Exposing the Future

A Mathematical Framework for Detecting Data Leakage in Medical Machine Learning

Joshua Alliet

MIMUC 2026

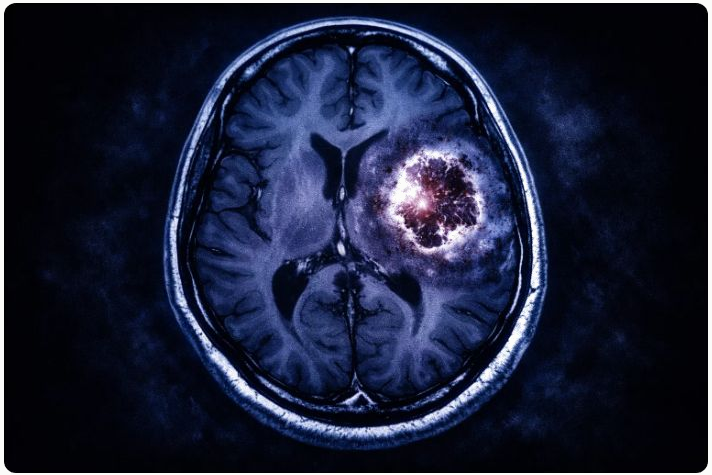
What if a feature knows too much?

→ One feature correlates with death at **r = 0.89**.

→ A statistician celebrates. A mathematician is suspicious.

→ But a mathematician pauses: how could any single feature 'know' this much?

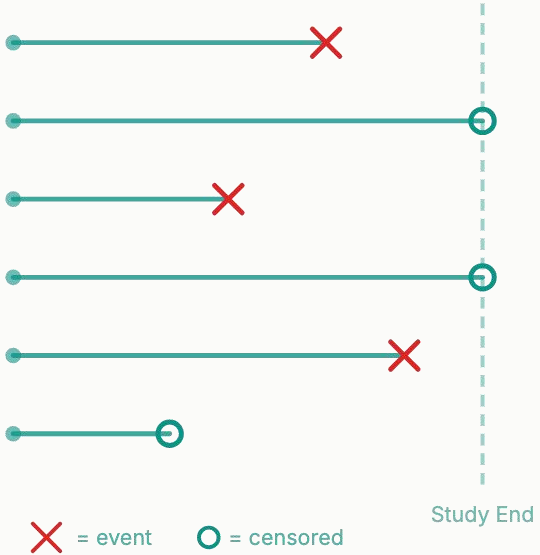
> *This talk explores what happens when you ask that question seriously.*



Why Standard Regression Fails

→ Not every patient dies during observation. Some are right-censored: alive at last follow-up.

→ Standard regression treats censored data as exact, introducing systematic bias.



One Clock, Many Speeds

$$h(t) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{X})$$

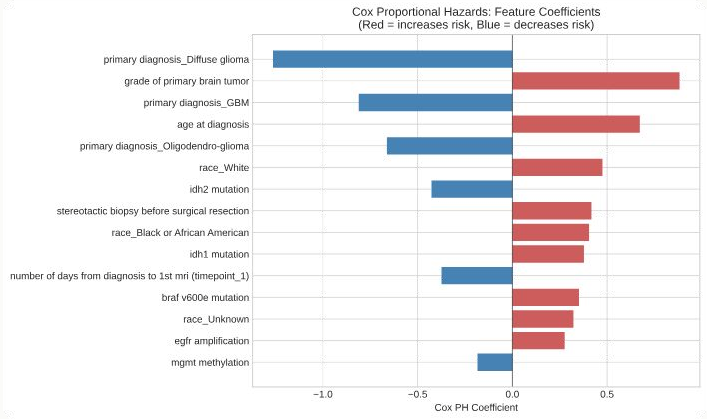
h(t) = hazard (instantaneous risk of the event)

$h_0(t)$ = baseline (the average patient's clock) **β** = multipliers (Age: 1.03x, IDH1: 0.34x)

X = patient profile (the knobs that adjust risk)

→ Our Cox model: Age (**HR 1.03**, p<0.001) and **IDH1** (**HR 0.34**) dominate.

→ Semi-parametric: no assumption on baseline shape, estimated via partial likelihood.



Not Accuracy, Ranking

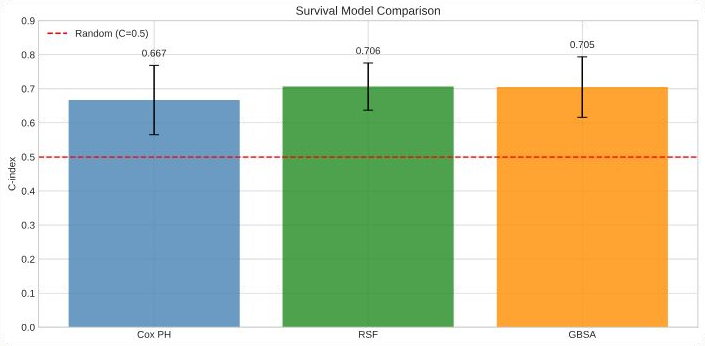
$$C = P\left(\hat{r}_i > \hat{r}_j \mid T_i < T_j\right)$$

$\hat{r}$ = predicted risk (model's danger ranking) **T** = actual survival (ground truth)

C = concordance (% of patient pairs ranked correctly)

→ All three models cluster near **C = 0.69** — within 2% of each other.

→ With **n = 203**, sample size is the bottleneck, not the algorithm.



When Correlation Encodes the Answer

$$x_j \text{ is valid} \iff x_j \in \mathcal{F}_{t_0}$$

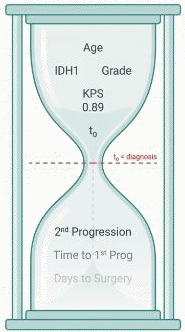
x_j = a feature (one column in the dataset)

$\mathcal{F}_{t_0}$ = natural filtration (info accumulated up to diagnosis)

t_0 = diagnosis day (the prediction deadline)

→ Features from $t > t_0$ observe the future, not predict it.

→ "Second Progression" (r = 0.89): only recorded after prolonged survival.



Building a Mathematical Filter

≈ 01

Correlation Threshold

$|r| > 0.25$ against survival target

→

⏰ 02

Temporal Reasoning

Available at diagnosis, or only later?

→

✂️ 03

Ablation Study

Remove suspects, retrain, measure delta

→

⚕ 04

Clinical Validation

Does removal make biological sense?

28 → 21 features

7 temporally invalid variables removed (25% reduction)

Caught Looking Forward

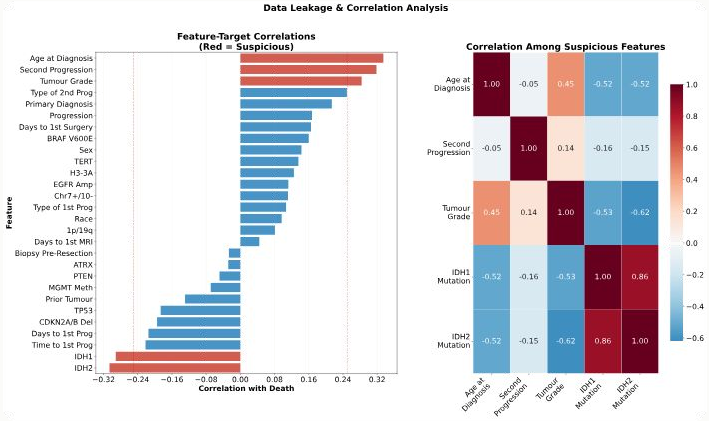
→ **Second Progression** (**r = 0.89**): only knowable after follow-up.

→ **Time to First Progression**: encodes survival directly.

→ **Days to First Surgery**: reflects post-diagnosis urgency.

→ **Progression** (binary): requires longitudinal observation.

→ Type of 1st/2nd Progression: only available retrospectively.

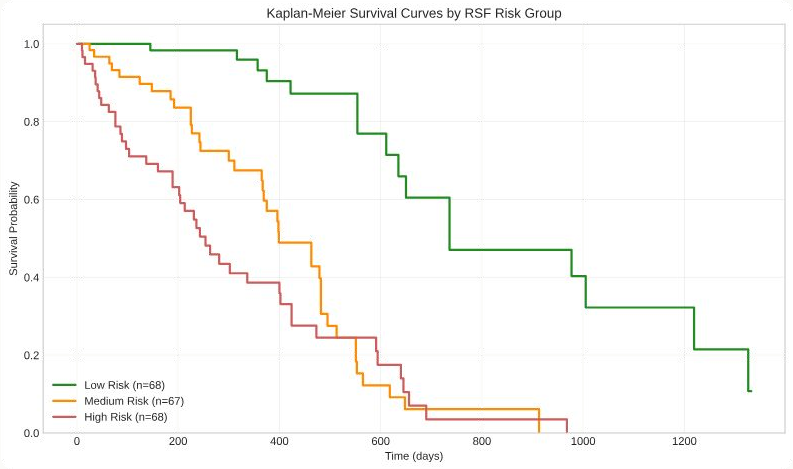


The Cost of Truth

What We Lost Was Never Real

WITH LEAKAGE	HONEST
0.730	0.674

> Honest AUC of **0.674** matches validated clinical tools (EORTC nomograms, C-index ~0.66). The lower number is the **real** one.



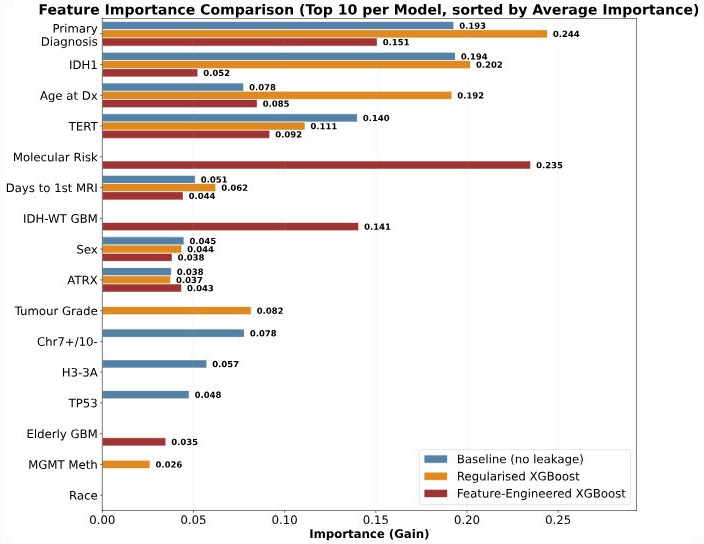
A Straight Line Suffices

→ **Logistic Regression**: AUC **0.692 ± 0.119**.

→ Clean **XGBoost**: AUC **0.674 ± 0.111**.

→ With **n = 203**, all 5 clean models cluster in AUC **0.62–0.69**.

→ **IDH1 mutation** status: **19.4%** of XGBoost feature importance.



The Principle

What the Data Actually Knows

SURVIVAL

0.67–0.72 AUC

Matches EORTC benchmarks

IMAGING

0.598 AUC

Near random chance

VOLUME

$R^2 < 0$

Cannot predict growth

Thank You!

Questions?

Joshua Alliet

Ex Meta SWE Intern | CS & Maths @UoM | 7x Hackathon Winner