Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early Postoperative Trajectories for Six-Month Outcome Prediction after Endoscopic Sinus Surgery: A Leakage-Safe Comparison of Transformer and Conventional Machine-Learning Models

Jia Liu

liujia_tsguas@163.com

Tianjin Sino-German University of Applied Sciences

Shibo Jin

Tianjin Sino-German University of Applied Sciences

Zhenyu Deng

Tianjin Sino-German University of Applied Sciences

Xueliang Zhang

Tianjin Sino-German University of Applied Sciences

Research Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11604497/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Early Postoperative Trajectories for Six-Month Outcome Prediction after Endoscopic Sinus Surgery: A Leakage-Safe Comparison of Transformer and Conventional Machine-Learning Models

Jia Liu[1*], Shibo Jin[1], Zhenyu Deng[1], Xueliang Zhang[1]

[1]School of Software and Communication, Tianjin Sino-German University of Applied Sciences, Tianjin, 300350, China.

*Corresponding author(s). E-mail(s): liujia_tsguas@163.com;
Contributing authors: 15122148909@126.com; deng_zhen_yu@163.com;
13102840552@163.com;

Abstract

Early recovery after endoscopic sinus surgery may contain prognostic information that is not available preoperatively. We investigated whether trajectories measured through one postoperative month predict six-month outcomes in chronic rhinosinusitis and whether a trajectory-specific Transformer outperforms conventional models. This secondary analysis used 95 participants from a prospective randomized study (49 control; 46 biodegradable steroid-eluting stent). Predictors comprised baseline covariates and endoscopy, five-item visual analogue scale (VAS), and 22-item Sinonasal Outcome Test (SNOT-22) measurements at baseline, two weeks, and one month. Six-month continuous outcomes were predicted using repeated nested cross-validation (10 repetitions of five folds). For six-month SNOT-22 prediction from one-month data, Extra Trees achieved a mean absolute error (MAE) of 2.044 (95% CI 1.934–2.156), compared with 2.055 (1.953–2.149) for a clinically restricted ridge model and 2.624 (2.482–2.777) for the proposed EPT-Former. EPT-Former was significantly worse than the horizon-specific best comparator at every horizon (Holm-adjusted $\boldsymbol{p < 0.001}$). Adding early postoperative observations improved all principal models, but tree and regularized linear models retained an advantage. Early trajectories were informative, whereas increased architectural complexity did not improve prediction in this small structured cohort. External validation is required before clinical use.

Keywords: chronic rhinosinusitis, endoscopic sinus surgery, SNOT-22, longitudinal data, Transformer, machine learning

Acknowledgments. The authors thank the patients and clinical staff who contributed to the parent study.

1 Introduction

Chronic rhinosinusitis (CRS) is a heterogeneous inflammatory disease characterized by persistent sinonasal symptoms and objective evidence of mucosal disease. It substantially affects health-related quality of life, and endoscopic sinus surgery (ESS) is commonly considered when guideline-directed medical therapy does not provide adequate control [1, 2]. Although average symptom and endoscopic outcomes improve after ESS, individual recovery varies. Reliable early identification of patients likely to have an unfavorable longer-term course could support closer follow-up, timely adjustment of topical treatment, and better-informed shared decision-making.

Most outcome prediction in CRS has focused on preoperative variables. Baseline SNOT-22, computed tomography, endoscopy, comorbidity, and inflammatory markers have shown prognostic value, but their predictive effects are inconsistent across cohorts [3–5]. Recent machine-learning studies have predicted postoperative disease-control categories or clinically meaningful improvement from preoperative clinical and biomarker data [6–8]. These studies establish the feasibility of data-driven prediction, yet they do not directly address a different clinical question: once surgery has been performed, how much prognostic information is added by the patient's early postoperative trajectory?

Measurements at two weeks and one month form a short, multimodal sequence. They combine patient-reported symptom burden, endoscopic findings, and postoperative events, and may therefore reveal the direction and rate of recovery. A Transformer can represent interactions among time, modality, and change-from-baseline tokens [9, 10]. However, deep learning is not uniformly superior for small tabular datasets; tree ensembles and regularized regression often provide stronger inductive bias and lower variance [11, 12]. A credible clinical study must therefore compare a trajectory-specific neural model against strong conventional baselines under a leakage-safe validation protocol rather than assuming that a newer architecture will be better.

We performed a secondary prediction analysis of a prospective randomized study of 95 patients undergoing ESS, with or without a biodegradable mometasone-eluting sinus stent. The primary aim was to predict six-month SNOT-22 as a continuous outcome using information available preoperatively, at two weeks, or at one month. Secondary outcomes were the six-month five-item VAS composite and Lund–Kennedy endoscopy score. We designed an Early Postoperative Trajectory Transformer (EPT-Former), compared it with linear, kernel, tree-ensemble, boosting, multilayer perceptron, LSTM, and tabular Transformer baselines, and evaluated calibration, robustness, ablations, subgroups, learning curves, and feature-group occlusion. Reporting follows TRIPOD+AI principles [13].

The principal finding is intentionally conservative: early postoperative trajectories materially improved prediction, but EPT-Former did not outperform simpler approaches. This result delineates the value of the data from the value of architectural complexity and provides a reproducible benchmark for future multicenter validation.

2 Materials and Methods

2.1 Study Design and Reporting Framework

This was a prespecified secondary machine-learning analysis of a single-center, prospective, randomized controlled study conducted in the Department of Otorhinolaryngology–Head and Neck Surgery at the Third Affiliated Hospital of Nanjing Medical University (Changzhou No. 2 People's Hospital), China, between September 2023 and July 2025. The parent study evaluated a biodegradable drug-eluting sinus stent placed during ESS. Of 103 randomized participants, eight were lost to follow-up (three in the stent group and five in the control group), leaving 95 participants with complete six-month data (completion rate, 92.23%). The prediction study used only these 95 complete cases.

An independent researcher generated the parent-study allocation sequence in Microsoft Excel using random numbers between 0 and 1 (< 0.5, control; ≥ 0.5, stent). Assignments were placed in opaque sealed envelopes and opened by the circulating nurse on the day of surgery. Post-allocation blinding details of the parent trial were not available for this secondary analysis. The parent-study sample size was based on a 10-per-group pilot, two-sided $\alpha = 0.05$, 80% power, and the largest estimate across six-month endoscopy, VAS, and SNOT-22 outcomes (42 participants per group). No separate machine-learning sample-size calculation was performed because the present work was a secondary analysis of the completed trial.

The protocol was approved by the institutional ethics committee of Changzhou No. 2 People's Hospital (approval code YLJSA149), and all participants provided written informed consent. A prospective clinical-trial registration number was not available to the authors of this secondary analysis at the time of writing. The analysis was structured around the TRIPOD+AI reporting framework and considered risks highlighted by PROBAST, including overfitting, data leakage, and lack of external validation [13, 14].

2.2 Participants

Eligible participants were 18–75 years old, met the 2024 Chinese diagnostic guideline for CRS, had bilateral pansinusitis on computed tomography, had persistent disease despite at least 12 weeks of medical treatment, and had an indication for bilateral functional ESS [2]. Major exclusions were acute sinonasal infection; severe septal deviation requiring concurrent correction; recurrent or fungal sinusitis; severe systemic disease precluding general anesthesia; allergy or contraindication to stent materials or mometasone; long-term oral corticosteroid or immunosuppressive treatment; immunodeficiency or autoimmune disease; pregnancy; psychiatric illness or anticipated poor adherence; and incomplete clinical data.

Allergy was determined by skin-prick testing or serum-specific IgE. Eosinophilic disease was defined histologically as at least 10 eosinophils per high-power field at 400× magnification. Demographic and clinical covariates included age, sex, disease duration, treatment assignment, asthma, allergy, eosinophilic phenotype, and the preoperative Lund–Mackay computed-tomography score.

2.3 Surgery and Postoperative Treatment

All patients received prednisone 20 mg/day for seven days before surgery and underwent standardized bilateral ESS by the same senior surgeon using a Messerklinger approach. Diseased tissue was removed while macroscopically normal mucosa was preserved; all involved sinuses were opened, and nasal packing was removed after 48 h. The stent group received a bilateral biodegradable "Xiangtong" implant (Puyi [Shanghai] Biotechnology, Shanghai, China), containing 652 μg mometasone furoate and designed to release approximately 20 μg/day while degrading over approximately 30 days. The control group received mometasone furoate nasal spray for one month after packing removal. Both groups received a third-generation cephalosporin for seven days, twice-daily saline irrigation, endoscopic cleaning at one week, and follow-up at two weeks and one, three, and six months. Exact catalogue identifiers of the implant were not re-verified for this secondary analysis; the device manufacturer had no role in the design of the present prediction analysis, data analysis, interpretation, writing, or the decision to submit the manuscript.

2.4 Predictors and Outcomes

Three information horizons were defined (Figure 1). The preoperative horizon used 15 variables: eight static covariates, the Lund–Kennedy endoscopy score, five symptom-specific VAS items, and total SNOT-22. The two-week horizon added endoscopy, five VAS items, SNOT-22, and nasal dryness (23 variables in total). The one-month horizon additionally included endoscopy, five VAS items, SNOT-22, vesicle formation, and adhesion (32 variables in total). No three-month or six-month variable was permitted in a predictor set.

The primary outcome was total SNOT-22 at six months (range, 0–110), a validated 22-item disease-specific quality-of-life measure [15]. Secondary continuous outcomes were (i) the six-month sum of five 0–10 VAS items—nasal obstruction, rhinorrhea, hyposmia, facial pain/pressure, and overall symptoms—with a possible range of 0–50, and (ii) the bilateral Lund–Kennedy endoscopy score (range, 0–12) [16]. The Lund–Mackay baseline score ranged from 0 to 24 [17].

2.5 Data Quality Control and Preprocessing

The source workbook contained 95 rows and 53 columns. Patient identifiers were unique; categorical variables used valid binary codes; and continuous scores were within their prespecified ranges. There were no missing values among the analyzed predictors and outcomes. The stored VAS totals were audited against the sum of their five components. Six preoperative stored totals were inconsistent with their components and were replaced by the reproducible component sum; no other stored total affected the analysis. Source-document confirmation of these six corrections against original case-report forms is recommended before any clinical translation of the models.

For conventional pipelines, median imputation and, for scale-sensitive models, standardization were fitted within the relevant training fold. Although the verified dataset was complete, the imputer was retained to make the pipeline robust to future

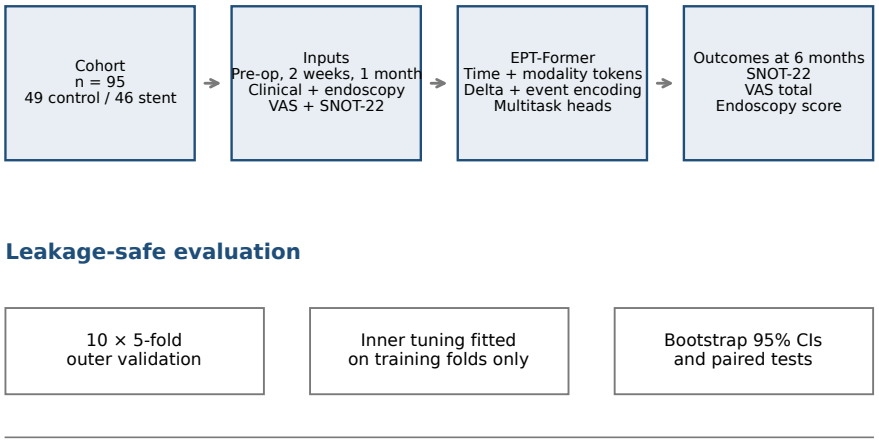


Fig. 1 Study design and leakage-safe modeling workflow. Static covariates and longitudinal endoscopy, five-item VAS, SNOT-22, and postoperative-event measurements were made available sequentially at the preoperative, two-week, and one-month prediction horizons. All model fitting, preprocessing, and hyperparameter selection were restricted to each outer training fold. EPT-Former, Early Postoperative Trajectory Transformer; VAS, visual analogue scale.

missing values. For neural models, continuous and binary predictors were standardized using means and standard deviations estimated from the fitting subset only. Outcomes were standardized during neural-network training and transformed back to their original scales for evaluation.

2.6 EPT-Former

EPT-Former was designed for a short clinical trajectory rather than a long, densely sampled sequence. Let $\mathbf{s}_i$ denote the static feature vector for patient i, and let $\mathbf{x}_{i,t}^{(m)}$ denote measurements at time $t \in \{0, 1, 2\}$ for modality $m \in \{\text{endo}, \text{VAS}, \text{SNOT}\}$. Numeric tokenizers mapped each input block to a d-dimensional embedding:

$$\mathbf{z}_{i,t}^{(m)} = \text{LN}\Big(\text{GELU}\Big(\mathbf{W}_m \mathbf{x}_{i,t}^{(m)} + \mathbf{b}_m\Big)\Big). \tag{1}$$

For postoperative observations, a separately parameterized delta token encoded change from baseline, $\Delta\mathbf{x}_{i,t}^{(m)} = \mathbf{x}_{i,t}^{(m)} - \mathbf{x}_{i,0}^{(m)}$, and was added to the corresponding value token. Learned time and modality embeddings were then added. A static token, postoperative-event tokens, and a learned classification token were concatenated with the modality tokens and passed through two pre-normalized Transformer encoder layers. Each layer used two attention heads, hidden dimension $d = 32$, feed-forward dimension 64, GELU activation, and dropout 0.20. The classification-token representation was layer-normalized and passed to three linear regression heads for six-month SNOT-22, VAS composite, and endoscopy score.

The multitask objective was the unweighted mean of standardized smooth-L_1 losses:

$$\mathcal{L} = \frac{1}{3} \sum_{k \in \{\text{SNOT,VAS,Endoscopy}\}} \text{SmoothL1}\left(\hat{y}_{i,k}^{(z)}, y_{i,k}^{(z)}\right). \tag{2}$$

Models were optimized with AdamW (learning rate 10^{-3}; weight decay 10^{-4}), gradient-norm clipping at 2.0, and a maximum of 400 full-batch epochs. Within each outer training set, a random 20% subset (minimum 10 patients) was reserved for early stopping on primary-outcome MAE with patience 50. Hyperparameters were fixed before the repeated evaluation rather than selected on outer test data.

2.7 Comparators

Comparators covered clinically interpretable and flexible model families: training-set mean; last-observation persistence; a clinically restricted ridge model; ordinary linear regression; ridge; lasso; elastic net; radial-basis-function support-vector regression; random forest; extremely randomized trees (Extra Trees); histogram gradient boosting; XGBoost; LightGBM; CatBoost; multilayer perceptron; a trajectory LSTM; and FT-Transformer [10, 18–25]. The clinically restricted ridge model used treatment group and the available historical measurements from the same outcome domain, with a fixed ridge penalty of 10. Conventional model hyperparameters were selected by four-fold inner cross-validation minimizing MAE. Neural comparator architecture and optimization budgets matched EPT-Former where applicable. TabPFN was not included among the final comparators.

2.8 Validation and Statistical Analysis

Generalization was estimated by 10 repetitions of five-fold outer cross-validation (50 outer test folds). Compound strata combined treatment assignment and ranked outcome halves; folds therefore approximately preserved both treatment allocation and outcome distribution. All preprocessing and conventional hyperparameter searches occurred within outer training data. This repeated nested design was selected because a single train–test split would be unstable in a 95-patient cohort [26].

The primary metric was MAE. Secondary metrics were root mean squared error (RMSE), coefficient of determination (R^2), Pearson and Spearman correlations, and calibration slope/intercept from a linear regression of observed on predicted values. Metrics were calculated separately in each outer fold and summarized by their mean. Ninety-five percent confidence intervals were generated by 2000 bootstrap resamples of the 50 fold-level estimates. EPT-Former was compared with the lowest-MAE comparator at each horizon by a paired split-level bootstrap; the three horizon-wise p values were adjusted using Holm's method. Values printed as zero by finite bootstrap sampling are reported as $p < 0.001$. Calibration and error plots averaged each patient's out-of-fold prediction across the 10 repetitions. Calibration is essential because strong correlation alone does not ensure numerically reliable predictions [27].

Secondary analyses included feature-group occlusion, five prespecified subgroups, test-time corruption, learning curves, ablation, and a label-permutation test. Occlusion replaced a feature family with its training-fold standardized mean and reported

the increase in MAE. Subgroups were treatment, sex, eosinophilic phenotype, allergy, and asthma; these estimates were descriptive. Robustness and learning-curve experiments used one five-fold outer repetition. Missingness corruption randomly replaced trajectory values by the standardized mean; Gaussian noise was expressed in standard-deviation units. A 500-permutation two-sided test compared the observed patient-level Pearson correlation with a null distribution created by permuting predictions. All analyses used a fixed seed (20260804).

2.9 Software and Reproducibility

The experiments used Python, scikit-learn [28], PyTorch 2.8, XGBoost, LightGBM, and CatBoost on a CUDA-enabled server. The complete run produced 153,900 patient-level outer-fold prediction rows. Split manifests, data-cleaning audits, hyperparameter records, source data for figures, and vector figures were retained. Exact package and hardware versions used for the final run are recorded in the analysis package and are available from the corresponding author upon reasonable request.

During manuscript preparation, generative AI tools were used for language editing and organization assistance; the authors reviewed and edited all content and take full responsibility for the scientific statements. All analyses and numerical results were produced from the authors' verified code and data.

3 Results

3.1 Cohort Characteristics

The analyzed cohort included 49 control and 46 stent-treated patients. Baseline variables were broadly similar between groups (Table 1). The mean age was 45.8 years overall, 71 of 95 patients were male, and 45 had an eosinophilic phenotype. No analyzed predictor or six-month outcome was missing.

Table 1 Baseline characteristics of the analyzed cohort.

Characteristic	**Control** ($n = 49$)	**Stent** ($n = 46$)	p **Value**
Age, years	43.98 ± 15.12	47.65 ± 14.61	0.232
Disease duration, years	3.19 ± 7.61	3.64 ± 4.78	0.731
Lund–Mackay CT score	16.98 ± 4.51	16.74 ± 4.48	0.795
Male sex, n (%)	35 (71.4)	36 (78.3)	0.444
Eosinophilic phenotype, n (%)	22 (44.9)	23 (50.0)	0.619
Allergy, n (%)	17 (34.7)	17 (37.0)	0.819
Asthma, n (%)	4 (8.2)	6 (13.0)	0.439
Preoperative endoscopy score	8.51 ± 2.13	9.13 ± 2.27	0.174
Preoperative SNOT-22	44.88 ± 18.51	41.28 ± 16.97	0.326

Values are mean $\pm$ standard deviation unless otherwise stated. CT, computed tomography; SNOT-22, 22-item Sinonasal Outcome Test. Continuous covariates are summarized as mean $\pm$ standard deviation as reported in the analysis tables; skewed variables may alternatively be summarized with medians in subsequent clinical reports.

3.2 Main Prediction Performance

At the one-month horizon, Extra Trees had the lowest mean SNOT-22 MAE (2.044; 95% CI 1.934–2.156), closely followed by the clinically restricted ridge model (2.055; 1.953–2.149) and lasso (2.122; 2.020–2.213) (Table 2). Extra Trees explained 66.6% of fold-level outcome variance and had a calibration slope of 1.047. EPT-Former achieved MAE 2.624 (2.482–2.777), RMSE 3.282 (3.109–3.468), $R^2 = 0.523$ (0.473–0.565), and Pearson $r = 0.764$. FT-Transformer performed less well (MAE 2.912), whereas the trajectory LSTM was intermediate (2.533). Thus, explicit trajectory architectures did not displace the strongest regularized and tree baselines.

Table 2 Prediction of six-month SNOT-22 from information available at one month.

Model	MAE (95% CI)	RMSE	R^2	Pearson r	Calibration Slope
Mean predictor	3.860 (3.685–4.026)	4.915	−0.030	—	—
Clinical-Ridge	2.055 (1.953–2.149)	2.646	0.689	0.853	1.056
Lasso	2.122 (2.020–2.213)	2.730	0.670	0.846	1.154
Extra Trees	**2.044 (1.934–2.156)**	2.752	0.666	0.842	1.047
XGBoost	2.145 (2.045–2.241)	2.870	0.632	0.820	0.977
LightGBM	2.229 (2.124–2.332)	2.982	0.599	0.807	0.956
CatBoost	2.137 (2.043–2.230)	2.786	0.655	0.835	1.071
Trajectory-LSTM	2.533 (2.373–2.693)	3.215	0.527	0.762	1.162
FT-Transformer	2.912 (2.763–3.068)	3.758	0.368	0.707	0.987
EPT-Former	2.624 (2.482–2.777)	3.282	0.523	0.764	1.181

Estimates are means across 50 outer folds. Confidence intervals are 2000-resample bootstrap intervals across fold-level estimates. MAE, mean absolute error; RMSE, root mean squared error.

The same pattern was observed for secondary outcomes (Figure 2; Table 3). At one month, Extra Trees achieved MAEs of 1.935 for the six-month five-item VAS composite and 0.683 for endoscopy, compared with 2.245 and 0.872 for EPT-Former. The clinically restricted ridge model remained competitive for all three outcomes.

Table 3 One-month-horizon MAE across the three six-month outcomes.

Model	SNOT-22	VAS Composite	Endoscopy
Clinical-Ridge	2.055	1.949	0.688
Extra Trees	**2.044**	**1.935**	**0.683**
EPT-Former	2.624	2.245	0.872

Values are mean MAE across 50 outer folds. Outcome scales differ and values should not be compared vertically between columns.

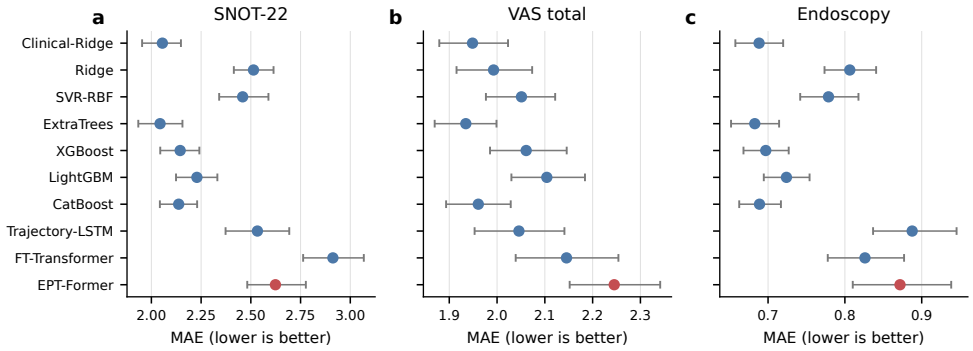


Fig. 2 One-month-horizon model comparison for the three six-month outcomes. Points show mean outer-fold MAE and horizontal bars show 95% bootstrap confidence intervals. EPT-Former is highlighted in red. Lower values indicate better prediction.

3.3 Calibration, Error Pattern, and Paired Comparisons

When each patient's 10 repeated out-of-fold EPT-Former predictions were averaged, the patient-level Pearson correlation with observed six-month SNOT-22 was 0.816. Calibration and residual plots showed useful rank ordering but non-negligible individual error and some dispersion at higher outcome values (Figure 3). The fold-level calibration slope of 1.181 indicates that EPT-Former predictions were somewhat under-dispersed on average.

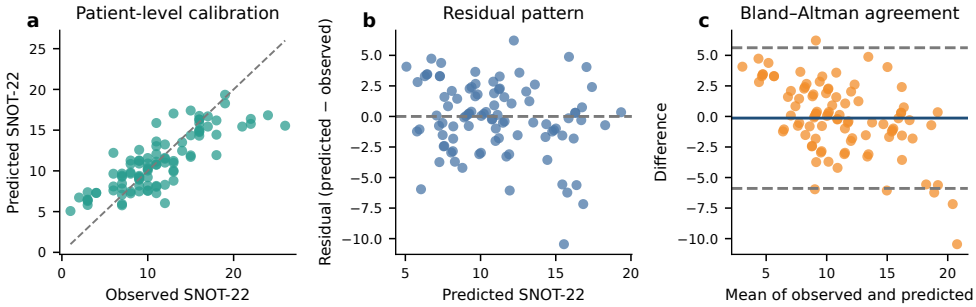


Fig. 3 Patient-level EPT-Former prediction diagnostics for six-month SNOT-22 at the one-month horizon. Predictions were averaged over 10 repeated out-of-fold estimates per patient. (A) Observed versus predicted values with the identity line; (B) residuals (predicted minus observed); and (C) Bland–Altman agreement with mean bias and 95% limits of agreement [29].

EPT-Former had significantly higher MAE than the best comparator selected at each horizon (Table 4). The mean excess MAE was 0.498 preoperatively, 0.539 at two weeks, and 0.580 at one month; all Holm-adjusted p values were < 0.001.

Table 4 Paired comparison of EPT-Former with the horizon-specific lowest-MAE comparator for six-month SNOT-22.

Horizon	Reference	MAE Difference	95% CI	Holm-adjusted p
Preoperative	LightGBM	0.498	0.346–0.641	< 0.001
Two weeks	CatBoost	0.539	0.400–0.680	< 0.001
One month	Extra Trees	0.580	0.444–0.719	< 0.001

Positive differences favor the reference model. Pairing was by repeated outer fold.

3.4 Information Horizons and Ablation

Later information horizons generally improved SNOT-22 prediction (Figure 4). For Extra Trees, MAE decreased from 2.423 preoperatively to 2.230 at two weeks and 2.044 at one month. Clinical-Ridge decreased from 2.430 to 2.320 and 2.055, while EPT-Former decreased from 2.902 to 2.741 and 2.624. These trends indicate that early postoperative measurements added predictive information regardless of architecture.

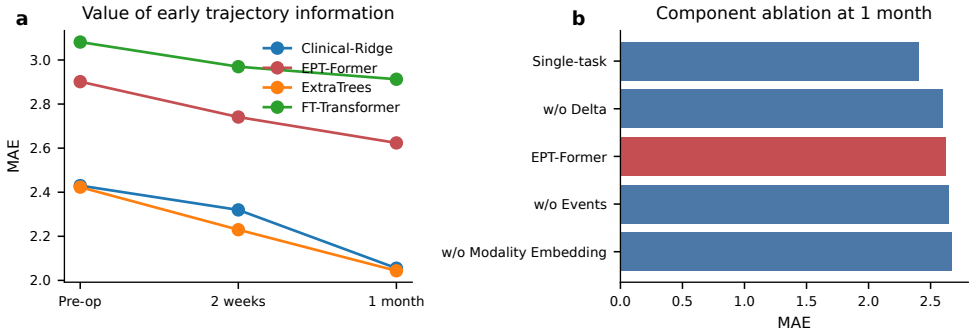


Fig. 4 Information-horizon and ablation analyses for six-month SNOT-22. (A) Mean MAE using predictors available preoperatively, at two weeks, and at one month. (B) One-month EPT-Former component ablations. Lower MAE is better.

The single-task EPT-Former achieved the best one-month result among EPT variants (MAE 2.404 versus 2.624 for multitask EPT-Former), suggesting negative transfer from the auxiliary outcomes in this sample (Table 5). Removing delta encoding slightly improved one-month MAE to 2.600 but worsened two-week MAE to 2.834. Removing events or modality embeddings caused small and inconsistent changes. Therefore, the ablation study did not support a claim that every proposed component independently improved performance.

Table 5 EPT-Former ablation results for six-month SNOT-22.

Variant	Two-week MAE	Two-week R^2	One-month MAE	One-month R^2
EPT-Former	2.741	0.482	2.624	0.523
Single-task	**2.685**	0.495	**2.404**	0.600
Without delta tokens	2.834	0.463	2.600	0.548
Without event tokens	2.724	0.498	2.645	0.518
Without modality embeddings	2.732	**0.501**	2.670	0.520

Values are means across 50 outer folds; full bootstrap intervals are retained in the analysis package.

3.5 Robustness, Sample Size, and Subgroups

Under one repeated five-fold analysis, EPT-Former MAE increased from 2.834 without corruption to 3.285 when 30% of trajectory values were replaced by their training-fold means and to 3.419 when Gaussian noise with standard deviation 1.0 was added (Table 6; Figure 5A). The learning curve favored Extra Trees at every evaluated training fraction (Figure 5B). With the complete fold training sample, Extra Trees achieved mean MAE 1.966 versus 2.802 for EPT-Former in this one-repeat secondary analysis. Both models improved with additional data, but the neural model showed the larger remaining data requirement.

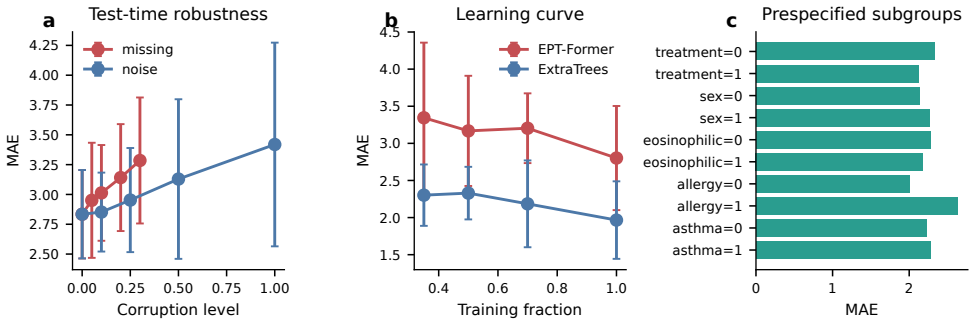


Fig. 5 Secondary generalization analyses for one-month prediction of six-month SNOT-22. (A) EPT-Former test-time corruption; error bars show standard deviations across five folds. Missingness denotes random replacement with the training-fold standardized mean, and noise is measured in standard-deviation units. (B) Learning curves for Extra Trees and EPT-Former. (C) Descriptive EPT-Former MAE in prespecified subgroups; binary level definitions follow the dataset coding (0, absent/control/female; 1, present/stent/male, as applicable). These one-repeat or small-subgroup analyses are exploratory.

EPT-Former subgroup MAEs ranged from 2.008 in patients without allergy to 2.634 in patients with allergy (Table 7). Treatment-group MAE was 2.333 in controls and 2.124 in stent recipients. The asthma-positive subgroup contained only 10 patients; none of the subgroup differences should be interpreted as evidence of algorithmic equivalence or disparity without a larger external cohort.

Table 6 EPT-Former test-time robustness at the one-month horizon.

Corruption	Level	MAE Mean ± SD	RMSE Mean ± SD	R^2 Mean ± SD
None	0	2.834 ± 0.371	3.600 ± 0.664	0.435 ± 0.085
Missing	0.05	2.951 ± 0.483	3.721 ± 0.717	0.401 ± 0.078
Missing	0.10	3.013 ± 0.401	3.813 ± 0.622	0.369 ± 0.033
Missing	0.20	3.141 ± 0.448	3.954 ± 0.536	0.315 ± 0.078
Missing	0.30	3.285 ± 0.528	4.034 ± 0.602	0.280 ± 0.149
Noise	0.10	2.852 ± 0.331	3.613 ± 0.664	0.435 ± 0.060
Noise	0.25	2.953 ± 0.437	3.693 ± 0.707	0.414 ± 0.036
Noise	0.50	3.129 ± 0.669	3.867 ± 0.829	0.357 ± 0.112
Noise	1.00	3.419 ± 0.854	4.156 ± 1.058	0.253 ± 0.226

Results are from one five-fold repetition and are not directly interchangeable with the 50-fold primary estimates.

Table 7 Descriptive patient-level EPT-Former subgroup performance at the one-month horizon.

Subgroup	Level	n	MAE	Pearson r
Treatment	Control	49	2.333	0.761
Treatment	Stent	46	2.124	0.683
Sex	Female	24	2.139	0.845
Sex	Male	71	2.263	0.810
Eosinophilic phenotype	Absent	50	2.278	0.798
Eosinophilic phenotype	Present	45	2.180	0.850
Allergy	Absent	61	2.008	0.825
Allergy	Present	34	2.634	0.823
Asthma	Absent	85	2.227	0.810
Asthma	Present	10	2.272	0.902

Patient-level predictions were averaged across 10 repeated out-of-fold estimates. No hypothesis tests were performed because subgroup analyses were underpowered.

3.6 Interpretability and Validity Checks

Occlusion of the SNOT-22 trajectory produced the largest mean EPT-Former MAE increase (0.996), followed by the VAS trajectory (0.691), one-month snapshot (0.506), and endoscopy trajectory (0.275) (Figure 6A). Occluding postoperative events had minimal effect (0.016), and occluding static covariates slightly reduced MAE (-0.060), consistent with their limited incremental value or overfitting in the neural model. Absolute error was highest in the upper observed-outcome tertile (mean 2.671 versus 2.169 and 1.764 in the low and middle tertiles; Figure 6B). The label-permutation test yielded an observed patient-level correlation of 0.816 and an empirical two-sided $p = 0.002$ with 500 permutations (Figure 6C), confirming that predictions contained signal beyond chance while not establishing transportability.

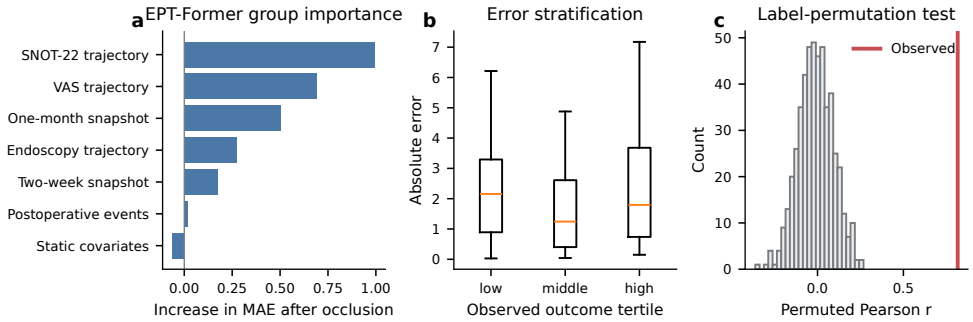


Fig. 6 EPT-Former interpretability and validity checks. (A) Mean increase in one-month-horizon MAE after foldwise feature-group occlusion; negative values indicate that occlusion improved performance. (B) patient-level absolute errors across observed SNOT-22 tertiles. (C) null distribution of Pearson correlation from 500 label permutations, with the observed correlation shown in red.

4 Discussion

4.1 Principal Findings

This study addressed two separable questions: whether early postoperative trajectories are useful, and whether a trajectory-specific Transformer is the best way to use them. The answer to the first question was yes. Across Clinical-Ridge, Extra Trees, FT-Transformer, and EPT-Former, prediction improved as two-week and one-month observations became available. The answer to the second question was no in this cohort. Extra Trees and a clinically restricted ridge model yielded the lowest one-month SNOT-22 errors, while EPT-Former was significantly worse than the strongest comparator at all three horizons.

The distinction matters. A model paper centered only on the proposed architecture might have obscured the clinically relevant finding that the trajectory itself was informative. The competitive performance of Clinical-Ridge also indicates that much of the signal was concentrated in treatment assignment and recent measurements of the same outcome domain. At one month, Extra Trees improved MAE only marginally over Clinical-Ridge (2.044 versus 2.055), a difference that is unlikely to justify greater implementation complexity without external evidence. The results support a simple-model-first strategy for small postoperative cohorts.

4.2 Relation to Previous Work

Previous CRS prediction studies have mainly used preoperative information to classify postoperative control, recurrence, or clinically important improvement. Gata et al. used random forests in 85 patients with CRSwNP and reported higher classification accuracy after adding microRNA markers [7]. Zhang et al. compared nine models in 89 followed patients and found that logistic regression combining patient-reported items and selected clinical features performed best [8]. Raghavan et al. studied 242 patients and reported strong discrimination for an ensemble predicting SNOT-22 minimal clinically important difference [5]. Lötsch et al. used machine-learning feature

selection to characterize changes associated with surgery rather than forecast a six-month continuous outcome [6]. Our analysis complements these studies by focusing on dynamic information observed after surgery and by evaluating continuous SNOT-22, VAS, and endoscopy targets.

The finding that tree and regularized models outperformed Transformers is consistent with broad tabular-learning benchmarks [11, 12]. With only 95 patients and 32 one-month predictors, flexible attention models can learn unstable interactions. The learning curve, negative effect of multitask training, and negligible contribution of several token groups all point in the same direction: the present data volume does not support the additional degrees of freedom. Multitask learning may also have introduced negative transfer because the three outcomes have different scales, noise characteristics, and clinical constructs despite standardization.

4.3 Clinical Interpretation

This analysis is a model-development study, not a clinical deployment study. Predictions should not be used to change postoperative care until they have been externally validated, recalibrated, and linked to a prespecified decision pathway. Nevertheless, the horizon analysis suggests a pragmatic future workflow: collect standardized SNOT-22, symptom-specific VAS, and endoscopy data at two weeks and one month; update the predicted six-month outcome; and flag patients with both high predicted burden and unfavorable observed recovery for clinical review. Such a system should display prediction uncertainty and recent trajectory rather than a single opaque score.

Treatment assignment was a legitimate predictor because it was known at every forecast horizon. However, its importance means that transport to hospitals using different implants, topical regimens, surgical techniques, or follow-up schedules cannot be assumed. The commercial implant itself was evaluated in the parent trial, and the current prediction model does not estimate a causal treatment effect. It forecasts outcomes conditional on observed treatment and clinical course. Any future decision-support system should distinguish these prognostic and causal questions.

4.4 Strengths and Limitations

Strengths include prospective standardized data collection, multiple clinically relevant outcomes, explicit separation of information horizons, strong conventional baselines, repeated nested validation, retained split manifests, fold-restricted preprocessing, paired comparisons, calibration assessment, ablation, robustness, learning curves, and permutation testing. The analysis did not suppress the negative architectural result, and all reported values derive from completed experiment files.

Several limitations are substantial. First, this is a single-center complete-case cohort of only 95 patients; no external validation was performed. Repeated cross-validation reduces dependence on a single split but cannot create independent patients, and bootstrap intervals over correlated repeated folds may be too narrow. Second, eight randomized participants were lost to follow-up, creating potential selection bias. Third, deep models used a random internal validation subset and fixed hyperparameters, whereas conventional models received inner-grid tuning; different tuning budgets

could affect relative performance, although the learning curve suggests that sample size remains the main constraint. Fourth, the robustness, occlusion, and learning-curve analyses used only one outer repetition and are exploratory. Fifth, subgroup sizes were small, especially for asthma, and no fairness conclusion is supported. Sixth, six preoperative VAS totals required correction from component scores and still require source-document confirmation. Seventh, the outcomes were observed six months after surgery; performance for recurrence, longer-term disease control, or treatment decisions was not evaluated. Finally, a prospective trial-registration number was unavailable for this secondary analysis, and public deposition of patient-level data remains subject to institutional ethics approval.

Future work should prospectively validate the simple Clinical-Ridge and Extra Trees models in temporally and geographically distinct cohorts before investing in a larger Transformer. A multicenter sample should include site-level validation, missing-data mechanisms, model recalibration, decision-curve analysis tied to a concrete clinical action, and comparison with clinician judgment. If the cohort becomes sufficiently large, pretraining or self-supervised trajectory representations may be reasonable; the present results do not justify that complexity yet.

5 Conclusions

Clinical measurements collected during the first postoperative month improved six-month outcome prediction after ESS for CRS. In a rigorous comparison on 95 patients, however, a purpose-built trajectory Transformer did not outperform Extra Trees, lasso, or a clinically restricted ridge model. The evidence therefore supports the prognostic value of early recovery trajectories, not the superiority of deep attention in small structured clinical datasets. Simple models should be prioritized for external validation, and none of the evaluated models is ready for clinical use.

Abbreviations

CRS	Chronic rhinosinusitis
EPT-Former	Early Postoperative Trajectory Transformer
ESS	Endoscopic sinus surgery
MAE	Mean absolute error
RMSE	Root mean squared error
SNOT-22	22-item Sinonasal Outcome Test
VAS	Visual analogue scale

Statements and Declarations

Funding

This research received no external funding.

Competing interests

The authors declare no conflicts of interest. The device manufacturer (Puyi [Shanghai] Biotechnology) had no role in the design of this secondary prediction analysis, data analysis, interpretation, manuscript writing, or the decision to publish.

Ethics approval

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Changzhou No. 2 People's Hospital (protocol code YLJSA149).

Consent to participate

Written informed consent was obtained from all subjects involved in the study.

Data availability

The analysis package contains split manifests, de-identified modeling tables, fold-level predictions, metrics, data-cleaning audits, and source data for all figures. De-identified analysis materials are available from the corresponding author upon reasonable request, subject to institutional ethics and privacy restrictions on patient-level clinical variables.

Author contributions

Conceptualization, J.L. and X.Z.; methodology, J.L., S.J., Z.D. and X.Z.; software, J.L., S.J. and Z.D.; validation, J.L. and S.J.; formal analysis, J.L., S.J. and Z.D.; investigation, J.L., S.J., Z.D. and X.Z.; resources, X.Z.; data curation, J.L. and S.J.; writing—original draft, J.L.; writing—review and editing, J.L., S.J., Z.D. and X.Z.; visualization, J.L. and Z.D.; supervision, X.Z.; project administration, J.L. and X.Z. All authors have read and agreed to the published version of the manuscript.

Appendix A Model Hyperparameters and Reproducibility Checklist

Table A1 lists the model and evaluation settings.

Table A1 Prespecified EPT-Former and evaluation settings.

Setting	Value
Outer validation	10 repetitions × 5 folds
Inner validation	4 folds for conventional model grids
Bootstrap iterations	2000
Permutation iterations	500
Random seed	20260804
Embedding dimension / heads / layers	32 / 2 / 2
Feed-forward dimension	64
Dropout	0.20
Optimizer	AdamW
Learning rate / weight decay	10^{-3} / 10^{-4}
Maximum epochs / patience	400 / 50
Loss	Standardized smooth-L_1 (mean across tasks)
Gradient clipping	Global norm 2.0

References

[1] Fokkens, W.J., Lund, V.J., Hopkins, C., *et al.*: European Position Paper on Rhinosinusitis and Nasal Polyps 2020. Rhinology **58**(Suppl. S29), 1–464 (2020) https://doi.org/10.4193/Rhin20.600

[2] Rhinology Group, Editorial Board of Chinese Journal of Otorhinolaryngology Head and Neck Surgery, Rhinology Group, Chinese Society of Otorhinolaryngology Head and Neck Surgery: Chinese Guideline for Diagnosis and Treatment of Chronic Rhinosinusitis (2024). Chin. J. Otorhinolaryngol. Head Neck Surg. **60**, 221–249 (2025). In Chinese

[3] Hopkins, C., Rudmik, L., Lund, V.J.: The Predictive Value of the Preoperative Sinonasal Outcome Test-22 Score in Patients Undergoing Endoscopic Sinus Surgery for Chronic Rhinosinusitis. Laryngoscope **125**, 1779–1784 (2015) https://doi.org/10.1002/lary.25318

[4] Brooks, S.G., Trope, M., Blasetti, M., *et al.*: Preoperative Lund–Mackay Computed Tomography Score Is Associated with Preoperative Symptom Severity and Predicts Quality-of-Life Outcome Trajectories after Sinus Surgery. Int. Forum Allergy Rhinol. **8**, 668–675 (2018) https://doi.org/10.1002/alr.22109

[5] Raghavan, A., Sage, E., Al-Ghezi, M., Aboueisha, M., Prohnitchi, I., Giliberto, J.P., Humphreys, I., Jafari, A., Abuzeid, W.M.: Predicting Surgical Outcomes in Chronic Rhinosinusitis from Preoperative Patient Data: A Machine Learning

Approach. Int. Forum Allergy Rhinol. **16**, 239–246 (2026) https://doi.org/10.1002/alr.70068

[6] Lötsch, J., Hintschich, C.A., Petridis, P., Pade, J., Hummel, T.: Machine-Learning Points at Endoscopic, Quality of Life, and Olfactory Parameters as Outcome Criteria for Endoscopic Paranasal Sinus Surgery in Chronic Rhinosinusitis. J. Clin. Med. **10**, 4245 (2021) https://doi.org/10.3390/jcm10184245

[7] Gata, A., Raduly, L., Budişan, L., Bajcsi, A., Ursu, T.-M., Chira, C., Dioşan, L., Berindan-Neagoe, I., Albu, S.: Machine Learning Model Predicts Postoperative Outcomes in Chronic Rhinosinusitis with Nasal Polyps. Clin. Otolaryngol. **49**, 776–784 (2024) https://doi.org/10.1111/coa.14208

[8] Zhang, C., *et al.*: Patient-Reported Outcome Measures as Predictive Tools for Disease Control in Chronic Rhinosinusitis with Nasal Polyps: A Prospective Study. Clin. Transl. Allergy **15**, 70119 (2025) https://doi.org/10.1002/clt2.70119

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Red Hook, NY, USA (2017)

[10] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting Deep Learning Models for Tabular Data. In: Advances in Neural Information Processing Systems 34, pp. 18932–18943. Curran Associates, Red Hook, NY, USA (2021)

[11] Shwartz-Ziv, R., Armon, A.: Tabular Data: Deep Learning Is Not All You Need. Inf. Fusion **81**, 84–90 (2022) https://doi.org/10.1016/j.inffus.2021.11.011

[12] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? In: Advances in Neural Information Processing Systems 35, pp. 507–520. Curran Associates, Red Hook, NY, USA (2022)

[13] Collins, G.S., Moons, K.G.M., Dhiman, P., *et al.*: TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. BMJ **385**, 078378 (2024) https://doi.org/10.1136/bmj-2023-078378

[14] Moons, K.G.M., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S.: PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann. Intern. Med. **170**, 1–33 (2019) https://doi.org/10.7326/M18-1377

[15] Hopkins, C., Gillett, S., Slack, R., Lund, V.J., Browne, J.P.: Psychometric Validity of the 22-Item Sinonasal Outcome Test. Clin. Otolaryngol. **34**, 447–454 (2009) https://doi.org/10.1111/j.1749-4486.2009.01995.x

[16] Lund, V.J., Kennedy, D.W.: Staging for Rhinosinusitis. Otolaryngol. Head Neck Surg. **117**, 35–40 (1997) https://doi.org/10.1016/S0194-5998(97)70005-6

[17] Lund, V.J., Mackay, I.S.: Staging in Rhinosinusitus. Rhinology **31**, 183–184 (1993)

[18] Breiman, L.: Random Forests. Mach. Learn. **45**, 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[19] Geurts, P., Ernst, D., Wehenkel, L.: Extremely Randomized Trees. Mach. Learn. **63**, 3–42 (2006) https://doi.org/10.1007/s10994-006-6226-1

[20] Cortes, C., Vapnik, V.: Support-Vector Networks. Mach. Learn. **20**, 273–297 (1995) https://doi.org/10.1007/BF00994018

[21] Zou, H., Hastie, T.: Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. B **67**, 301–320 (2005) https://doi.org/10.1111/j.1467-9868.2005.00503.x

[22] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785 . 13–17 August 2016

[23] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30, pp. 3146–3154. Curran Associates, Red Hook, NY, USA (2017)

[24] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased Boosting with Categorical Features. In: Advances in Neural Information Processing Systems 31, pp. 6638–6648. Curran Associates, Red Hook, NY, USA (2018)

[25] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**, 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[26] Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine Learning Algorithm Validation with a Limited Sample Size. PLoS ONE **14**, 0224365 (2019) https://doi.org/10.1371/journal.pone.0224365

[27] Van Calster, B., McLernon, D.J., Smeden, M., Wynants, L., Steyerberg, E.W.: Calibration: The Achilles Heel of Predictive Analytics. BMC Med. **17**, 230 (2019) https://doi.org/10.1186/s12916-019-1466-7

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.*: Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

[29] Bland, J.M., Altman, D.G.: Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet **1**, 307–310 (1986) https://doi.org/10.1016/S0140-6736(86)90837-8