

Interpolated Supervision Can Change What Regression Benchmarks Reward

Siddhant Agarwal

siddhant4583agarwal@gmail.com

Research Article

Keywords: Target construction, Interpolated supervision, Regression benchmarks, Model evaluation, Label provenance, Parkinson's telemonitoring

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11400343/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Interpolated Supervision Can Change What Regression Benchmarks Reward

Anonymous author

Abstract

Dense regression labels may be constructed from sparse temporal observations rather than independently measured at each prediction time. We investigate how this changes what a benchmark rewards, even when training and evaluation observations are separated. An exact squared-risk identity expresses pairwise ranking changes as the alignment between label-construction discrepancy and prediction differences. A conservative norm bound supplies a sufficient ranking-preservation check. We audit the explicitly interpolated Unified Parkinson’s Disease Rating Scale targets in a 42-participant Parkinson’s telemonitoring release, without interpreting them as independent daily clinical ratings. Two measured-reference sensor cases then isolate evaluation-only changes and learning under interpolated supervision across 30 distinct anchor configurations and eight candidate regressors. In air-quality reference-trained comparisons, changing only evaluation targets reverses 74 of 140 candidate pairs at a 12-row anchor spacing; corresponding energy comparisons show 18 of 140 reversals at a 36-row spacing. The sufficient bound has no observed violations but frequently cannot certify preserved rankings. Across 320 synthetic cohort settings, periodic components omitted by anchors produce systematic ranking differences, while affine-target controls preserve every ranking. Reference-label validation audits yield mixed, sometimes adverse, selection changes. These experiments demonstrate target-dependent model comparison rather than clinical efficacy or a new prediction architecture. The clinical case cannot quantify daily symptom error without independent off-anchor assessments. Code, saved predictions, numerical checks and source provenance support reproducibility.

Keywords: Target construction, Interpolated supervision, Regression benchmarks, Model evaluation, Label provenance, Parkinson’s telemonitoring

1 Introduction

Dense regression labels need not represent independently observed responses. Expanding sparse temporal measurements by interpolation constructs a target that describes a rule as well as the underlying process. A predictor can reproduce that rule while missing between-anchor variation; a predictor following measured variation can score worse against the constructed target. Separating training and evaluation observations does not resolve this distinction when both share the construction.

We ask whether interpolation changes which of several fixed prediction vectors is preferred, keeping their input rows and evaluation observations unchanged. We distinguish a reference response y from a constructed response $\tilde{y}$. Interpolation may be appropriate when the intended response is a smooth trajectory. A mismatch arises when its evaluation loss supports a claim about another response, such as fluctuations between clinical assessments.

The University of California, Irvine (UCI) Parkinsons Telemonitoring release explicitly describes motor and total Unified Parkinson’s Disease Rating Scale (UPDRS) scores as

linearly interpolated [12, 13]. We audit released-label geometry and patient-disjoint regression, but independent clinical ratings at every voice timestamp are unavailable. To measure target-dependent rankings against an observable reference, we also construct interpolation inside chronological partitions of air-quality and appliance-energy records. Known-response simulations add affine controls and phase/cadence sensitivities. These are target-construction audits, not intervention or state-of-the-art clinical prediction claims.

The study makes three contributions:

1. Matched comparisons separate *evaluation-only* target changes from fitting and selection under constructed supervision.
2. An exact squared-risk identity and conservative preservation bound explain pairwise changes and are checked against saved predictions. They use elementary algebra, not a new general learning theorem.
3. Clinical provenance, two measured-reference cases and simulations are linked to an audit of validation selection, retaining adverse outcomes and stable controls.

Code, configurations, source checksums and predictions support the empirical comparisons. The scientific distinction is between reconstructing a released response and validating the response that an application claims to monitor.

2 Related work and position

Measurement and target operationalization.

Jacobs and Wallach [8] distinguish a theoretical construct from the measurement used to operationalize it. We adopt that distinction for a numerical response: the target supplied to a learner need not equal the response about which an application makes a claim. Here the operator is temporal interpolation. Dataset documentation, including the collection and preprocessing history advocated by Gebru et al. [4], is important because a dense target column alone does not disclose independent observation times.

Label quality and benchmark stability.

Northcutt et al. [9] show that test-label errors can affect comparisons between classifiers. Zhou et al. [18] study numerical-response errors and regression-based detection. These establish that evaluation labels matter. Our controlled perturbation is intentional interpolation with a known operator, rather than a label-cleaning or error-detection procedure.

Interpolation in regression augmentation.

RegMix and C-Mixup already examine limitations of interpolating continuous labels and propose training augmentation approaches [7, 17]. Thus, the broad observation that interpolated regression labels can be inappropriate is not new. Our experiments keep the original input rows and distinguish changing their scoring response from fitting on that response. We do not compare their augmentation algorithms.

Ambiguity beyond one held-out score.

D’Amour et al. [3] investigate pipelines that admit predictors with similar in-domain performance and different deployment behavior. The present problem can occur even with fixed predictions: replacing y with $\tilde{y}$ changes the metric and can reverse a ranking. Participant or temporal separation is maintained; the question is which target that evaluation rewards.

Clinical and tabular regression.

The original telemonitoring work investigates speech-based replication of symptom assessment [13]. More recent approaches address feature robustness in that task, including NoRo [11]. We do not reimplement those methods or claim that their reported results are incorrect. Our clinical regressors use a different, patient-disjoint protocol, so their scores are not a directly comparable leaderboard. The selected linear, ensemble and small neural baselines provide several inductive biases for a target audit. The broader tabular study of Grinsztajn et al. [5] motivates reporting more than one family, but this limited grid cannot support a universal comparison of tree and neural architectures.

3 Target construction and what the evaluation identifies

3.1 Observed anchors, constructed labels and reference responses

Let x_i be an input measured at time t_i . In a controlled reference case, a recorded response y_i is also available. An anchor set A contains selected times and their responses. Piecewise-linear interpolation produces $\tilde{y} = I_A[y]$. It preserves y_i at the anchors and replaces the response elsewhere by a line between adjacent anchor values. Define the construction discrepancy

$$d_i = \tilde{y}_i - y_i. \quad (1)$$

This discrepancy is defined relative to the selected reference. It is not automatically a measurement error or a clinical invalidity score. The sensor reference may itself contain noise, and some applications may intentionally target a smoothed response. The analyst must first state which quantity is of interest.

In the clinical release, the available target is already a constructed response. We observe that response and the voice features, but not y_i as an independently collected rating at every recording. We therefore cannot estimate d_i in the Parkinson case. Fitting a curve to the released scores or reconstructing them from a sparse subset does not recover the missing clinical reference. The measured-reference and synthetic experiments address the ranking question that the clinical release alone cannot answer.

For a fixed prediction vector $f = (f_1, \dots, f_n)$, define weighted squared risk

$$R_y(f) = \sum_{i=1}^n w_i (f_i - y_i)^2, \quad \|v\|_w^2 = \sum_{i=1}^n w_i v_i^2, \\ w_i \geq 0, \quad \sum_i w_i = 1. \quad (2)$$

The chronological sensor experiments use equal row weights. The clinical evaluation averages participant-level errors; the synthetic evaluation has the same number of test rows per patient, so participant-balanced and pooled squared risks agree. Root mean squared error (RMSE) is

reported in native units, while the ranking analysis uses squared risk.

3.2 A finite-sample ranking identity

Proposition 1 (Ranking change for a fixed pair) *For two fixed prediction vectors f, g evaluated on the same rows and weights,*

$$[R_y(f) - R_y(g)] - [R_{\tilde{y}}(f) - R_{\tilde{y}}(g)] = 2 \sum_i w_i d_i (f_i - g_i). \quad (3)$$

Consequently,

$$|[R_y(f) - R_y(g)] - [R_{\tilde{y}}(f) - R_{\tilde{y}}(g)]| \leq 2 \|d\|_w \|f - g\|_w. \quad (4)$$

Proof Expanding $(f_i - y_i - d_i)^2$ gives $R_{\tilde{y}}(f) = R_y(f) - 2 \sum_i w_i d_i (f_i - y_i) + \|d\|_w^2$. Subtract the analogous expression for g ; the $\|d\|_w^2$ terms cancel. Applying Cauchy-Schwarz to the remaining weighted inner product gives Equation 4. $\square$

Corollary 1 (Sufficient ranking preservation) *If $|R_{\tilde{y}}(f) - R_{\tilde{y}}(g)| > 2 \|d\|_w \|f - g\|_w$, then the two nonzero risk differences have the same sign.*

The interval around the constructed-target risk difference with radius $2 \|d\|_w \|f - g\|_w$ excludes zero under the condition, proving the corollary. Failure of the condition does not imply a reversal. It means this sufficient check cannot certify the pair. If $d = 0$, any non-tied ranking is preserved. If d aligns with $f - g$, the inner-product term is large; its direction determines which candidate is favored by the changed target.

The calculation conditions on already fitted predictions and observed evaluation labels. It does not supply a confidence interval for new samples, account for model-estimation uncertainty, or certify performance where the reference response is unavailable. In particular, the bound cannot be numerically applied to the unknown daily clinical response by substituting reconstruction error of the released UPDRS labels for $\|d\|_w$.

3.3 Off-anchor nonidentifiability

Proposition 2 (Interpolation does not identify an unconstrained response) *For a fixed set of observed anchors, let $h(t)$ vanish at every anchor but be nonzero*

at at least one unobserved time. Then $I_A[y + h] = I_A[y]$. Without an additional restriction linking inputs or trajectory shape to the response, the anchors and constructed targets do not determine the off-anchor reference trajectory.

Proof The interpolation operator uses only the values at the anchors. Those values are unchanged by h , so every interpolated value is unchanged. Both trajectories remain consistent with the observed anchors and constructed target. Keeping the recorded input features fixed does not eliminate this ambiguity unless a specified response–feature relation rules out one of the trajectories. $\square$

This is an elementary limitation of sparse observation, not evidence that interpolation is unsuitable for every clinical task. Smoothness constraints, domain models or independent assessments could restrict the admissible h . We do not estimate such constraints from this release. The proposition explains why a large number of interpolated training rows cannot, by itself, validate otherwise unobserved fine-scale symptom changes.

4 Executed study design

4.1 Use of generative artificial intelligence

OpenAI ChatGPT assisted with study design, experimental code development and execution, interpretation of results, and manuscript drafting and editing. The supplied scripts were executed, and saved predictions and numerical checks accompany the study. ChatGPT is not an author. The human author retains responsibility for the study design, code, interpretation, citations and submitted conclusions.

4.2 Three sources with different roles

Table 1 summarizes the sources and their distinct roles. The Parkinson release is the motivating clinical case, while the other two datasets support controlled comparison against retained measured responses. The air-quality reference is recorded carbon monoxide concentration, and predictors are the five physical sensor channels plus temperature and humidity variables [15, 16]. Missing target codes are excluded; missing predictors remain

for training-fitted imputation. No other reference-analyzer concentration is supplied as a predictor. The appliance-energy case uses temperature, humidity and weather predictors [1, 2]. Light-energy readings and the source’s two random variables are excluded. Calendar features describe elapsed time and cyclic time-of-day/week; they are distinct from the target and its interpolation anchors.

Each public object is downloaded with a retained SHA-256. The actual file dimensions and timestamp counts are reported from parsing rather than substituted from repository summary metadata. UCI identifies the releases under CC BY 4.0. Its air-quality description additionally states research-only use; this work uses it for research and does not assert permission for a commercial deployment. The original investigators retain attribution. No new clinical recordings, annotations or human participants were collected.

4.3 Clinical geometry and patient-separated regressors

Two clinical analyses use different information. First, a *released-label reconstruction diagnostic* selects sparse labels from each person’s released sequence on 7-, 14-, 28- or 56-day grids, with five offset settings. The closest available recording time is used for each grid point, and the first and last available times are anchors. All other unique recording times are evaluated. Exact duplicate times are averaged only for this geometric diagnostic. Interpolation here sees anchor labels on both sides of an evaluated time: it is an offline reconstruction experiment, not an admissible online clinical predictor or a new-patient model. These anchors are sampled released values, not recovered clinic-visit ratings.

A descriptive straight-line and one-knot fit is also computed for each participant/target. The knot search uses the interior 15–85% of available recording times. Its error is descriptive and in sample. A fitted knot is not identified as an original assessment date. These calculations measure the shape of the released target and do not estimate clinical truth between assessments.

Second, five outer folds partition all 42 participants using a recorded shuffled seed. Within each outer training group, three patient-separated inner folds choose regression parameters by

Table 1 Distinct roles of the sources. Counts refer to the actual parsed analysis inputs. Reference means a recorded response, not error-free latent truth

Source	Rows	Independent units	Role and response
Parkinson telemonitoring	5,875	42 participants	Clinical provenance audit; released interpolated motor/total UPDRS; 16 voice predictors.
Air Quality	7,674	One time series	Recorded CO reference (mg/m^3); eight physical/environmental predictors.
Appliances Energy	19,735	One building	Recorded appliance energy (Wh); 24 environmental/weather predictors.
Simulation	24,200/cohort	200 patients/cohort	Known trend plus periodic residual; 150 train/50 held out; 20 seeds in each of 16 regimes.

participant-macro mean absolute error (MAE). Candidate inputs are the 16 voice measures, age/recorded-sex/time metadata, or their union. The subject identifier and the other UPDRS target are excluded. We fit separate motor and total models using a training-mean baseline, ridge, random forest, histogram gradient boosting and a small multilayer perceptron (MLP). Every outer test person’s recordings are absent from training and selection. The clinical table concerns prediction of the *released interpolated scores*; it is not error against independent per-recording clinical ratings.

4.4 Chronological measured-reference experiments

For each sensor dataset, the first 60% of retained chronological observations train the models, the next 20% select parameters, and the final 20% evaluate predictions. We interpolate targets *separately within each partition*. No validation or test anchor supplies a training label. Every partition includes its first and last row as anchors, plus a regular stride with a predefined offset. Input observations and training rows are unchanged across target representations. Calendar variables are timestamp functions; predictor imputation and scaling use training observations only.

The air-quality strides are 1, 3, 12 and 24 retained observed-target rows; appliance-energy strides are 1, 6, 36 and 72 rows. Stride one is the exact-reference control. A stride is an observation count, not necessarily a fixed elapsed-time gap after missing air targets are removed. Nontrivial strides use up to five distinct offsets, 0 through

$\min(4, k - 1)$. We retain true timestamps for interpolation. The paper reports all evaluated settings rather than selecting the largest observed reversal as a representative rate.

For *evaluation-only* perturbation, we hold predictions from reference-trained models fixed and score the same test rows against both responses. For *interpolated training*, candidates are trained on the constructed training response and their parameters chosen against constructed validation responses. Models are not subsequently refitted with validation observations. Their saved test predictions are then evaluated against both reference and constructed responses. Each experiment distinguishes a changed target from a changed predictor.

4.5 Candidate families and numerical fitting

Table 2 lists the candidate family grids and fitting budgets. All learned candidates use a training-fitted median imputer and input standardization. Targets are standardized within training through a transformed-target regressor, then predictions returned to native units. The MLP has one 32-unit rectified linear unit (ReLU) layer, Adam, batches of at most 128 and a 200-iteration limit; it has no internal early stopping. Boosting permits the implementation’s automatic training-internal early stopping when its training-set size triggers it; the external chronological validation block still selects parameters. Actual iteration counts are retained. Reaching an iteration budget is not claimed to prove optimizer convergence.

The sensor study has eight candidate feature/family combinations: mean; calendar ridge

Table 2 Candidate family grids for the clinical and measured-reference cases. All predictor imputation, predictor scaling and response scaling are fit on training rows only. The simulation fixes ridge $\alpha = 0.1$ and boosting $L_2 = 10$

Family	Fixed budget/settings	Tuned values
Ridge	Linear, standardized predictors and response	$\alpha \in \{0.1, 10, 1000\}$
Random forest	100 trees, feature fraction 0.8	Leaf minimum $\{5, 20\}$
Histogram boosting	Maximum 120 iterations, 15 leaves, leaf minimum 20; library automatic early stopping	$L_2 \in \{1, 10\}$
MLP	One hidden layer, 32 ReLU units, Adam, batch 128, maximum 200 iterations; early stopping disabled	$\alpha \in \{0.001, 0.1\}$
Mean	Training-row response mean	None

and boosting; sensor ridge, boosting, forest and MLP; and combined calendar/sensor boosting. The synthetic study uses calendar ridge, sensor ridge, combined ridge and combined boosting. These fixed configurations do not establish a best clinical algorithm. Software versions are pinned [6, 10, 14].

4.6 Reference-validation audits and ranking statistics

Among the candidates trained under interpolated supervision, the proxy choice minimizes constructed validation mean squared error (MSE). We also choose candidates using 5%, 10%, 25% or all non-anchor validation reference values, with a nested seeded sample of timestamps. These reference values were retained from the source but withheld from the constructed-supervision experiment. They are not newly collected expert labels. A nominal audit fraction is not an estimate of clinical annotation cost or independent effective sample size.

The chosen candidate is evaluated on the held-out chronological test block. For descriptive selection regret, a test-reference oracle is the best of the same fitted candidates on that test block. The oracle only defines an analysis reference; it is not used to choose the deployed candidate. Comparing proxy and audited validation choices is the actionable model-selection experiment. We report adverse as well as favorable changes, since a finite validation audit can choose a candidate that performs worse in the subsequent period.

Pairwise reversal occurs when two non-tied MSE differences, evaluated on the same prediction vectors, have opposite signs under the two targets. Comparisons within 10^{-8} of a tie are not counted as strict reversals. For each pair, we compute

Equation 3 and the sufficient ranking check. Offset settings reuse the same time series and are not independent experimental cohorts. We describe their spread without treating it as a population confidence interval.

4.7 Controlled simulations

For patient s and integer time $t \in \{0, \dots, 120\}$, the reference response is

$$\begin{aligned} y_s(t) &= b_s + a_s t + \gamma u_s \sin(2\pi t/60 + \phi_s), \\ x_s(t) &= y_s(t) + \epsilon_s(t). \end{aligned} \quad (5)$$

We draw $b_s \sim N(25, 5^2)$, $a_s \sim N(0.03, 0.01^2)$ and $u_s \sim U(0.5, 1.5)$. The sensor error is $\epsilon \sim N(0, \sigma^2)$. Calendar candidates receive b_s and $a_s t$, an intentionally idealized observed trend comparator. These generator parameters are not claimed available as clinical covariates. Sensor candidates receive $x_s(t)$ and combined candidates receive both. There are 200 simulated patients; the first 150 train and the remaining 50 are held out in each seeded generation.

The core grid uses amplitude $\gamma \in \{0, 2, 5, 10\}$, sensor noise $\sigma \in \{0.2, 1, 3\}$, aligned phase zero and anchor stride 60. The sinusoidal component vanishes at these anchors, making its omission explicit. A sensitivity grid fixes amplitude five and noise one while varying aligned/random phases and anchor strides 30/40. Twenty independently seeded cohorts are generated per setting; matching seeds across settings share generator draws. Both reference and interpolated training responses are examined with fixed family parameters, without test-driven hyperparameter search. The simulations illustrate a possible mechanism, not the prevalence or magnitude of hidden symptom variation in Parkinson’s disease.

5 Results

5.1 Clinical label geometry and patient-disjoint prediction

The 5,875 recordings represent 42 participants. Across participants and five offsets, mean non-anchor reconstruction MAE for released motor UPDRS is 0.000274 points at the seven-day nominal schedule, 0.005050 at 14 days, 0.027440 at 28 days and 0.110776 at 56 days. Total-UPDRS errors are 0.000281, 0.005246, 0.029244 and 0.120759. Mean selected-anchor counts are approximately 24.92, 13.97, 7.96 and 4.93. Table 3 gives the complete summaries.

These small errors quantify reconstruction of the released target using a subset of that target. They do not show that assessments seven days apart are clinically sufficient or that the interpolated curve represents independently observed daily severity. Both interior anchors and endpoints are drawn from the released response; later values are available to the reconstruction. The procedure is an offline diagnostic, not prospective monitoring. Figure 1 displays three fixed participant IDs, without inferred clinical-visit markers.

Patient-disjoint prediction is substantially harder. Metadata ridge has participant-balanced MAE 6.999 for motor UPDRS and 8.780 for total UPDRS. Voice-only ridge obtains 7.035 and 9.015; the training mean obtains 7.125 and 8.898. Flexible voice-only families range from 7.419 to 7.719 motor points and 9.392 to 10.016 total points. Combined predictors do not uniformly help: combined MLP reaches 8.318 motor and 12.051 total points. Figure 2 and Table 4 report every configuration.

Intervals are broad: motor metadata ridge has [5.939, 8.114], voice ridge [5.996, 8.147], and the mean [6.024, 8.269]. These descriptive intervals do not establish superiority for the small numerical differences. The geometry task receives released scores from the evaluated person, whereas the learning task receives no training scores from outer-test people. Their errors are not competing estimates of deployable prediction. A dense released target can be highly reconstructable from sparse released values while new-person prediction remains difficult; neither task identifies independently observed off-anchor symptom error.

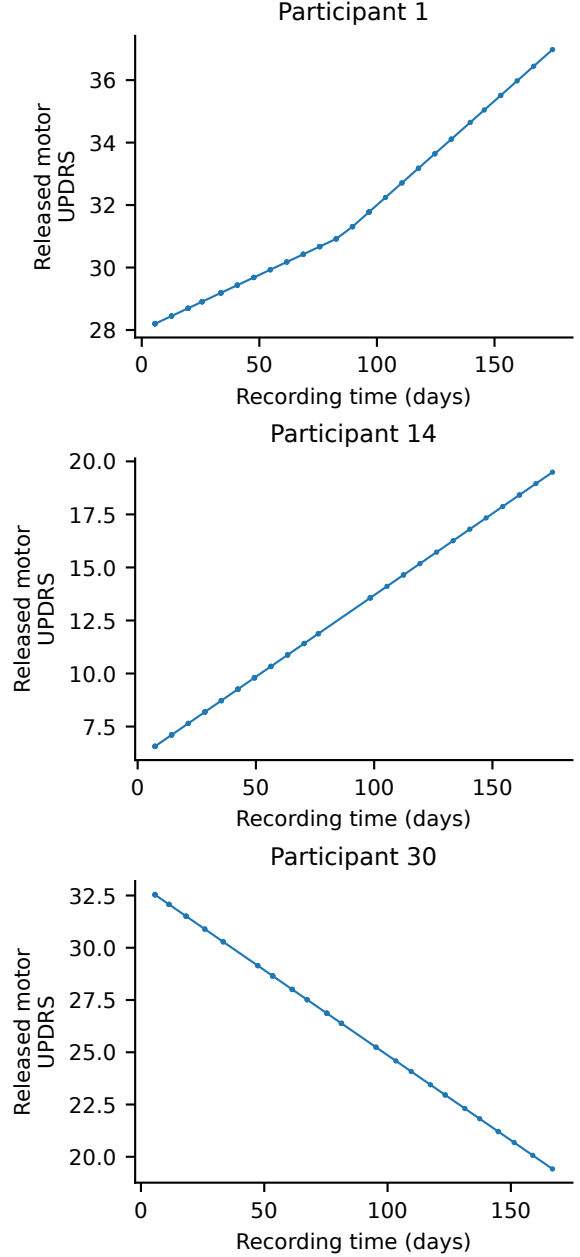


Fig. 1 Released motor-UPDRS curves for fixed illustrative participant IDs 1, 14 and 30. Points are labels at unique voice-recording times, not independent daily clinical assessments. Geometry and sparse reconstruction concern the released response only

5.2 Evaluation-only target changes reverse rankings

Reference-trained sensor candidates hold both fitting and predictions fixed while changing the scoring response. No-construction controls preserve all

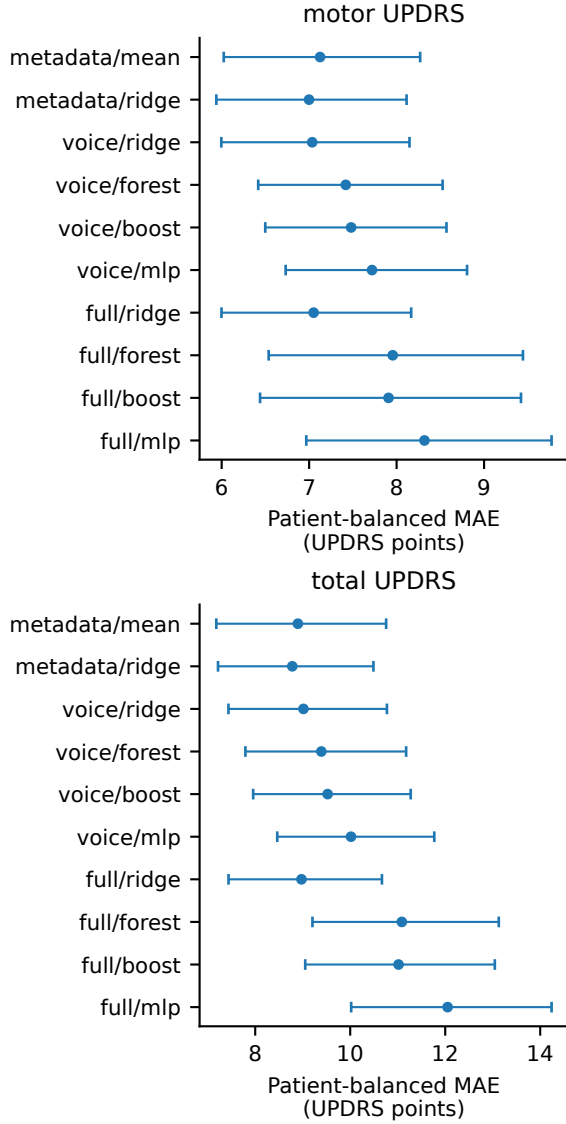


Fig. 2 Patient-balanced out-of-fold MAE against released interpolated targets. Intervals resample 42 participants 10,000 times conditional on fixed predictions; they do not include refitting or outer-split uncertainty. “Full” combines voice with age, sex and recording time

28 model pairs in both sources. Air Quality spacing three reverses six of 84 pairs across three distinct offsets (7.1%). Spacing 12 reverses 74 of 140 (52.9%), and spacing 24 reverses 69 of 140 (49.3%). Mean Kendall rank correlation is -0.057 at spacing 12 and 0.014 at spacing 24. A separated future test block therefore does not make rankings invariant to response construction.

Energy is more stable in this grid but still changes: four of 140 pairs reverse at spacing six

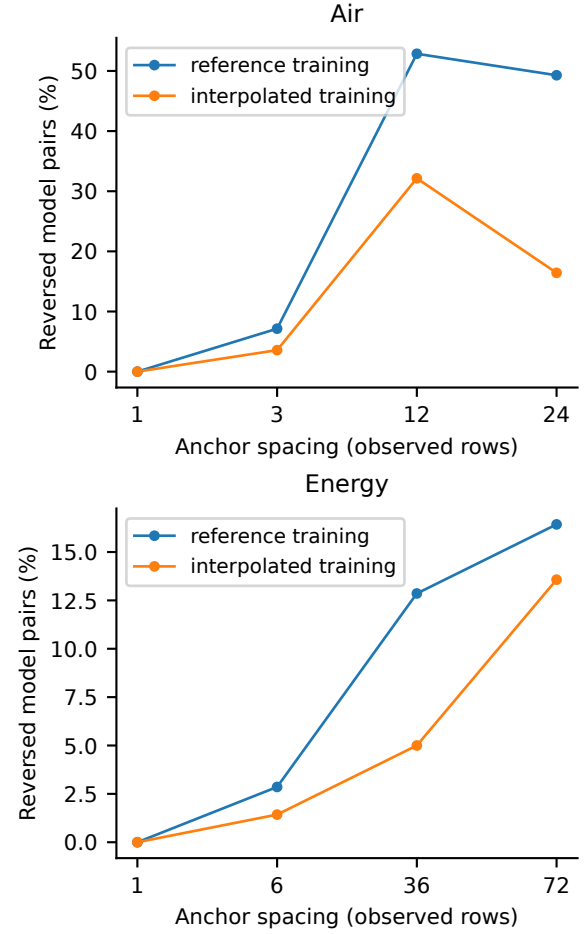


Fig. 3 Strict pairwise reversals between recorded-reference and constructed-target evaluation, aggregated over distinct selected offsets within each series. Reference-trained curves isolate scoring changes. Interpolated-trained curves use separately fitted/tuned candidate collections. Offsets are dependent sensitivities, not independent cohorts

(2.9%), 18 of 140 at spacing 36 (12.9%), and 23 of 140 at spacing 72 (16.4%). Mean correlations are 0.943, 0.743 and 0.671. These are differences between two selected series and grids, not a general comparison between all energy and air-quality tasks.

Mean construction RMSE is 0.534, 1.284 and 1.394 mg/m^3 for Air spacings 3/12/24 and 67.318, 90.028 and 102.538 Wh for Energy spacings 6/36/72. Reversal frequency need not grow monotonically with discrepancy. Air spacing 12 has more evaluation-only reversals than spacing 24 despite its smaller discrepancy. Equation 3 depends on alignment with prediction differences, not construction RMSE alone.

5.3 Constructed training changes the candidate collection

Candidates fitted and tuned under constructed supervision exhibit different sensitivities. Air has three of 84 reversals at spacing three (3.6%), 45 of 140 at spacing 12 (32.1%), and 23 of 140 at spacing 24 (16.4%). Energy has two, seven and 19 of 140 at spacings 6/36/72 (1.4%, 5.0%, 13.6%). The appendix retains both training modes and all offsets.

Fewer reversals do not establish better reference prediction. Learning a smoother target can make candidates more similar; similarly biased predictors may preserve their internal ordering. Actual reference losses must be inspected. Appendix Figure 7 displays every candidate’s two native-unit losses at the largest spacing and offset zero, a schedule-based illustration rather than the maximum-reversal case. Inputs exclude other reference gas columns in Air and lights/random features in Energy. Physical sensors still intentionally co-vary with the response; the controlled change is the artificial target construction.

5.4 The bound is correct but often uninformative

All saved measured-reference prediction tables satisfy the risk identity and bound within tolerance. Separate recomputation gives maximum absolute identity error 1.67×10^{-11} in squared native units. No certified pair reverses.

Excluding no-construction controls, the sufficient condition certifies 13 of 364 Air reference-trained pairs and zero of 420 Energy reference-trained pairs. Interpolated-trained collections certify 38 of 364 Air pairs and 53 of 420 Energy pairs. Most Energy rankings are preserved despite not being certified. Failure of the check is not a prediction that a pair will reverse.

This check uses reference discrepancy on the same test rows. It explains that table, without supplying a future-population confidence statement. The independent dense clinical reference is unavailable, so the check cannot certify daily symptom rankings in the Parkinson release. Extrapolating a validation-estimated discrepancy to a future test period would require additional assumptions absent here.

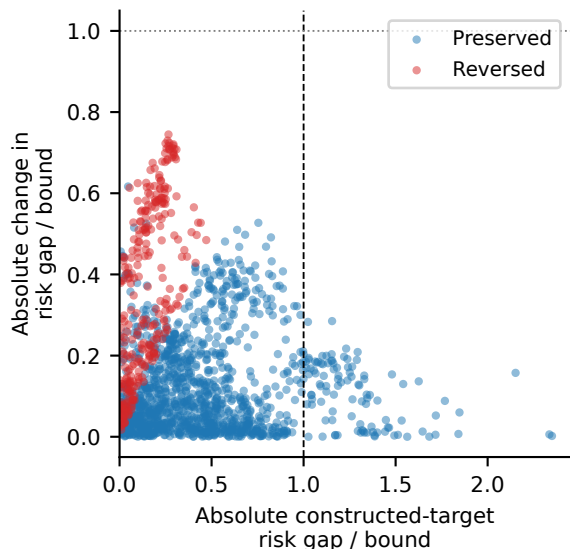


Fig. 4 Risk-gap changes relative to the deterministic bound. Horizontal ratios above one certify preservation. Vertical ratios do not exceed one within numerical tolerance. Many preserved pairs are left of the certification threshold. The horizontal display is limited to five for legibility

5.5 Validation reference audits give mixed selection results

Audits select among the same eight candidates already trained/tuned under constructed supervision. They do not retrain candidates and never select using future test reference values. A test-reference oracle describes the best available candidate after the fact only.

Across 13 constructed Air cases, proxy validation selection has mean reference test RMSE 0.965 mg/m^3 . Five-percent, 25-percent and full non-anchor reference audits leave the selected candidate unchanged in all cases. Ten-percent audits improve two cases and worsen none, reducing mean RMSE by 0.00238 mg/m^3 . More audited rows need not force a different choice: a nested subset can change the validation winner and later return to the original one.

Across 15 Energy cases, proxy selection has mean reference test RMSE 96.855 Wh . Five-percent audits yield 97.615 Wh : two cases improve, six worsen, seven remain unchanged. Ten-percent audits yield 92.625 Wh (two improved, three worse), 25-percent audits 92.873 Wh (three improved, four worse), and full audits

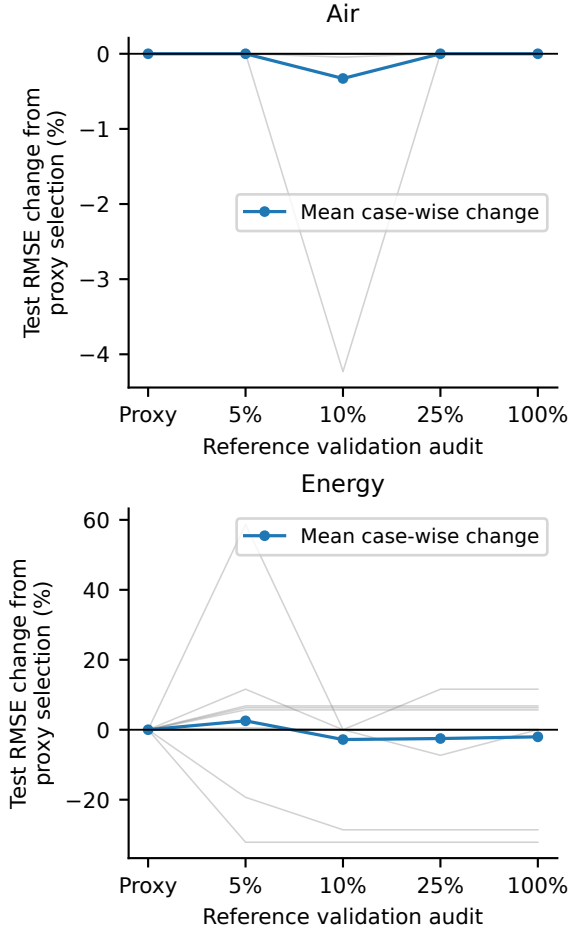


Fig. 5 Future reference test RMSE change after selecting fixed constructed-trained candidates with different validation responses. Gray lines are cases; blue averages case-wise relative changes. Positive is worse. Percentages refer to non-anchor validation rows, not independently collected clinical annotations or annotation cost

93.331 Wh (two improved, four worse). The ten-percent mean reduction of 4.230 Wh reflects case magnitudes; mean case-wise relative change is -2.80% , versus $+2.54\%$ for the five-percent audit.

Reference auditing aligns selection with an intended response but does not ensure temporal transfer. Candidate quality, drift and validation variability still matter. Unsuccessful Air audits and adverse Energy changes are part of the evidence. Intervention personalization would require observing and validating intervention outcomes, beyond this selection experiment.

5.6 Known-response simulations expose the mechanism and controls

All simulated test patients are unseen during training. Zero-amplitude trajectories are affine; interpolation reproduces their response and no pair reverses in any of the 60 seeded zero-amplitude scenarios under either training mode. This checks that the analysis does not manufacture changes when construction leaves the target unchanged.

Nonzero amplitude with aligned stride 60 removes the periodic component at the anchors. The idealized trend comparator can then be favored by constructed evaluation even though the sensor contains omitted-component information. At amplitude five/noise one, reference-trained candidates show 69 reversals among 120 pairs across 20 cohorts; interpolated-trained candidates show 104 of 120. The reference and constructed winners differ in all 20 cohorts for both training modes. Those 120 comparisons are six correlated model pairs per cohort, not 120 independent tasks.

At amplitude five/noise one, interpolated-trained reversals are 104/120 for aligned stride 30, 79/120 for aligned stride 40, 62/120 for random-phase stride 30 and 71/120 for random-phase stride 40. Reference-trained counts are 69, 70, 74 and 78. The exact core phase/cadence is therefore not the only tested setting with reversals, but magnitude depends strongly on scenario. Table 9 retains all controls.

These generators establish possibility, not clinical prevalence. They do not show that symptoms follow a sinusoid, that clinical scheduling suppresses a biological rhythm, or that trend parameters are available as covariates. The limited general mechanism is that an input can carry response information which a constructed evaluation has removed.

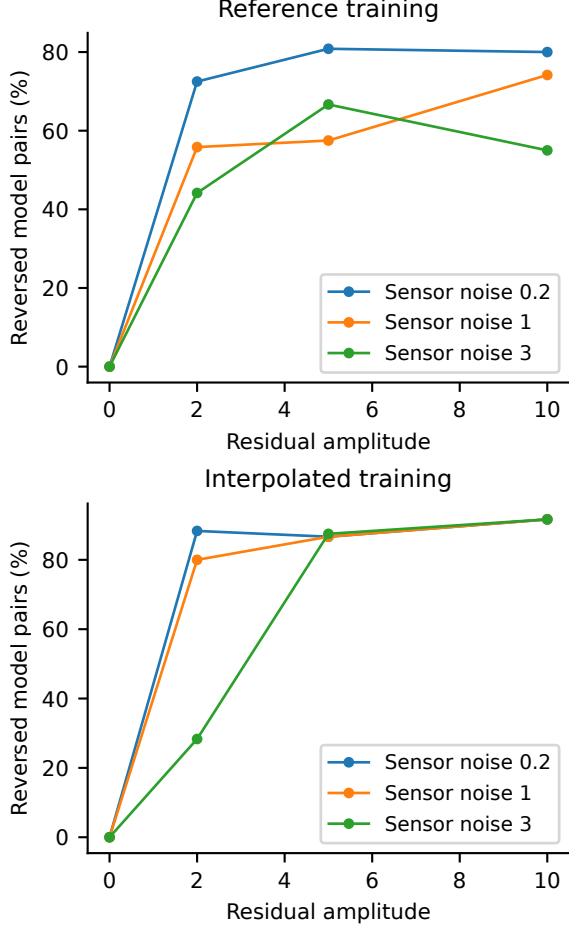


Fig. 6 Synthetic core grid with aligned phase/stride 60. Six pairs per cohort are aggregated over 20 independently seeded cohorts. Zero amplitude preserves every ranking. Frequencies vary with amplitude, noise and training response. Trend predictors are idealized generator observables

6 Interpretation, recommendations and limits

6.1 What a constructed-target ranking means

A ranking change need not indicate a poorly optimized estimator. Equation 3 shows that the alignment of construction discrepancy with prediction differences determines the change. Similar prediction vectors can remain similarly ranked despite substantial discrepancy; differences following an omitted component can reverse the ordering.

Reconstructing a constructed response and tracking an independently assessed process are

separate claims. The latter requires observations linking the construction to that process. Training/evaluation separation does not supply this evidence. Label availability also matters: an earlier timestamp’s target may use a later assessment, so timestamp order does not establish when its label was knowable.

Smoothing can define an appropriate estimand, such as average progression. Recorded sensor responses can themselves be noisy. These experiments therefore identify differences in rewarded behavior, rather than a universally preferable target.

6.2 A concrete target-provenance audit

Record the intended response, independently observed rows, anchor values/times, construction operator, its future-observation requirements, and evaluation definition. Keep unavailable information explicitly unknown. Reconstructing released clinical scores does not recover independent off-anchor assessments.

Where a measured reference exists, first evaluate fixed predictions against both definitions to isolate scoring changes. Then assess constructed training and validation separately. Reference-response selection should reserve independent validation measurements, including non-anchor times. Our audits improve, preserve or worsen future test error; they are not a guaranteed mitigation.

This workflow audits task specification, not clinical efficacy. A release lacking independent dense ratings can support constructed-target research, but cannot certify unknown daily symptom accuracy through released-label reconstruction.

6.3 Limitations and boundaries of the evidence

There is one Parkinson cohort and two single-series sensor cases. The clinical sample has 42 independent people, despite thousands of recordings. The sensor cases contain one monitored environment/building each. Chronological separation limits direct future-row leakage but does not create independent sites, remove sensor drift or guarantee transfer to a new environment. More clinical releases with observed rating times and

independently evaluated non-anchor assessments are needed to measure how often the mechanism matters in practice.

The interpolation schedules in the sensor study are artificial stress tests. Their spacing and offsets are selected settings, not an estimate of real clinical appointment patterns. Sensor targets are recorded interval summaries, not error-free latent states. The simulated trends and periodic residuals are stylized, and the ideal trend covariates make one comparator deliberately informative. Phase and cadence controls restrict the inference: a constructed target can preserve useful comparisons, and rank reversal is not inevitable.

The model grids are small, with finite neural/boosting budgets and one clinical outer split seed. Participant-bootstrap intervals condition on saved fits. Sensor offset summaries do not include independent-series uncertainty. Validation audits use temporally related points; their effective independent information may be smaller than their row count. The finite-sample certificate is deterministic for a given evaluation table and supplies no future-population risk bound.

No causal, diagnostic or intervention-response claim is estimated. Released scores do not establish independent daily symptom accuracy or support care recommendations. Clinical translation would require prospective outcome provenance, clinician-defined usefulness and independent validation.

7 Conclusion

Dense interpolated supervision can change a regression benchmark’s target and the behavior its score rewards. The clinical release illustrates a provenance limitation; measured-reference stress tests and controlled simulations quantify how candidate rankings and validation choices can change. An exact risk-difference identity explains the comparison, while a sufficient bound identifies some preserved pairs. The practical lesson is to identify the response being evaluated and distinguish reconstruction of a constructed label from validation of an independently assessed process. More fitted models or more interpolated rows do not resolve an unspecified target.

A Execution, partition details and computational checks

The archive separates acquisition, fitting, analysis and numerical verification. The internal protocol was written before model results and is not an external preregistration. Modular offsets are deduplicated by schedule identity: stride three has three distinct offsets. Initially repeated Air fits at offsets three/four were excluded without using their outcomes. The final analysis has 14 Air and 16 Energy cases, including one no-construction control in each source.

The checker recomputes 20 clinical configurations from saved predictions, checks all outer patient separations and predictor exclusions, and recomputes 480 sensor model records from 30 case files. It checks the pair identity, bound and certificate in all sensor tables. It independently regenerates/refits eight models in one fixed representative synthetic setting; all 2,560 synthetic fits were not rerun in this verification. All 1,680 clinical participant/target/schedule/offset reconstruction records are recomputed. These computational checks assess numerical consistency; they do not establish scientific validity or replace author responsibility.

A.1 Exact chronological partitions and features

Air parses 9,471 CSV rows, of which 9,357 have timestamps and 7,674 have CO responses after source missing codes are handled. Train/validation/test sizes are 4,604/1,535/1,535. Test dates are 2005-01-29 07:00 through 2005-04-04 14:00. Predictors are the five PT08 sensors plus temperature, relative and absolute humidity. Other recorded-reference gases are excluded. Interpolation uses actual time; row stride does not imply an exact number of hours after missing-target removal.

Energy has 19,735 rows and sizes 11,841/3,947/3,947. Test dates are 2016-04-30 08:20 through 2016-05-27 18:00. Predictors are nine temperature/humidity pairs, outdoor temperature/humidity, dew point, pressure, wind speed and visibility. Lights and two random variables are excluded. Calendar inputs are elapsed days and sine/cosine hour/weekday. First/last labels in each partition remain anchors. No

anchor crosses a partition boundary, but later anchors within a partition are used: this is offline supervision.

Clinical inner tuning combines three-fold GroupKFold predictions before averaging participant absolute error. Outer splits use shuffled five-fold participant IDs with seed 20261014. Age, sex and recording time comprise metadata; full inputs add 16 voice features. Neither the other outcome nor participant identifier is a predictor. The mean baseline weights training recordings equally; evaluation weights participants equally.

Sensor candidates minimize validation MSE against their training-response definition. Validation data are not added to training. Audits select from fixed candidates using nested seeded non-anchor subsets, with at least ten rows. The rows are temporally correlated and are retained source responses, not newly adjudicated gold labels. Clinical bootstrap intervals condition on saved fits. Sensor offset ranges are descriptive; synthetic seeds are independent within regime and matched across regimes.

B Complete clinical geometry and predictive results

Table 3 Released-target non-anchor reconstruction averaged over participant and offset. Nominal day spacing selects nearby recording times; this is an offline diagnostic

Target	Days	MAE	Mean anchors
motor_UPDRS	7	0.000274	24.92
motor_UPDRS	14	0.005050	13.97
motor_UPDRS	28	0.027440	7.96
motor_UPDRS	56	0.110776	4.93
total_UPDRS	7	0.000281	24.92
total_UPDRS	14	0.005246	13.97
total_UPDRS	28	0.029244	7.96
total_UPDRS	56	0.120759	4.93

Table 4 All patient-disjoint configurations. Units are released-score points; bootstrap intervals condition on fixed predictions

Inputs	Family	MAE	95% interval	RMSE
<i>Target: motor_UPDRS</i>				
metadata	mean	7.125	[6.024, 8.269]	8.357
metadata	ridge	6.999	[5.939, 8.114]	8.164
voice	ridge	7.035	[5.996, 8.147]	8.270
voice	forest	7.419	[6.418, 8.526]	8.854
voice	boost	7.480	[6.498, 8.570]	8.934
voice	mlp	7.719	[6.732, 8.805]	9.265
full	ridge	7.051	[5.998, 8.167]	8.264
full	forest	7.956	[6.537, 9.444]	9.622
full	boost	7.908	[6.439, 9.422]	9.670
full	mlp	8.318	[6.967, 9.772]	10.117
<i>Target: total_UPDRS</i>				
metadata	mean	8.898	[7.182, 10.757]	11.007
metadata	ridge	8.780	[7.217, 10.489]	10.615
voice	ridge	9.015	[7.437, 10.773]	11.029
voice	forest	9.392	[7.794, 11.177]	11.651
voice	boost	9.525	[7.957, 11.273]	11.811
voice	mlp	10.016	[8.464, 11.770]	12.466
full	ridge	8.974	[7.438, 10.668]	10.831
full	forest	11.087	[9.205, 13.128]	13.286
full	boost	11.018	[9.053, 13.044]	13.232
full	mlp	12.051	[10.020, 14.239]	14.709

The in-sample geometry fits all labels of a person and does not estimate held-out clinical accuracy. For motor_UPDRS, median linear/one-knot RMSE is 0.8074/0.0173. For total_UPDRS, median linear/one-knot RMSE is 0.7632/0.0159. A fitted knot is a curve parameter, not an inferred assessment date.

C All measured-reference ranking configurations

Eight candidates give 28 pairs per row. Cert. counts sufficient preservation checks, rather than all preserved pairs. Offsets reuse one physical series.

Table 5 Air risk-rank comparison across all distinct schedules

Stride	Offset	Training	Rever- sals	Pairs	τ	Cert.
1	0	interpolated	0	28	1.000	28
1	0	reference	0	28	1.000	28
3	0	interpolated	0	28	1.000	5
3	0	reference	2	28	0.857	4
3	1	interpolated	1	28	0.929	4
3	1	reference	2	28	0.857	4
3	2	interpolated	2	28	0.857	5
3	2	reference	2	28	0.857	5
12	0	interpolated	3	28	0.786	0
12	0	reference	12	28	0.143	0
12	1	interpolated	9	28	0.357	0
12	1	reference	15	28	-0.071	0
12	2	interpolated	13	28	0.071	0
12	2	reference	16	28	-0.143	0
12	3	interpolated	9	28	0.357	3
12	3	reference	16	28	-0.143	0
12	4	interpolated	11	28	0.214	0
12	4	reference	15	28	-0.071	0
24	0	interpolated	5	28	0.643	0
24	0	reference	14	28	0.000	0
24	1	interpolated	5	28	0.643	3
24	1	reference	14	28	0.000	0
24	2	interpolated	4	28	0.714	7
24	2	reference	15	28	-0.071	0
24	3	interpolated	3	28	0.786	7
24	3	reference	13	28	0.071	0
24	4	interpolated	6	28	0.571	4
24	4	reference	13	28	0.071	0

Table 6 Energy risk-rank comparison across all distinct schedules

Stride	Offset	Training	Rever- sals	Pairs	τ	Cert.
1	0	interpolated	0	28	1.000	28
1	0	reference	0	28	1.000	28
6	0	interpolated	0	28	1.000	11
6	0	reference	0	28	1.000	0
6	1	interpolated	0	28	1.000	2
6	1	reference	1	28	0.929	0
6	2	interpolated	0	28	1.000	0
6	2	reference	2	28	0.857	0
6	3	interpolated	1	28	0.929	5
6	3	reference	0	28	1.000	0
6	4	interpolated	1	28	0.929	5
6	4	reference	1	28	0.929	0
36	0	interpolated	1	28	0.929	4
36	0	reference	5	28	0.643	0
36	1	interpolated	1	28	0.929	6
36	1	reference	2	28	0.857	0
36	2	interpolated	2	28	0.857	0
36	2	reference	5	28	0.643	0
36	3	interpolated	1	28	0.929	0
36	3	reference	3	28	0.786	0
36	4	interpolated	2	28	0.857	8
36	4	reference	3	28	0.786	0
72	0	interpolated	2	28	0.857	0
72	0	reference	4	28	0.714	0
72	1	interpolated	3	28	0.786	0
72	1	reference	5	28	0.643	0
72	2	interpolated	10	28	0.286	0
72	2	reference	6	28	0.571	0
72	3	interpolated	3	28	0.786	0
72	3	reference	5	28	0.643	0
72	4	interpolated	1	28	0.929	12
72	4	reference	3	28	0.786	0

D Validation audit results, including adverse changes

Summaries exclude stride-one controls. Cases are dependent schedules within a series. Lower native-unit RMSE delta is favorable; relative changes average individual case percentages.

Table 7 Reference-validation selection across all constructed cases

Source	Audit	Cases	RMSE	Delta	Relative (%)	Better/ worse
air	0%	13	0.965	0.000	0.00	0/0
air	5%	13	0.965	0.000	0.00	0/0
air	10%	13	0.962	-0.002	-0.33	2/0
air	25%	13	0.965	0.000	0.00	0/0
air	100%	13	0.965	0.000	0.00	0/0
energy	0%	15	96.855	0.000	0.00	0/0
energy	5%	15	97.615	0.760	2.54	2/6
energy	10%	15	92.625	-4.230	-2.80	2/3
energy	25%	15	92.873	-3.982	-2.51	3/4
energy	100%	15	93.331	-3.524	-2.03	2/4

Table 8 Individual reference test RMSE after validation selection. Columns 5–100 indicate non-anchor reference audit percentages; zero uses proxy validation. Test oracle is descriptive and never used for selection

Stride	Offset	0	5	10	25	100	Oracle
<i>Source: air</i>							
3	0	0.722	0.722	0.691	0.722	0.722	0.671
3	1	0.713	0.713	0.713	0.713	0.713	0.671
3	2	0.683	0.683	0.683	0.683	0.683	0.665
12	0	0.973	0.973	0.973	0.973	0.973	0.973
12	1	0.972	0.972	0.972	0.972	0.972	0.880
12	2	0.960	0.960	0.960	0.960	0.960	0.830
12	3	0.884	0.884	0.884	0.884	0.884	0.884
12	4	1.036	1.036	1.036	1.036	1.036	0.999
24	0	1.097	1.097	1.097	1.097	1.097	1.097
24	1	1.105	1.105	1.105	1.105	1.105	1.105
24	2	1.127	1.127	1.127	1.127	1.127	1.127
24	3	1.141	1.141	1.141	1.141	1.141	1.141
24	4	1.131	1.131	1.131	1.131	1.131	1.131
<i>Source: energy</i>							
6	0	89.435	141.939	89.435	89.435	89.435	86.785
6	1	90.997	90.997	90.997	90.997	90.997	86.888
6	2	86.533	86.533	86.533	86.533	86.533	86.533
6	3	92.277	92.277	92.277	92.277	92.277	87.471
6	4	88.934	94.988	94.988	94.988	94.988	87.977
36	0	92.458	98.276	98.276	98.276	98.276	88.222
36	1	91.318	101.904	91.318	101.904	101.904	88.065
36	2	93.997	93.997	93.997	87.132	93.997	87.132
36	3	96.910	96.910	96.910	96.910	96.910	87.040
36	4	95.561	95.561	95.561	95.561	95.561	86.373
72	0	88.796	89.299	88.796	88.796	88.796	88.796
72	1	121.182	97.775	86.516	86.516	86.516	86.516
72	2	142.467	96.651	96.651	96.651	96.651	91.882
72	3	90.915	90.915	90.915	90.915	90.915	89.978
72	4	91.043	96.207	96.207	96.207	96.207	89.372

E All synthetic regimes

Each row aggregates 20 independently seeded cohorts, with six correlated candidate pairs per

cohort. Top differs counts seed-wise winner disagreement. Amplitude/noise are generator settings, not estimated clinical parameters.

Table 9 All 16 regimes with both training response definitions

Amp.	Noise	Phase	Stride	Training	Reversals	Top Pairs	differs
0	0.2	aligned	60	interpolated	0	120	0
0	0.2	aligned	60	reference	0	120	0
0	1	aligned	60	interpolated	0	120	0
0	1	aligned	60	reference	0	120	0
0	3	aligned	60	interpolated	0	120	0
0	3	aligned	60	reference	0	120	0
2	0.2	aligned	60	interpolated	106	120	20
2	0.2	aligned	60	reference	87	120	20
2	1	aligned	60	interpolated	96	120	20
2	1	aligned	60	reference	67	120	20
2	3	aligned	60	interpolated	34	120	20
2	3	aligned	60	reference	53	120	20
5	0.2	aligned	60	interpolated	104	120	20
5	0.2	aligned	60	reference	97	120	20
5	1	aligned	30	interpolated	104	120	20
5	1	aligned	30	reference	69	120	20
5	1	aligned	40	interpolated	79	120	20
5	1	aligned	40	reference	70	120	20
5	1	aligned	60	interpolated	104	120	20
5	1	aligned	60	reference	69	120	20
5	1	random	30	interpolated	62	120	20
5	1	random	30	reference	74	120	20
5	1	random	40	interpolated	71	120	20
5	1	random	40	reference	78	120	20
5	3	aligned	60	interpolated	105	120	20
5	3	aligned	60	reference	80	120	20
10	0.2	aligned	60	interpolated	110	120	20
10	0.2	aligned	60	reference	96	120	20
10	1	aligned	60	interpolated	110	120	20
10	1	aligned	60	reference	89	120	20
10	3	aligned	60	interpolated	110	120	20
10	3	aligned	60	reference	66	120	20

E.1 Representative-regime native errors

Amplitude five/noise one/aligned phase/stride 60 is fixed for this display. Values average all 20 seeds; standard deviations reflect cohort generation, not clinical uncertainty. Verification refits this regime for the first seed only.

Table 10 Representative-regime mean and standard deviation

Training	Candidate	Reference RMSE	Constructed RMSE
interpolated	calendar_ridge	3.640 ± 0.159	0.000 ± 0.000
interpolated	full_boost	3.641 ± 0.161	0.299 ± 0.117
interpolated	full_ridge	3.640 ± 0.159	0.000 ± 0.000
interpolated	sensor_ridge	2.172 ± 0.236	3.146 ± 0.141
reference	calendar_ridge	3.468 ± 0.147	1.154 ± 0.107
reference	full_boost	0.955 ± 0.013	3.503 ± 0.129
reference	full_ridge	0.958 ± 0.009	3.520 ± 0.139
reference	sensor_ridge	0.983 ± 0.010	3.689 ± 0.150

F Provenance, permissions and environment

Source archive URL, dataset DOI, retrieval date, archive SHA-256 and analyzed input-file SHA-256 are saved in `data/source_manifest.json`. Parkinson and Energy pages display CC BY 4.0. Air displays that license but its descriptive text also specifies research/noncommercial use. Complete source input files are not redistributed here; saved evaluation tables contain source-derived responses for research checking. Acquisition requires a research-use flag for Air and checks archive/extracted hashes. Broader use requires resolving the permission ambiguity.

The manuscript uses the official Springer Nature journal authoring template, with its original class and bibliography-style license notices retained. Experiment code was written for this audit. Publications associated with the datasets and relevant methods are cited.

Execution used Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0 and scikit-learn 1.8.0. Seed is 20261014; clinical fitting adds outer-fold index. Forests use two jobs; BLAS/OpenMP variables are set to one in the shared script. Other defaults are those of the pinned library. Selected parameters and actual fitted iteration counts are saved to inspect automatic boosting early stopping and finite neural budgets. No claim of hardware-independent bitwise identity is made.

After acquisition, run clinical fitting, both sensor cases, and simulations. Then run analysis and numerical verification. Editable manuscript tables are provided in the separate manuscript source package. Saved results permit numerical/table regeneration without all fits, after reacquisition for checks that require raw inputs. Source data preparation removes missing Air responses rather than imputing them. These steps regenerate the computational record; responsibility for the design, evidence, novelty, code and citations rests with the author.

G Fixed-case native-unit errors

The following illustration retains every candidate in the largest-spacing, offset-zero cases specified in the main text. It is selected by schedule rather

than maximum observed reversal. Both plotted losses use the same fixed prediction vectors.

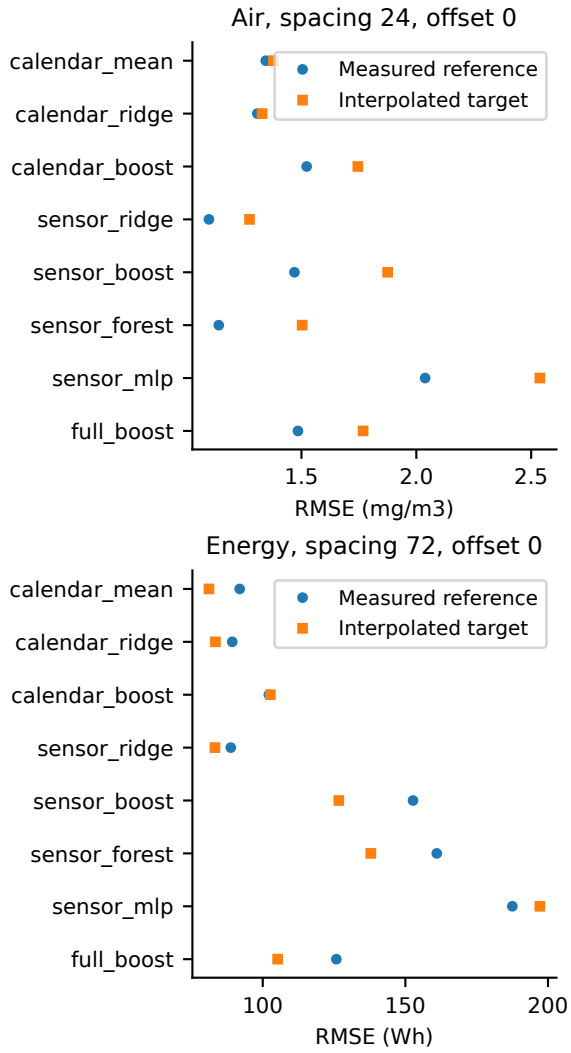


Fig. 7 Two test losses for the same interpolated-trained predictions in fixed cases: Air spacing 24/offset zero and Energy spacing 72/offset zero. Different magnitudes and orderings arise from response definition. Candidate names identify feature set and family, not test-selected hyperparameters

References

- [1] Candanedo L (2017) Appliances energy prediction. UCI Machine Learning Repository, <https://doi.org/10.24432/C5VC8G>, URL <https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction>
- [2] Candanedo LM, Feldheim V, Deramaix D (2017) Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings* 140:81–97. <https://doi.org/10.1016/j.enbuild.2017.01.083>
- [3] D’Amour A, Heller K, Moldovan D, et al (2022) Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23(226):1–61
- [4] Gebru T, Morgenstern J, Vecchione B, et al (2021) Datasheets for datasets. *Communications of the ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>
- [5] Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? In: *Advances in Neural Information Processing Systems*, pp 507–520
- [6] Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. *Nature* 585:357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- [7] Hwang SH, Whang SE (2022) Regmix: Data mixing augmentation for regression. *ArXiv preprint 2106.03374*, version 4
- [8] Jacobs AZ, Wallach H (2021) Measurement and fairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp 375–385, <https://doi.org/10.1145/3442188.3445901>
- [9] Northcutt CG, Athalye A, Mueller J (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. In: *Advances in Neural Information Processing Systems: Datasets and Benchmarks*, URL <https://arxiv.org/abs/2103.14749>
- [10] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830
- [11] Tang Z, Hou C, Zhang T, et al (2025) Enhancing noise robustness of Parkinson’s disease telemonitoring via contrastive feature augmentation. URL <https://arxiv.org/abs/2510.01588>, arXiv preprint 2510.01588
- [12] Tsanas A, Little M (2009) Parkinsons telemonitoring. UCI Machine Learning Repository, <https://doi.org/10.24432/C5ZS3N>, URL <https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring>
- [13] Tsanas A, Little MA, McSharry PE, et al (2010) Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering* 57(4):884–893. <https://doi.org/10.1109/TBME.2009.2036000>
- [14] Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- [15] Vito SD (2008) Air quality. UCI Machine Learning Repository, <https://doi.org/10.24432/C59K5F>, URL <https://archive.ics.uci.edu/dataset/360/air+quality>
- [16] Vito SD, Massera E, Piga M, et al (2008) On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* 129(2):750–757. <https://doi.org/10.1016/j.snb.2007.09.060>
- [17] Yao H, Wang Y, Zhang L, et al (2022) C-mixup: Improving generalization in regression. In: *Advances in Neural Information Processing Systems*
- [18] Zhou H, Mueller J, Kumar M, et al (2024) Detecting errors in a numerical response via any regression model. *Data-centric Machine Learning Research* 1(6):1–25