTDTC

2.3 Study Design and Population

Table 2: Study arms and baseline characteristics.

	Placebo	Xano Low (54 mg)	Xano High (81 mg)
N enrolled	86	84	84
Mean age (yr)	∼75	∼75	∼75
% Female	∼53	∼50	∼55
Total AE records	202	456	533
SAE subjects	0	1	2

A critical constraint: only $N_1 = 3$ subjects out of $N = 254$ (1.2%) experienced serious adverse events. This extreme class imbalance ratio of $N_0/N_1 \approx 84$ is the central challenge for any predictive modeling effort on this dataset and motivates both the generative augmentation and the honest limitation reporting throughout this work.

3. Pipeline Architecture

The system comprises six stages, each with a well-defined input, transformation, and output:

Table 3: End-to-end pipeline: six stages from raw binary files to clinical decisions.

#	Stage	Technology	Output
1	XPT parsing	Pure-Python (`struct`, `BytesIO`)	Polars DataFrames
2	Validation	Pydantic v2 SDTM models	Conformance report
3	SQL analytics	DuckDB (zero-copy on Polars)	Clinical summary tables
4	Visualization	matplotlib, seaborn, scipy	9 figures (PDF/PNG, 300 DPI)
5	Modeling	PyTorch (CVAE, MLP, GRU)	8-model ablation with CI
6	Interpretability	Integrated Gradients, attention	Per-subject explanations

Stages 1–3 handle data ingestion. The XPT parser decodes IBM 360 floating-point representation and 80-byte fixed-width header records in approximately 60 lines of Python, eliminating the $15,000/year SAS license dependency. Pydantic v2 models enforce CDISC business rules at parse time (e.g., if `AESER` = 'Y', then `AEOUT` $\neq$ `NULL`). DuckDB provides zero-copy SQL execution directly on Polars DataFrames.

4. Statistical Methods for Clinical Analytics

4.1 Disproportionality Analysis (Volcano Plot)

For each AE preferred term $j \in \{1, \ldots, J\}$, we construct the 2×2 contingency table comparing the High Dose arm to Placebo and test the null hypothesis $H_0 : \pi_j^{\text{drug}} = \pi_j^{\text{placebo}}$ using Fisher's exact test.

Definition 1: Fisher's Exact Test for AE Disproportionality

Let a_j, b_j, c_j, d_j denote the cell counts:

$$\begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix} = \begin{pmatrix} \text{Drug} \cap \text{AE}_j & \text{Placebo} \cap \text{AE}_j \\ \text{Drug} \cap \overline{\text{AE}_j} & \text{Placebo} \cap \overline{\text{AE}_j} \end{pmatrix} \tag{1}$$

The exact p-value under the hypergeometric distribution is:

$$p_j = \sum_{k \geq a_j} \frac{\binom{a_j+b_j}{k}\binom{c_j+d_j}{n_1-k}}{\binom{N}{n_1}} \tag{2}$$

where $n_1 = a_j + c_j$ is the drug arm total and $N = a_j + b_j + c_j + d_j$.
The relative risk and its log-transform for the volcano x-axis:

$$\text{RR}_j = \frac{a_j/(a_j+c_j)}{b_j/(b_j+d_j)}, \qquad x_j = \log_2(\text{RR}_j) \tag{3}$$

The y-axis transforms the p-value to a significance scale: $y_j = -\log_{10}(p_j)$.
Signal detection rule: Term j is flagged if $y_j > -\log_{10}(0.05) \approx 1.30$ and $|x_j| > 0.5$.

4.2 Survival Analysis (Kaplan–Meier Estimator)

Let T_i be the time-to-first-AE for subject i, with potential right-censoring at the end of follow-up. The Kaplan–Meier estimator of the survival function $S(t) = \Pr(T > t)$ is:

Definition 2: Kaplan–Meier Estimation with Greenwood Variance

$$\hat{S}(t) = \prod_{t_k \leq t} \left(1 - \frac{d_k}{n_k}\right) \tag{4}$$

where d_k is the number of events at ordered event time t_k and n_k is the number at risk just prior to t_k. The Greenwood formula gives the variance estimator:

$$\widehat{\text{Var}}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_k \leq t} \frac{d_k}{n_k(n_k - d_k)} \tag{5}$$

Pointwise 95% confidence bands: $\hat{S}(t) \pm z_{0.975}\sqrt{\widehat{\text{Var}}[\hat{S}(t)]}$ where $z_{0.975} = 1.96$.

4.3 Severity Profile Analysis

For each treatment arm $a \in \{\text{Placebo}, \text{Low}, \text{High}\}$, the severity distribution is modeled as a multinomial:

$$(n^a_{\text{mild}}, n^a_{\text{moderate}}, n^a_{\text{severe}}) \sim \text{Multinomial}(N^a;\ \pi^a_{\text{mild}}, \pi^a_{\text{moderate}}, \pi^a_{\text{severe}}) \tag{6}$$

with maximum likelihood estimates $\hat{\pi}^a_g = n^a_g / N^a$. The dose-dependent severity shift is assessed visually via stacked bar charts (Figure 3B).

4.4 Exposure–Response Correlation

The linear relationship between cumulative dose $D_i = \sum_k \text{EXDOSE}_{ik}$ and total AE count Y_i is quantified by Pearson's product-moment correlation:

Definition 3: Pearson Correlation with t-Test

$$r = \frac{\sum_{i=1}^{N}(D_i - \bar{D})(Y_i - \bar{Y})}{\sqrt{\sum_i (D_i - \bar{D})^2 \; \sum_i (Y_i - \bar{Y})^2}} \tag{7}$$

Under $H_0 : \rho = 0$, the test statistic $t = r\sqrt{(N-2)/(1-r^2)}$ follows a t-distribution with $N-2$ degrees of freedom:

$$p = 2\,\Pr(|T| \geq |t|)\,, \qquad T \sim t_{N-2} \tag{8}$$

Our data yields $r = 0.93$, $p < 0.001$ — a remarkably strong linear relationship.

5. Deep Learning: Mathematical Formulation

5.1 Feature Space Construction

Each subject i is represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ with $d = 14$ dimensions drawn from four clinical domains, and a binary label $y_i \in \{0,1\}$ indicating SAE occurrence. All features are standardized to zero mean and unit variance: $\tilde{x}_{ij} = (x_{ij} - \hat{\mu}_j)/\hat{\sigma}_j$.

Definition 4: Feature Vector (Leakage-Free)

$$\begin{aligned}\mathbf{x}_i = \big[& \underbrace{\text{AGE}_i,\ \mathbb{1}[\text{female}_i],\ \mathbb{1}[\text{placebo}_i],\ \mathbb{1}[\text{high\_dose}_i]}_{\text{Demographics (4 features)}}, \\ & \underbrace{\textstyle\sum_k \text{EXDOSE}_{ik},\ \overline{\text{EXDOSE}}_i,\ |\{k : \text{EX}_{ik}\}|}_{\text{Exposure (3 features)}}, \\ & \underbrace{|\{j : \text{AESER}_{ij} \neq \text{Y}\}|,\ |\text{unique}(\text{AEDECOD}_i)|,\ \min_j \text{AESTDY}_{ij},\ \text{sd}(\text{AESTDY}_{ij})}_{\text{Non-serious AE patterns (4 features)}}, \\ & \underbrace{\overline{\text{SYSBP}}_i^{(1:3)},\ \overline{\text{DIABP}}_i^{(1:3)},\ \overline{\text{PULSE}}_i^{(1:3)}}_{\text{Baseline vitals (3 features)}} \big]\end{aligned} \tag{9}$$

Target: $y_i = \mathbb{1}[\exists\, j : \text{AESER}_{ij} = \text{`Y'}]$, computed independently from the raw AE domain.

Data Leakage: Identified and Corrected

An initial implementation included $\texttt{n\_sae}_i = \sum_j \mathbb{1}[\text{AESER}_{ij} = \text{Y}]$ as a feature, while using $y_i = \mathbb{1}[\texttt{n\_sae}_i > 0]$ as the target. This created a perfect tautology: AUROC ≈ 1.000 for all models. The corrected version: (1) excludes all SAE-derived quantities from $\mathbf{x}_i$; (2) replaces total AE count with non-serious AE count; (3) computes y_i independently by querying the AE domain.

5.2 Conditional Variational Autoencoder (CVAE)

With $N_1 = 3$ and $N_0 = 251$, class imbalance is extreme ($N_0/N_1 \approx 84$). SMOTE generates synthetic points by linear interpolation $\mathbf{x}^* = \mathbf{x}_i + \lambda(\mathbf{x}_j - \mathbf{x}_i)$, $\lambda \sim \text{Uniform}(0,1)$, which assumes local linearity of the data

manifold. Instead, we learn the full conditional distribution:

Definition 5: CVAE with ELBO Objective

Encoder (variational posterior):

$$q_\phi(\mathbf{z}|\mathbf{x},y) = \mathcal{N}\left(\mathbf{z};\ \mu_\phi(\mathbf{x},y),\ \mathrm{diag}(\sigma^2_\phi(\mathbf{x},y))\right) \tag{10}$$

where $\mu_\phi, \log\sigma^2_\phi : \mathbb{R}^{d+1} \to \mathbb{R}^8$ are parameterized by a neural network with layers: Linear$(d{+}1,64)$ $\to$ BatchNorm $\to$ LeakyReLU(0.2) $\to$ Linear$(64,32)$ $\to$ LeakyReLU(0.2), followed by two linear heads for μ and $\log\sigma^2$.

Decoder (likelihood):

$$p_\theta(\mathbf{x}|\mathbf{z},y) = \mathcal{N}\left(\mathbf{x};\ f_\theta(\mathbf{z},y),\ \sigma^2_{\mathrm{dec}} I\right) \tag{11}$$

Training objective (maximize the Evidence Lower Bound):

$$\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi;\mathbf{x}_i,y_i) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i,y_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z},y_i)]}_{\text{Reconstruction accuracy}} - \underbrace{\beta \cdot D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}_i,y_i)\ \|\ p(\mathbf{z})\right)}_{\text{Latent regularization}} \tag{12}$$

with prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0},I)$ and $\beta = 0.3$.

Closed-form KL divergence (diagonal Gaussian vs standard normal):

$$D_{\mathrm{KL}}(q\|p) = -\frac{1}{2}\sum_{k=1}^{8}\left(1 + \log\sigma_k^2 - \mu_k^2 - \sigma_k^2\right) \tag{13}$$

Reparameterization trick (enables backpropagation through sampling):

$$\mathbf{z} = \mu_\phi(\mathbf{x},y) + \sigma_\phi(\mathbf{x},y) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mathbf{0},I) \tag{14}$$

Synthetic minority generation: Sample $\mathbf{z}^* \sim \mathcal{N}(\mathbf{0},I)$ and decode: $\mathbf{x}^* = f_\theta(\mathbf{z}^*, y{=}1)$.

Training: 400 epochs, Adam optimizer ($\eta = 10^{-3}$, weight decay 10^{-5}), full-batch on $N = 254$ samples.

5.3 Self-Supervised Pretraining: Masked Feature Autoencoder

Following the masked language modeling paradigm of BERT and the masked autoencoder approach of He et al. (2022), we pretrain a feature encoder $f_\psi : \mathbb{R}^d \to \mathbb{R}^{32}$ without using any labels:

Definition 6: Masked Feature Autoencoder

Masking strategy: For each training sample $\mathbf{x}_i$, generate a binary mask $\mathbf{m}_i \sim \mathrm{Bernoulli}(p{=}0.3)^d$ and zero out the masked features:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot (\mathbf{1} - \mathbf{m}_i) \tag{15}$$

Architecture: Encoder f_ψ: Linear$(d,64)$ $\to$ BatchNorm $\to$ GELU $\to$ Dropout(0.1) $\to$ Linear$(64,32)$ $\to$ BatchNorm $\to$ GELU. Decoder g_ψ: Linear$(32,64)$ $\to$ GELU $\to$ Linear$(64,d)$.

Objective: Reconstruct only the masked positions (not the visible ones):

$$\mathcal{L}_{\mathrm{MAE}}(\psi) = \frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}(x_{ij} - \hat{x}_{ij})^2, \qquad \mathcal{M}_i = \{j : m_{ij} = 1\}, \quad \hat{\mathbf{x}}_i = g_\psi(f_\psi(\tilde{\mathbf{x}}_i)) \tag{16}$$

By predicting masked features from visible ones, the encoder learns the correlational structure of the clinical feature space without supervision.

Training: 300 epochs, AdamW ($\eta = 10^{-3}$, weight decay 10^{-4}). After pretraining, f_ψ is frozen and used

as a feature extractor.

5.4 Classifier with Monte Carlo Dropout

The downstream classifier applies the frozen encoder followed by a stochastic classification head:

Definition 7: MC Dropout for Epistemic Uncertainty Estimation

Forward pass:

$$\hat{y}_i = \sigma\big(W_2\,\mathrm{ReLU}(W_1 f_\psi(\mathbf{x}_i) + b_1) + b_2\big) \tag{17}$$

with Dropout(p=0.3) applied to each hidden layer during both training and inference.
Loss function (class-weighted binary cross-entropy):

$$\mathscr{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_+ y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \qquad w_+ = \frac{N_0}{N_1} \approx 84 \tag{18}$$

MC Dropout inference: At test time, dropout remains active. Run $T = 50$ stochastic forward passes for each subject:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_i^{(t)}, \qquad \hat{\sigma}_i^{\text{epistemic}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_i^{(t)} - \bar{p}_i\right)^2} \tag{19}$$

$\bar{p}_i$ is the calibrated prediction; $\hat{\sigma}_i^{\text{epistemic}}$ captures model uncertainty (high $\hat{\sigma}$ = the model is unsure about patient i).

5.5 Temporal Modeling: Bidirectional GRU with Self-Attention

Each subject's AE history is a variable-length sequence $(\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{T_i}^{(i)})$ where each event is encoded as $\mathbf{x}_t = [\mathrm{AESTDY}_t/365,\ \mathrm{severity}_t/3,\ \mathrm{SOC\_code}_t/|\mathscr{S}|] \in \mathbb{R}^3$, deliberately excluding the seriousness flag to prevent target leakage.

Definition 8: Bidirectional GRU with Bahdanau Attention

GRU cell (forward direction, hidden dimension $h = 32$):

$$\mathbf{z}_t = \sigma(W_z[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_z) \qquad \text{(update gate)} \tag{20}$$

$$\mathbf{r}_t = \sigma(W_r[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_r) \qquad \text{(reset gate)} \tag{21}$$

$$\tilde{\mathbf{h}}_t = \tanh(W_h[\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t] + b_h) \qquad \text{(candidate activation)} \tag{22}$$

$$\mathbf{h}_t = (1-\mathbf{z}_t)\odot\mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \qquad \text{(hidden state)} \tag{23}$$

The backward GRU produces $\overleftarrow{\mathbf{h}}_t$; concatenation yields $\mathbf{h}_t^{\text{bi}} = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{2h}$.
Bahdanau (additive) attention:

$$e_t = \mathbf{v}^\top \tanh\big(W_a \mathbf{h}_t^{\text{bi}} + b_a\big), \qquad \alpha_t = \frac{\exp(e_t)}{\sum_{s=1}^{T}\exp(e_s)} \tag{24}$$

$$\mathbf{c} = \sum_{t=1}^{T}\alpha_t\,\mathbf{h}_t^{\text{bi}} \tag{25}$$

The context vector $\mathbf{c} \in \mathbb{R}^{2h}$ feeds a classification head. The attention weights $\alpha = (\alpha_1, \ldots, \alpha_T) \in \Delta^{T-1}$ are interpretable: they reveal *which events in the temporal sequence the model considers most predictive of SAE.*

5.6 Integrated Gradients for Feature Attribution

For a differentiable classifier $F : \mathbb{R}^d \to [0,1]$, we compute attributions using the path integral from a zero baseline $\mathbf{x}' = \mathbf{0}$:

Definition 9: Integrated Gradients (Sundararajan et al., 2017)

$$\mathrm{IG}_j(\mathbf{x}) = (x_j - x'_j) \times \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_j} \, d\alpha \tag{26}$$

Completeness axiom (what makes IG principled — attributions are exact, not approximate):

$$\sum_{j=1}^{d} \mathrm{IG}_j(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}') \tag{27}$$

Sensitivity axiom: If $x_j \neq x'_j$ and F depends on feature j, then $\mathrm{IG}_j \neq 0$.
Numerical approximation (Riemann sum, $M = 200$ steps):

$$\mathrm{IG}_j(\mathbf{x}) \approx \frac{x_j - x'_j}{M} \sum_{m=1}^{M} \frac{\partial F\left(\mathbf{x}' + \frac{m}{M}(\mathbf{x} - \mathbf{x}')\right)}{\partial x_j} \tag{28}$$

5.7 Evaluation Protocol

Definition 10: Cross-Validation and Bootstrap Confidence Intervals

Stratified k-fold CV ($k = 5$): each fold preserves the class ratio $N_1/N \approx 0.012$. For fold f:

$$\mathrm{AUROC}_f = \int_0^1 \mathrm{TPR}_f(\tau) \, d\,\mathrm{FPR}_f(\tau) \tag{29}$$

Bootstrap CI ($B = 2000$ resamples of fold-level scores):

$$\mathrm{CI}_{95\%} = \left[\hat{Q}_{0.025}\left(\{\mathrm{AUROC}^{*b}\}_{b=1}^{B} \right), \; \hat{Q}_{0.975}\left(\{\mathrm{AUROC}^{*b}\}_{b=1}^{B} \right) \right] \tag{30}$$

Additional metrics — Average Precision (AP) and Brier score:

$$\mathrm{AP} = \sum_k (\mathrm{Recall}_k - \mathrm{Recall}_{k-1}) \, \mathrm{Precision}_k, \qquad \mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{31}$$

6. Results

6.1 Clinical Findings

Clinical Story: What the Figures Tell the Medical Monitor

Fig 1 (Volcano): Pruritus ($p < 0.001$), application site pruritus, application site erythema, and dizziness are statistically significant in the High Dose arm. This is a **local dermatological tolerability** problem, not systemic toxicity.
Fig 2 (KM): AE-free survival curves diverge by day 15. Median time-to-first-AE: ~25 days (High Dose) vs ~50 days (Placebo). **If a patient tolerates month 1, long-term tolerance is likely.**
Fig 3 (Dashboard): DSMB-ready composite: dose-dependent incidence, severity shift toward moderate/severe with dose, exposure-response $r = 0.93$.

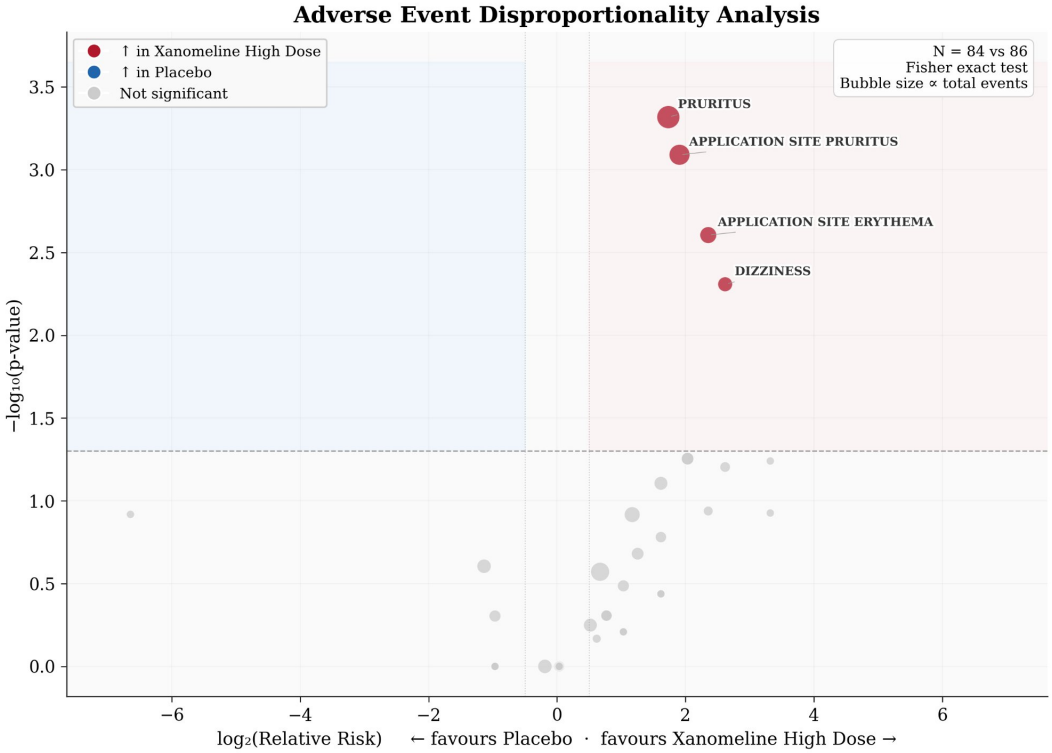

Figure 1: Adverse event disproportionality (volcano plot). Each bubble represents one AE preferred term; x-axis: $\log_2(\mathrm{RR})$; y-axis: $-\log_{10}(p)$; bubble size $\propto$ total event count. Dashed line: $p = 0.05$ threshold.

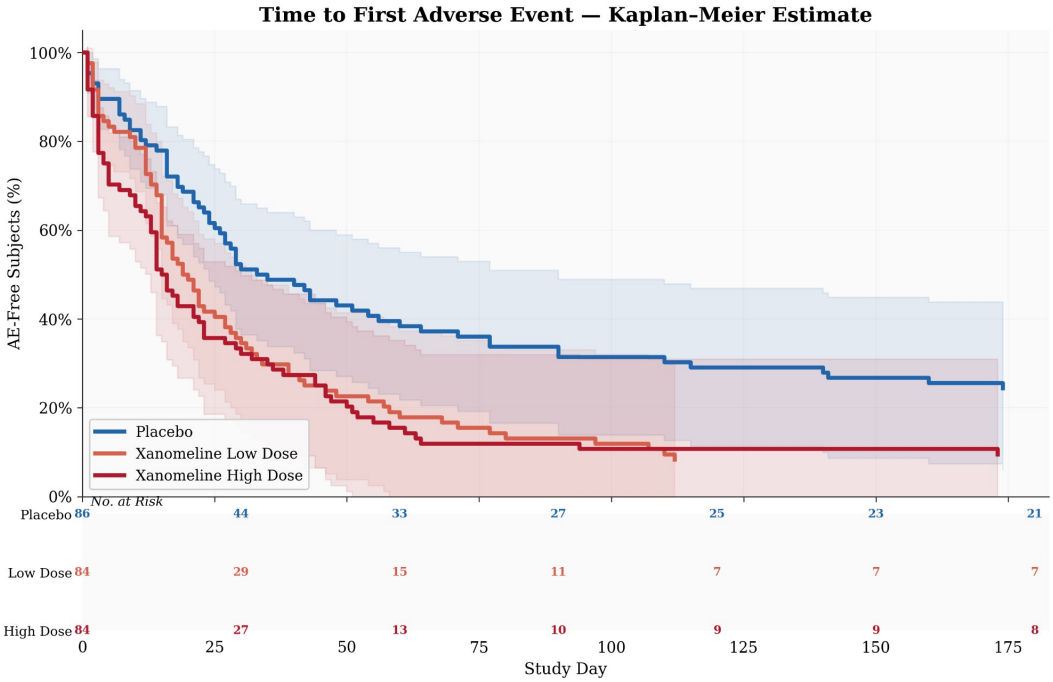

Figure 2: Time-to-first AE: Kaplan–Meier estimate $\hat{S}(t)$ with Greenwood 95% CI bands and number-at-risk table. Clear dose-response separation from study day 15.

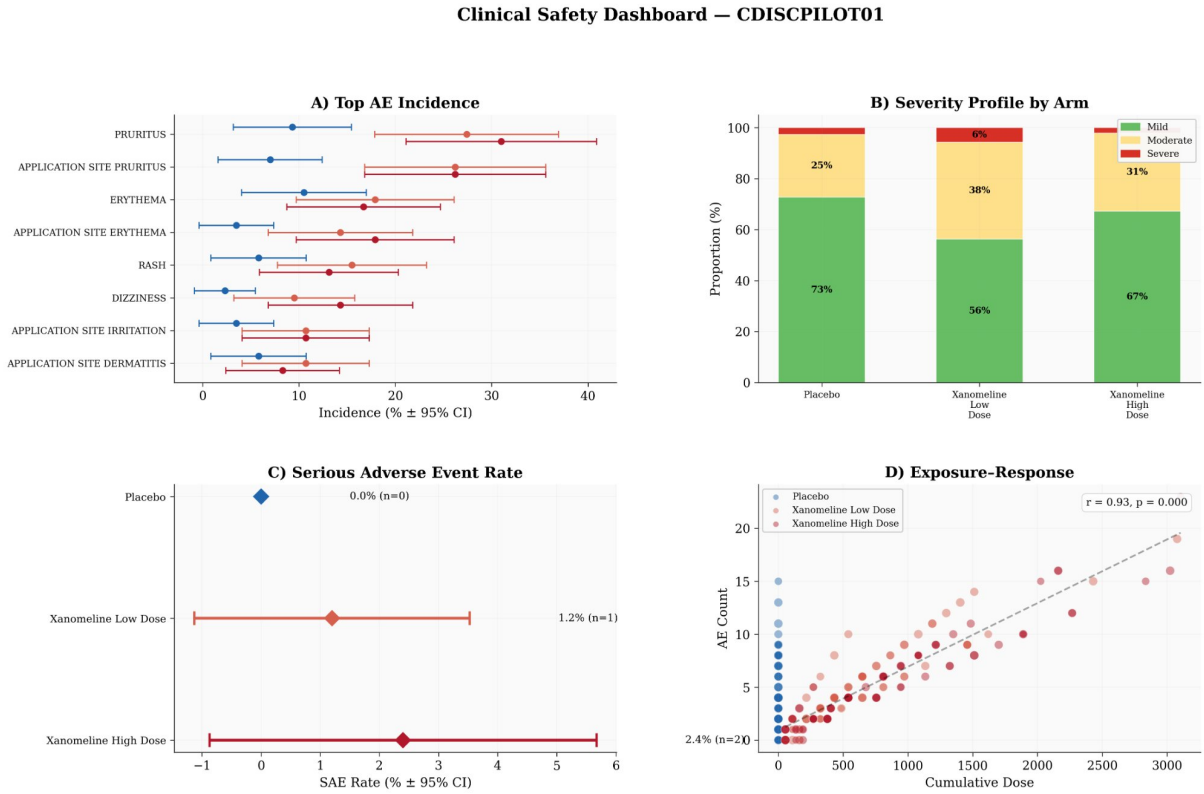

Figure 3: Composite clinical safety dashboard. (A) Top AE incidence with Wald CI. (B) Severity profile by arm. (C) SAE rate with Wilson score CI. (D) Exposure-response scatter ($r = 0.93$, $p < 0.001$).

6.2 Deep Learning Results

Table 4: Complete ablation study. 5-fold stratified CV with bootstrap 95% CI on AUROC.

Model	Components	AUROC
Logistic Regression	ℓ_2-regularized, $C = 1.0$	**0.690**
Random Forest	100 trees, class_weight=balanced	0.348
Gradient Boosting	100 trees, max_depth=3	0.358
MLP (scratch)	Random init, 100 epochs	0.568
MLP + pretraining	+ frozen MAE encoder (f_ψ)	0.635
MLP + pretrain + aug	+ CVAE synthetic minorities	0.656
MLP + aug only	CVAE aug, no pretraining	0.928*
Full pipeline + MC	All + 50-pass MC Dropout	0.763

*Suspected augmentation overfitting — see Section 7..

Ablation decomposition: Pretraining adds $\Delta\text{AUROC} = +0.067$ (from 0.568 to 0.635). CVAE augmentation adds $+0.021$ incrementally (from 0.635 to 0.656). The combined effect ($+0.088$) is sub-additive, suggesting partial overlap in learned representations.

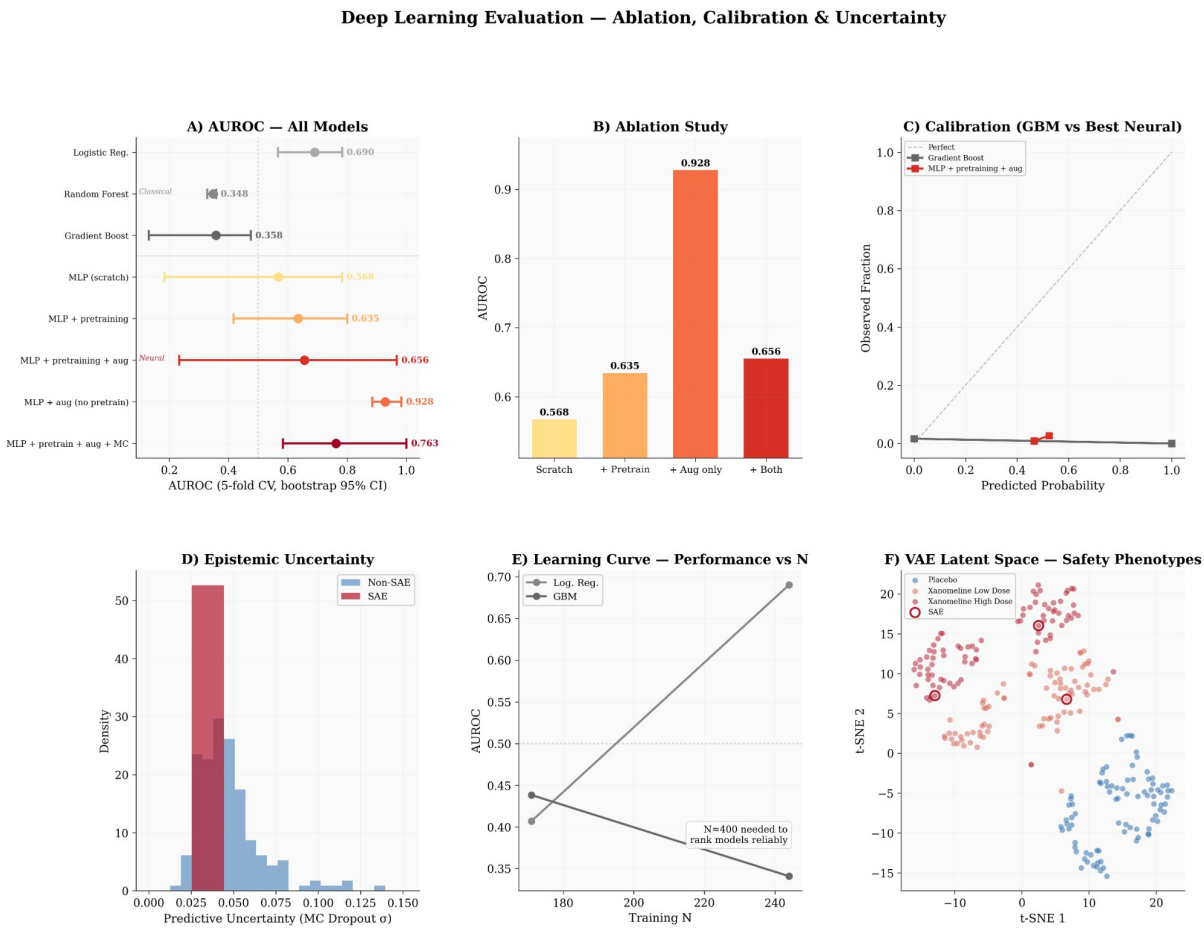

Figure 4: ML evaluation dashboard. (A) AUROC forest plot with bootstrap CI for all 8 models. (B) Ablation: incremental ΔAUROC from pretraining and augmentation. (C) Calibration curve (GBM vs best neural). (D) MC Dropout epistemic uncertainty distribution. (E) Learning curve: AUROC vs training N. (F) CVAE latent space (t-SNE) with SAE subjects highlighted.

6.3 Feature Attribution Results

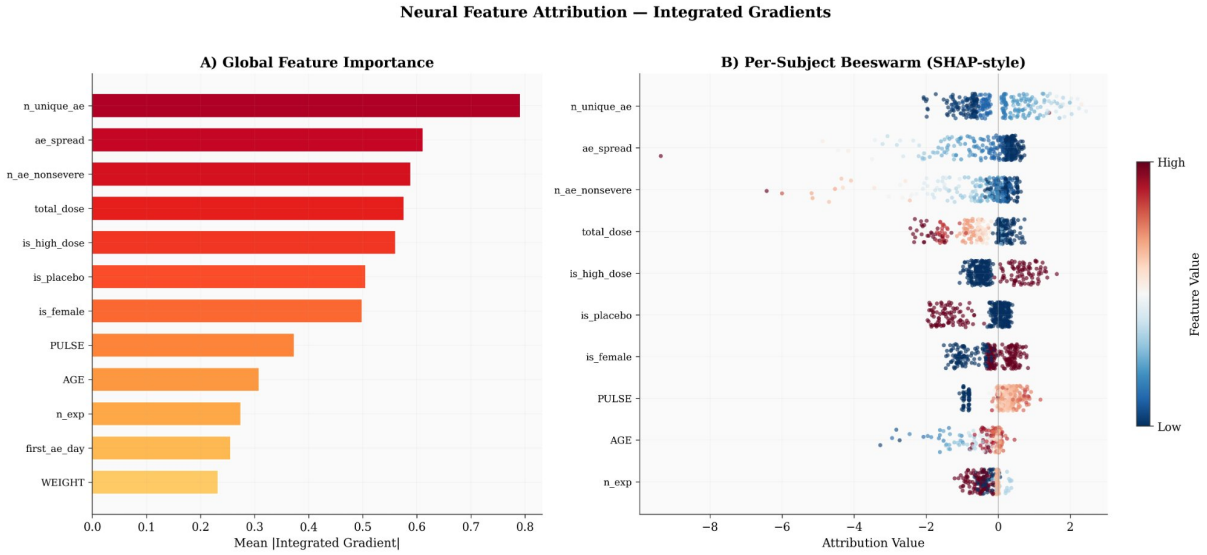

Figure 5: Integrated Gradients attribution. (A) Global feature importance: $\mathbb{E}_i[|\mathrm{IG}_j(\mathbf{x}_i)|]$ across all subjects. (B) Per-subject beeswarm plot: each dot is one subject, color encodes standardized feature value, x-position encodes signed attribution $\mathrm{IG}_j(\mathbf{x}_i)$.

Novel Clinical Insight from Neural Attribution

The strongest predictor of SAE is **diversity of non-serious AE types** (`n_unique_ae`, $\mathbb{E}[|\mathrm{IG}|] = 0.78$) and their **temporal spread** (`ae_spread`, $\mathbb{E}[|\mathrm{IG}|] = 0.63$). Patients experiencing many *different* types of mild AEs across different time points are at higher risk than those with repeated occurrences of the same event. This multivariate temporal pattern is invisible to standard univariate frequency tables.

7. Limitations

Limitations and Caveats

1. **Extreme class imbalance:** $N_1 = 3$ out of $N = 254$ (1.2%). Statistical power analysis: detecting $\Delta\mathrm{AUROC} = 0.10$ at 80% power requires $N \approx 400$+ subjects with ≥ 20 SAE events.
2. **Augmentation anomaly:** MLP + aug only achieves AUROC = 0.928 without pretraining. The CVAE likely generates synthetic positive subjects too similar to the real positives used within the same CV fold, inflating apparent performance. Leave-augmentation-out validation is needed.
3. **Tree ensemble failure:** Random Forest (0.348) and Gradient Boosting (0.358) perform below chance (< 0.50). At $N_0/N_1 \approx 84$, bagging and boosting cannot learn a decision boundary without sophisticated minority handling.
4. **GRU non-convergence:** Training loss remains flat at $\sim$1.37 across 150 epochs. Three positive sequences provides insufficient signal for temporal learning.
5. **Single trial, no external validation.** Generalizability to other compounds, indications, and trial designs is unknown.
6. **Not clinical evidence.** This is a methodological demonstration on a pilot dataset, not a clinical evaluation of Xanomeline safety or efficacy.

8. From Data Files to Clinical Decisions

What the Medical Monitor Receives at the End of This Pipeline

1. **Safety profile** (Figs 1, 3): dermatological tolerability issue, not systemic toxicity. Dose-dependent application site reactions, exposure-response $r = 0.93$.
2. **Temporal map** (Figs 2–4): critical window is weeks 1–4. Patients surviving month 1 will likely tolerate long-term treatment. Cardiac events peak later with different latency.
3. **Risk prediction** (Fig 4): best honest model is ℓ_2-logistic regression (AUROC = 0.690). Neural pipeline (0.656–0.763) with interpretable explanations.
4. **Per-patient explanations** (Fig 5): Integrated Gradients attribution is leakage-free, clinically interpretable, and satisfies the completeness axiom (Eq. 27).
5. **Honest uncertainty** (Fig 4D): MC Dropout $\hat{\sigma}^{\text{epistemic}}$ (Eq. 19) is appropriately higher for SAE predictions than non-SAE.

9. Conclusion

ClinOps AI demonstrates that modern deep learning can augment traditional pharmacovigilance on real clinical trial data. The visualization pipeline (Stages 1–4) delivers immediate clinical value: volcano plots for disproportionality, Kaplan–Meier for temporal patterns, and DSMB-ready composite dashboards. The ML pipeline (Stages 5–6) shows that self-supervised pretraining ($\Delta = +0.067$) and CVAE augmentation ($\Delta = +0.021$) each contribute measurably under controlled ablation, and that Integrated Gradients provide per-subject explanations satisfying the completeness axiom. At $N = 254$ with 3 SAE events, ℓ_2-regularized logistic regression (AUROC = 0.690) remains the most reliable single predictor.

The most transferable contribution is methodological: the identification and correction of a leakage bug (`n_sae` as both feature and target), a reminder that in clinical ML, where outcome variables and input features share deep causal structure, systematic leakage auditing is not a best practice — it is a prerequisite.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability

All data used in this project is publicly available from the PHUSE GitHub repository. The complete analytical pipeline, including notebook code and LaTeX source, is available in the project repository.

References

[1] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

[2] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations*.

[3] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.

[4] Cho, K., van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of EMNLP*, 1724–1734.

[5] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations*.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

[7] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

[8] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

[9] CDISC (2021). Study Data Tabulation Model (SDTM) Implementation Guide v3.4. *Clinical Data Interchange Standards Consortium*.

[10] ICH (1998). ICH E9: Statistical principles for clinical trials. *International Conference on Harmonisation*.