Prediction of Speech Outcomes with Cochlear Implants

Limitations and Opportunities for Clinical Practice

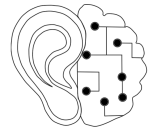
Dr. Marta Campi Prof. Alexander Huber Prof. Tobias Goehring

Deep Hearing Lab · University Hospital Zurich

Blunch Series, Cambridge, 20 March 2026

University of Zurich[UZH]

USZ Universitäts Spital Zürich



Deep Hearing Lab

Why Prediction Matters

Deep Hearing Lab

"How well will I hear after the implant?"

Every CI candidate asks this question. Clinicians need to answer it for:

Pre-operative counselling	Set realistic expectations
Surgical decision-making	Especially borderline candidates
Post-operative monitoring	Flag patients who underperform expectations
Resource allocation	Target rehabilitation where it matters most

The problem: despite decades of research, no model reliably predicts individual CI speech outcomes from pre-operative data.

The Prediction Problem in CI Rehabilitation

All studies below predict word recognition score (WRS) from routine pre-operative clinical data.

Study	Model	N	Test	Validation	Best result
Kim (2018)	random forest	120	Korean mono	LOO + cross-site	MAE 4.8 → 17.1
Shafieibavani (2021)	ensemble	2,489	CNC/Frei/CVC	cross-site	MAE 18–25 pp
Bertschinger (2024)	linear + sel.	360	Freiburg mono	in-sample	$R^2 = 0.27$
Demyanchuk (2025)	decision tree	2,571	Freiburg mono	held-out + temp.	MAE 17 pp
Ollermann (2025)	lin., tree, ens.	236	Freiburg mono	held-out test	$R^2 < 0$
Anthonisen (2026)	8 models	2,251	multi-site WRS	held-out (10×)	$R^2 = 0.11$

Caution: Mo et al. (2025; 16 studies, 5,058 patients): most ML studies lack proper validation. All predict monosyllabic WRS but use different word lists (Freiburg = DE/CH, CNC = US, CVC = AU).

Validation hierarchy: in-sample (optimistic) → held-out → LOO / CV (honest) → **cross-site** (hardest).

Why does this ceiling exist? What makes WRS so hard to predict?

The Dataset: Swiss National CI Database, Zurich Centre

Deep Hearing Lab

Cohort & sample size flow

Source	Swiss Nat. CI DB
Centre	Univ. Hosp. Zurich
Period	2000–2024
All patients	771
Adults ($\geq$18 yr)	538
With post-op Mono1	431
Pre + post Mono1	**266**
(165 too deaf pre-op)	
With trajectory data	136
(4 time points, Mono1)	
With 3 post-op tests	387
(3 speech tests, post-op)	

Bias: N=266 excludes the most severely impaired (no pre-op speech possible).

Variables (257 columns, 85 usable pre-op)

Pure Tone Audiometry — 30 frequencies, CI + contralateral
Pre-op speech tests — Mono1, FEMMAL, Numbers, V08, C12
Demographics — age at surgery, sex
Etiology & onset — duration of deafness, pre/post-lingual
Communication — articulation, lip reading, phone use
Socioeconomic status — education, employment
Surgical — CI model, electrode type, surgeon
Imaging — CT, MRI (qualitative only)
Quality of life — SSQ12, THI, IOI-HA

Our Study: What Are We Trying to Predict?

Deep Hearing Lab

The prediction problem

Given pre-operative data X_{pre}, predict post-operative speech outcome y:

$$\hat{y} = f(X_{\text{pre}})$$

X_{pre}: demographics, audiometry, speech, etiology (∼85 features)

y: post-operative Mono1 (%)

Primary outcome: Mono1

Freiburg monosyllabic word recognition score (WRS). Open-set, 20 words at 65 dB SPL, CI-only. The Swiss clinical standard — widest dynamic range, only 16% at ceiling.

Five post-operative speech tests

Test	What it measures	r_{Mono1}	Ceiling?
FM	Voice discrim. (M vs F)	0.48	Yes (68%)
Num	Number recognition	0.63	Yes (81%)
V08	Vowel discrimination	0.68	No
C12	Consonant discrimination	0.79	No
Mono1	**Open-set words (WRS)**	—	No (16%)

r_{Mono1} = post-op correlation with Mono1

↑ difficulty: FM easiest → Mono1 hardest

Research Questions & Outline

Deep Hearing Lab

Objective: predict post-operative speech outcomes from pre-operative data — 771 patients, 3 analyses

1.	Can we predict post-operative speech outcomes?	**The Hard Ceiling**
2.	Why does pre-operative prediction fail?	**Three Reasons Why**
3.	Can we identify patients at risk?	**What We CAN Predict**

Part 1

Can we predict post-operative speech outcomes?

Eight Algorithms, One Answer

8 algorithms × 5 feature sets × 5 imputation strategies — all roads lead to the same ceiling

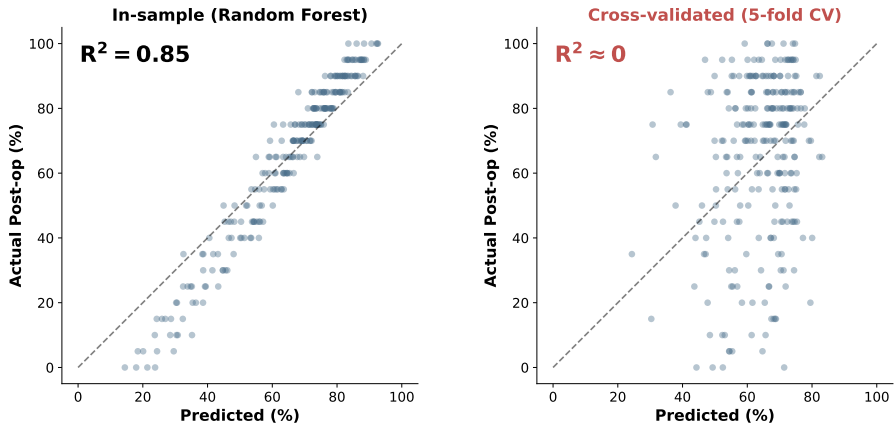
Model	**Type**	R^2 **(train)**	**CV** R^2	**CV MAE**
Linear regression	Linear	0.18	0.03	19.4
Ridge / Lasso	Regularised	0.17	0.02	20.9
ABESS (Bertschinger)	Subset sel.	0.18	0.03	19.4
Random Forest	Ensemble	0.85	−0.03	21.4
Gradient Boosting	Boosting	0.60	−0.02	21.3
XGBoost	Boosting	0.55	−0.00	21.2
LightGBM	Boosting	0.52	−0.01	21.2
TabPFN	**Foundation**	—	**0.10**	**20.1**

In-sample R^2 up to 0.85 — but **all collapse to** $R^2 \leq 0.10$ **under CV**. More features → **worse**, not better (3 → 85 → 226: R^2 degrades).

Conformal 90% PI: $\pm$**41 pp** (distribution-free) — "90% sure they'll score between 25% and 100%."

What Does $R^2 \approx 0$ Look Like?

In-sample vs cross-validated predictions — same data, different honesty



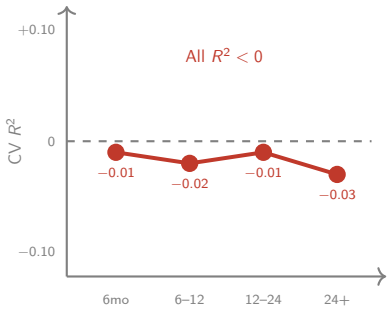
Left: model trained and tested on the same patients — fits noise. **Right:** each patient predicted by a model that never saw them — predictions collapse to ∼65%.

This Holds at Every Timepoint

Same pre-op features, different response variables — same answer

Outcome timepoint	CV R^2	N
→ 6 months	−0.01	166
→ 6–12 months	−0.02	185
→ 12–24 months	−0.01	224
→ 24+ months	−0.03	200

All worse than predicting the mean



The ceiling is not a measurement artefact. It appears **immediately** at 6 months and **never lifts** — not at 12mo, not at 24mo+.

The bottleneck is **information**, not timing.

Part 2

Why does prediction fail?

The CI Equalizer Effect

What happens to CI patients after surgery?

Deep Hearing Lab

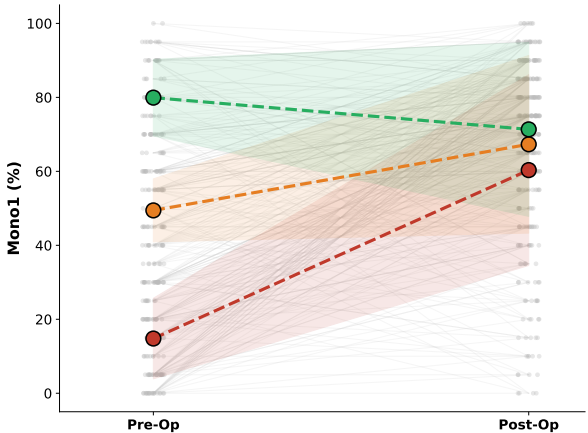
Pre-op → post-op convergence

Pre-op	N	Mean pre	Mean post
≤30%	119	15%	60%
30–60%	80	49%	67%
>60%	67	80%	71%

Means converge to ∼65%.

But post-op SD ≈ 25% **in every group**.

Means converge, **variance stays**. Individual outcomes remain unpredictable.



Same pattern for any grouping (post-op means): Age (62/66/67%) · Contra PTA (65/68/62%) · CI PTA (67/66/63%) · All SD ≈ 24%. **Exception:** pre-lingual onset (52%) vs post-lingual (69%).

Why Does $R^2 \approx 0$?

Deep Hearing Lab

General (any model)

Given a model $Y = f(X) + \varepsilon$,
where $f(X)$ estimates $E[Y|X]$:

Law of total variance:

$$\mathrm{Var}(Y) = \underbrace{\mathrm{Var}(E[Y|X])}_{\substack{\text{var. of predictions} \\ \text{explained}}} + \underbrace{E[\mathrm{Var}(Y|X)]}_{\substack{\text{avg. residual var.} \\ \text{unexplained}}}$$

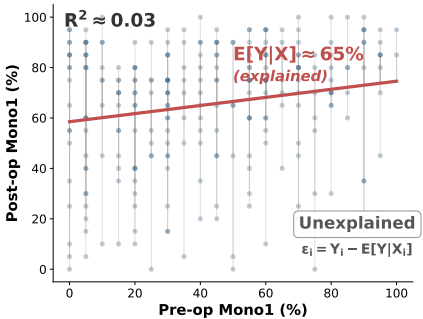
$$R^2 = \frac{\mathrm{Var}(E[Y|X])}{\mathrm{Var}(Y)}$$

Valid for **any** model (linear, nonlinear, AI).

If the expected outcome given all pre-op features is the same for all patients:

$$\mathrm{Var}(E[Y|X]) = 0 \quad \Rightarrow \quad R^2 = 0$$

For example, in a linear regression...



$E[Y|X] \approx 65\%$ for all X: the line is flat.

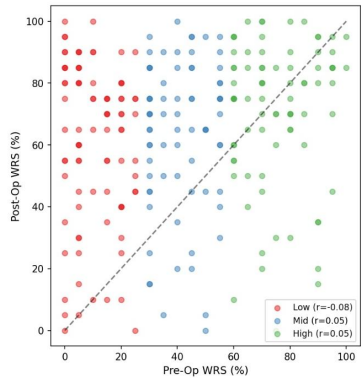
The whiskers (unexplained) dominate. $R^2 \approx 0.03$.

The Equalizer showed $E[Y|X] \approx 65\%$ for all patient profiles $\Rightarrow$ variance of predictions $\approx 0 \Rightarrow R^2 \approx 0$. Not a modelling failure — there is nothing to predict.

Simpson's Paradox in CI Outcomes

Stratum	N	r
Overall	266	**0.18***
Low (0–30%)	119	−0.03
Mid (30–60%)	80	+0.08
High (60–100%)	67	+0.12

Overall $r = 0.18$ — but within every stratum: $r \approx 0$.



Clinical meaning: Knowing a patient's pre-op score gives **no information** about their post-op score once stratum is accounted for. Pre-op score tells you the starting group — not where they land within it.

Mathematical Coupling: The $r = -0.70$ Trap

Why "improvement" from pre-op score is always spurious

Deep Hearing Lab

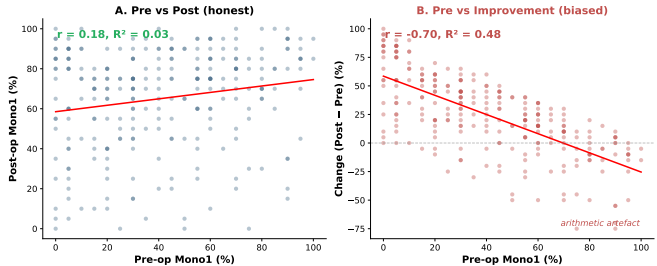
Predict **change** from Pre:

$$\mathrm{Cov}(X, Y - X) = \mathrm{Cov}(X, Y) - \mathrm{Var}(X)$$

Even if $X \perp Y$: correlation is **negative by construction**.

Question	Result
Pre → Change	$r = -0.70$
apparent R^2	0.48
Pre → Post	$r = 0.18$
honest R^2	0.03

Same data. Different question.
15× inflation from one formula.



Many published "predictors of improvement" are reporting B, not A.

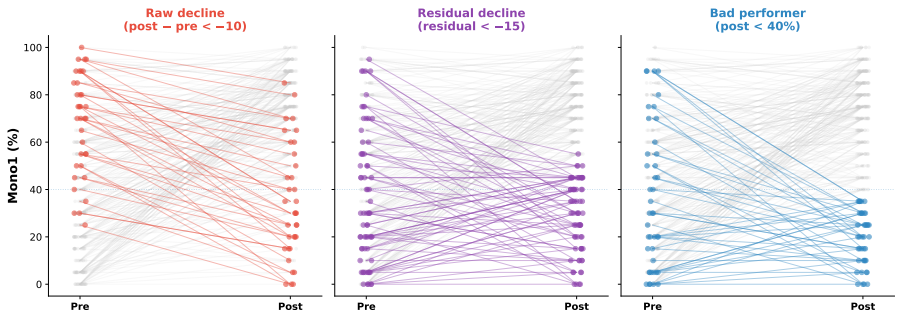
Part 3

What CAN we predict?

Reframing the Problem: From Regression to Risk

If we can't predict a score — can we predict who is at risk?

Target	Definition	Formula
Decline (raw)	>10 pp drop from pre-op	Post − Pre < −10
Decline (residual)	Drop beyond LOO-expected	Post − $\hat{y}_{\text{LOO}}$ < −15
Poor performer	Post-op Mono1 < 40%	$\text{Mono1}_{\text{post}}$ < 40



Different definitions flag different patients. Note: some residual decliners actually *improved* — but less than expected.

Two Definitions of Decline

The definition determines which patients you flag — and which you miss

Deep Hearing Lab

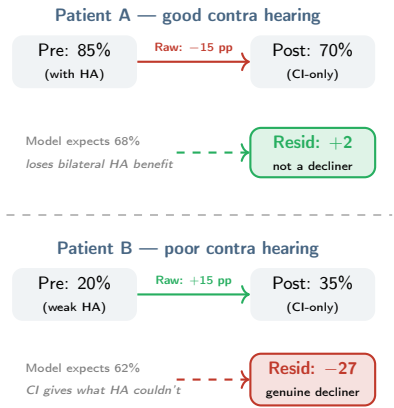
Raw decline

$\text{Post}_i - \text{Pre}_i < -10$ pp. Problem: pre-op with **hearing aids**, post-op **CI-only**. Drop may reflect test condition change.

Residual decline

1. Predict $\hat{y}_i = f(\text{age}, \text{contraPTA}, \ldots)$
2. Residual: $r_i = \text{Post}_i - \hat{y}_i$
3. Flag if $r_i < -15$ pp

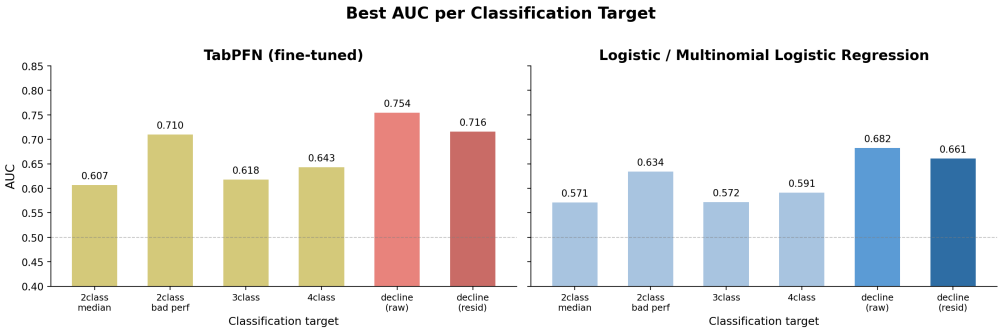
Good contra → good HA benefit → loses bilateral benefit with CI-only → $\hat{y}_i$ adjusted down.



Only patients who drop beyond what their profile predicts are flagged.

Risk Classification Results

Six targets — best AUC across model classes (TabPFN vs Logistic Regression)



Decline prediction is possible (AUC ≈ 0.72), even when regression fails ($R^2 \approx 0$).

Who Declines? First Look

Permutation importance (TabPFN) — raw decline definition, pre-op predictors only

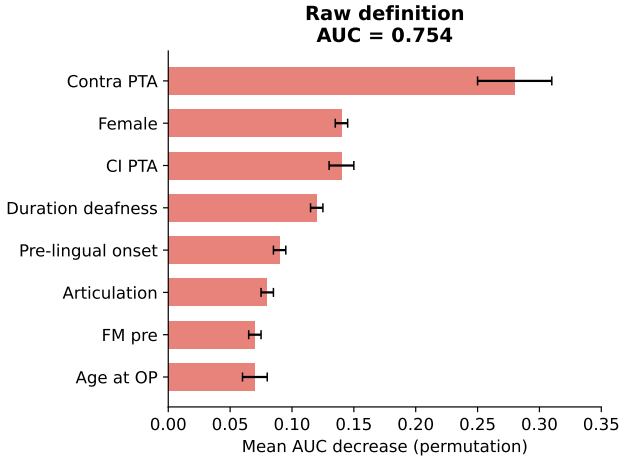
What drives the prediction?

Shuffle each feature, measure AUC drop.

- **Contralateral hearing** dominates
- Pre-lingual onset, duration of deafness also contribute
- CI PTA, articulation, age: lower-ranked but present

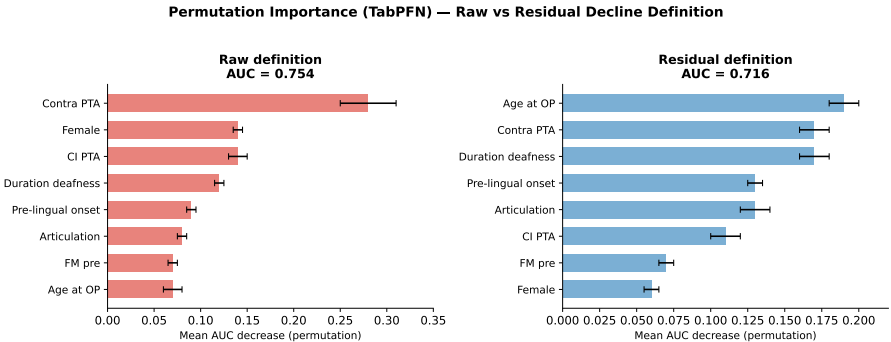
But is contralateral hearing a real risk factor?

Or is it the HA → CI test condition artefact?



Residual Decline: The Right Question

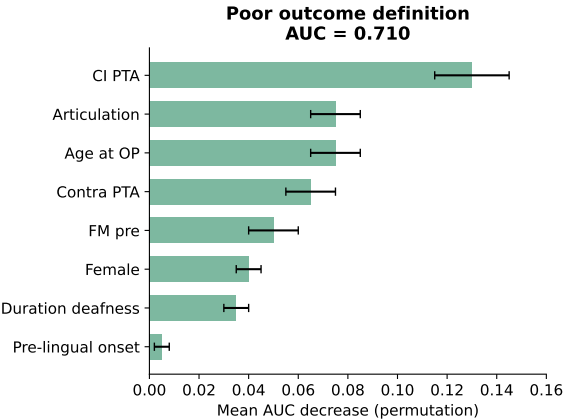
AUC remains strong (0.716) — predictor structure changes



Raw: AUC = 0.754 **Residual:** AUC = 0.716

What Predicts Poor Absolute Outcome?

Different target, different predictors — CI PTA dominates, not contralateral hearing



Decline: who underperforms their expectation? Confounded by HA → CI test condition.
Poor outcome: who is functionally impaired? Driven by the implanted ear (CI PTA).

Which Predictors Are Robust?

Comparing predictor importance across three outcome definitions

Deep Hearing Lab

Predictor	Raw decline	Residual decline	Poor performer	Robust?
Pre-lingual onset	Moderate	Moderate	Weak	**Yes (decline)**
Contralateral hearing	**Dominant**	Strong	Moderate	**Attenuated**
Duration of deafness	Moderate	Strong	Weak	**Yes (decline)**
CI PTA	Moderate	Weak	**Dominant**	**Yes (poor perf)**
Articulation	Weak	Moderate	**Strong**	Emerges
Age at OP	Weak	**Dominant**	**Strong**	**Emerges**

Decline (residual): age, duration, contralateral hearing, pre-lingual onset. **Poor outcome**: CI PTA, articulation, age.

Different questions → different risk factors → different interventions.

Discussion

Open questions for the community

Deep Hearing Lab

1. **Who should we target — decliners or poor performers?**
   They flag *different* patients with *different* predictors at *different* timepoints. A complete risk evaluation needs both — shall we target more?

2. **Is pre-operative prediction the right goal?**
   Early post-op trajectory (6 months) predicts final outcome 5× better than any pre-op variable ($R^2 = 0.36$ vs 0.03). We have preliminary results — should we invest more in post-op monitoring?

3. **What data would break the ceiling?**
   Electrode imaging, cortical responses, cognitive measures — at what cost and clinical feasibility?

Summary

Summary

Deep Hearing Lab

What we did

Swiss National CI Database, Zurich centre. 771 patients, 85 pre-op features, 5 speech tests.

Part 1 8 algorithms $\times$ 5 feature sets
Temporal validation, conformal PI

Part 2 CI Equalizer, Simpson's Paradox, mathematical coupling

Part 3 6 classification targets, raw vs residual decline

What we found

1. **Hard ceiling**: pre-op data cannot predict individual CI outcomes ($R^2 \leq 0.10$ across all models)
2. **Three reasons why**: CI Equalizer Effect, Simpson's Paradox, mathematical coupling
3. **Risk prediction works**: decline can be flagged pre-op (AUC = 0.72); pre-lingual onset is the key predictor
4. **Definition matters**: raw decline is biased — residual decline (LOO) is the right question

Dr. Marta Campi · marta.campi@uzh.ch · Deep Hearing Lab, University Hospital Zurich

Thank you

Questions?

Dr. Marta Campi

Deep Hearing Lab · University Hospital Zurich

marta.campi@uzh.ch

Prof. Alexander Huber · Prof. Tobias Goehring

Swiss National CI Database (Zurich centre, N = 771)

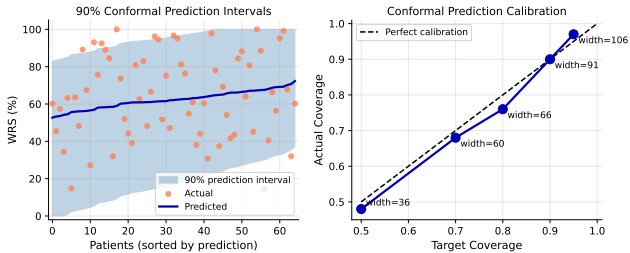
Appendix

Four Lines of Evidence That the Ceiling Is Real

Not artefact, not modelling failure — genuine information limit

Deep Hearing Lab

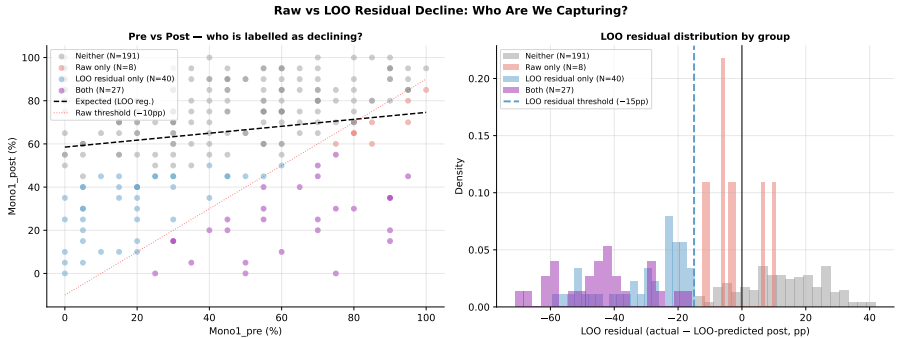
1. $R^2 < 0.03$ **across all models**
   LR, RF, GBM, TabPFN converge. If signal existed, one would find it.
2. **Features make it worse**
   $3 \rightarrow 85 \rightarrow 226$: CV R^2 degrades.
3. **Conformal intervals span the full scale**
   90% PI: ± 41 pp. Distribution-free. Honest.
4. **Fails at every timepoint**
   6mo, 12mo, 24mo+: CV $R^2 < 0$ everywhere. Equalizer is **immediate**.



PI band spans nearly the full scale (± 41 pp). Calibration curve confirms interval honesty.

The Raw Definition Gets It Wrong

8 false positives, 40 genuine decliners missed — same raw drop, different mechanisms



Group	N	Pre	Post	LOO Resid
Raw only	8	87%	69%	−3 pp *performed as expected — not true decliners*
LOO only	40	20%	32%	−30 pp *genuine deterioration — invisible to raw definition*
Both	27	62%	24%	−44 pp

What Should We Predict?

Outcome definition is a scientific choice — not a convention

Deep Hearing Lab

Definition	Clinical question	Limitation	Who?
Raw decline (>10 pp)	Who loses function?	Confounded by baseline	—
Residual decline (LOO)	Who underperforms?	Depends on regression	Marta
Bad performer (<40%)	Who needs support?	No change information	Tobi

Decline and poor performance flag **different patients** with **different predictors**.
A complete risk evaluation needs both — or perhaps more.

Next step: pre-operative risk stratification — pending external validation (Bern, Basel).