Can Tabular Foundation Models Improve Frailty Detection Using HRV?

Saman Parvaneh[1], Sadaf Moharreri[2], Shahab Rezaei[3], Nima Toosizadeh[4]

[1]Edwards Lifesciences, Irvine, California, USA
[2]Department of Biomedical Engineering, Kho.C., Islamic Azad University, Khomeinishahr, Iran
[3]Department of Biomedical Engineering, CT.C., Islamic Azad University, Tehran, Iran
[4]Department of Rehabilitation and Movement Sciences, School of Health Professions, Rutgers University, Newark, New Jersey, USA

Abstract

Tabular foundation models (TFMs) are pre-trained transformers reported to perform well on small tables, the regime typical of clinical studies. We asked whether a TFM improves frailty detection from heart rate variability (HRV) over conventional classifiers. We analyzed single-lead ECG recorded during a six-minute walk test from 67 patients after open-heart surgery (38 frail, 29 non-frail by the Edmonton Frail Scale). Thirty-six HRV features were extracted. We evaluated ten classifiers, including three TFMs (TabPFN, TabICL, and TabDPT) across seven predictor sets spanning HRV, demographics, clinical history, and physical function, using 10 repeats of 5-fold cross-validation with all preprocessing inside the training folds, participant-level bootstrap intervals, and label-permutation tests. No predictor set exceeded chance. The best result was CatBoost on HRV, AUC 0.626 (95% CI 0.511-0.778, permutation p = 0.083). All three TFMs ranked below CatBoost, which outperformed TabPFN in a paired test (p = 0.035). Clinical history was worse than a non-informative baseline, and combining predictor blocks did not help. This cohort supports 80% power only for a true AUC near 0.70, so the finding is an inconclusive null rather than evidence against an HRV-frailty association. Pre-trained tabular models did not compensate for a small, range-restricted cohort.

1. Introduction

Frailty is a syndrome of diminished physiological reserve and predicts adverse outcomes after cardiac surgery [1, 2]. It is consistently linked to impaired cardiac autonomic control, often reflected in reduced heart rate variability (HRV) [3]. HRV measured during physical activity may reveal dysregulation that resting measurements miss [4, 5], and a wearable ECG makes such measurement inexpensive.

The obstacle is sample size. Frailty cohorts are small, and HRV feature sets are large and collinear, a regime in which flexible models overfit and a single train-test split yields a confident but irreproducible estimate [6]. Tabular foundation models (TFMs) are motivated by exactly this setting: pre-trained across many tabular tasks, they perform in-context learning in a forward pass with no per-task tuning [7-9].

We previously reported CatBoost achieving AUC 0.82 on this cohort using a single data split [10]. Here we ask two questions under one protocol: does a TFM improve on conventional classifiers, and does any available predictor class detect frailty once evaluation removes split-to-split variance?

2. Methods

2.1. Data

Data come from the PhysioNet database of wearable-based signals during physical exercises from patients with frailty after open-heart surgery (coronary artery bypass, isolated valve, or combined) [11, 12] at Kulautuva Rehabilitation Hospital, Lithuania. A Polar H10 chest strap recorded single-lead ECG at 130 Hz during five standardized tests. We analyzed 67 participants with an analyzable six-minute walk test (6MWT) recording: 38 frail and 29 non-frail, aged 72.6±5.4 years (65-87), 34% female.

Frailty was assessed with the Edmonton Frail Scale (EFS), a nine-domain instrument (cognition, general health, functional independence, social support, medication use, nutrition, mood, continence, functional performance) scored 0–17 [13]. Subjects with EFS <= 5 and >= 6 were considered non-frail and frail, respectively. Enrolment required EFS >= 4, so the cohort spans only 4-9 and the non-frail group consists of participants scoring 4 or 5. The task is therefore to separate adjacent bands of a truncated scale rather than frail from robust individuals. The EFS score itself was excluded from every model, since the label is that score thresholded and retaining it yields AUC 1.000 trivially.

QRS detection used the Pan-Tompkins algorithm [14], followed by extraction of 36 HRV features spanning statistical, frequency-domain, Poincare, geometric Poincare, heart rate asymmetry, and heart rate

fragmentation families [3, 10, 15]. Demographic, clinical, and physical-function variables were taken from the database metadata.

2.2. Predictor sets

Seven predictor sets were evaluated: (a) demographics (age, sex, height, and weight), 4 variables; (b) HRV, 36; (c) HRV with demographics, 40; (d) clinical history, 11 (NYHA class, atrial fibrillation, COPD, depression, musculoskeletal and oncological disease, ACE inhibitors, beta blockers, calcium-channel blockers, surgery type, days since surgery); (e) physical function, 26 (six-minute walk distance and time, exercise-test duration, max load and max heart rate, and 21 instrumented gait measures); (f) clinical with functional, 37; and (g) all predictors, 77. A sensitivity analysis (h) repeated set (b) with in-fold principal components retaining 95% of variance.
Sets (e), (f) and (g) contain measurements that partly constitute the frailty construct, since the EFS scores timed mobility and the Fried phenotype defines frailty partly as slow gait [16]. We include them as an approximate upper bound, not as independent predictors; set (d) is the honest test of independent clinical information.

2.3. Models and Evaluation

We compared ten classifiers: a prior-predicting baseline, L2-regularised logistic regression, logistic regression with in-fold univariate selection, a decision tree, a random forest, XGBoost [17], CatBoost [18], and three TFMs (TabPFN v2 [7], TabICL [9], TabDPT [8]). The prior-predicting baseline ignores the features and assigns every held-out participant the class prior of its training fold; because stratified partitioning yields folds with slightly different priors, its pooled out-of-fold predictions take only three distinct values and its AUC is 0.470 rather than 0.500, so chance should be read as 0.5 and the permutation null, not this baseline, is the reference. We deliberately constrained tree and boosting models, since 67 observations with correlated predictors are past the point where unconstrained ensembles memorize their training folds. The three foundation models perform in-context learning by using the training fold as inference context, without gradient updates or hyperparameter search. TabPFN targets small samples through synthetic-task pretraining, TabICL uses column-then-row attention to scale to larger tables, and TabDPT uses real-data self-supervised pretraining with retrieval-based context selection.

We estimated performance using 10 repeats of stratified 5-fold cross-validation. Within each repeat, we pooled held-out predictions into one out-of-fold vector covering all participants and computed AUC once on that vector. We fitted all preprocessing (imputation, zero-variance filtering, scaling, and feature selection) only within the training folds. Confidence intervals came from bootstrap resampling of participants (2,000 resamples); significance from 120 label permutations evaluated through the complete pipeline, with Benjamini-Hochberg correction across arms; and model comparisons from a paired bootstrap scoring both models on the same resample.

3. Results

No predictor set achieved discrimination distinguishable from chance (Table 1-2 and Figure 1). The best result was CatBoost on HRV, AUC 0.626 (95% CI 0.511-0.778), with a permutation p-value of 0.083. After correction, every arm-level test remained non-significant. In the clinical arm, no trained model beat the non-informative baseline (0.470).

All three TFMs ranked below CatBoost on HRV (TabDPT 0.584, TabICL 0.568, TabPFN 0.535 versus 0.626). Of 15 paired TFM-versus-conventional comparisons, 13 had intervals spanning zero; the two exceptions point in opposite directions, CatBoost outperforming TabPFN (difference 0.103, 95% CI 0.009-0.193, p = 0.035) and TabDPT outperforming logistic regression (0.107, 0.007-0.209, p = 0.038), with the comparator itself at chance. Combining predictor blocks did not help. Adding demographics to HRV moved the best AUC from 0.626 to 0.596 and pooling all 77 predictors reached 0.583, but tested directly these differences were not significant (p = 0.365 and 0.399). Compressing the HRV panel by principal components also did not help (0.570). Calibration was no better than a constant prediction: a model predicting the cohort prevalence attains a Brier score of 0.245, which the best model in six of seven arms did not beat. Because the binary label is a thresholded ordinal score, we repeated the benchmark as regression on the EFS score. Three of seven arms became significant and survived correction: demographics (R2 = 0.175, Spearman rho = 0.426, p = 0.008), physical function, and all predictors (both p = 0.017, q = 0.039), whereas none was significant against the binary label. HRV alone remained null under both codings.

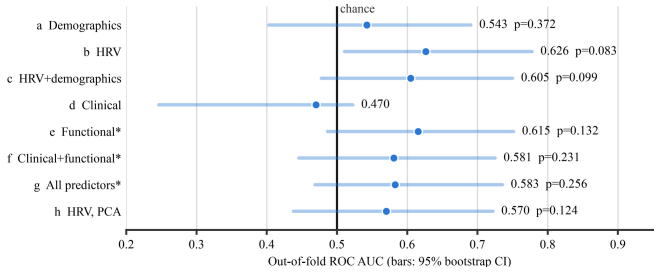

Figure 1. Best model in each predictor set. Points are mean out-of-fold AUC over 10 cross-validation repeats; bars are 95% bootstrap confidence intervals over participants. The vertical rule marks chance.

Table 1. Mean out-of-fold AUC for every model in every predictor set. Arm h omits the foundation models.

Model	a: demo (4)	b: HRV (36)	c: HRV+demo (40)	d: clinical (11)	e: functional* (26)	f: clin+func* (37)	g: all* (77)	h: HRV-PCA
CatBoost	0.493	**0.626**	0.596	0.375	0.570	0.538	**0.583**	0.570
Random forest	**0.543**	0.612	**0.605**	0.455	0.524	0.523	0.581	0.551
XGBoost	0.538	0.551	0.526	0.448	0.509	0.509	0.538	0.534
Decision tree	0.514	0.564	0.543	0.456	0.521	0.512	0.509	0.552
Logistic regression	0.448	0.504	0.484	0.315	0.569	0.490	0.406	0.481
Logistic + top-10 (in CV)	—	0.472	0.447	—	**0.615**	**0.581**	0.523	—
TabDPT	0.461	0.584	0.565	0.385	0.576	0.515	0.428	—
TabICL	0.517	0.568	0.542	0.367	0.557	0.511	0.445	—
TabPFN	0.388	0.535	0.504	0.403	0.577	0.552	0.506	—
Baseline (prior)	0.470	0.470	0.470	**0.470**	0.470	0.470	0.470	0.470

Table 2. Best model in each predictor set. Asterisks mark blocks containing predictors that partly constitute the frailty construct. p is the label-permutation p-value; no arm is significant before or after correction.

Predictor Set	N	Best model	AUC	95% CI	Permutation p
b: HRV	36	**CatBoost**	0.626	0.511–0.778	0.083
e: Functional*	26	Logistic + top-10 (in CV)	0.615	0.487–0.751	0.132
c: HRV + demographics	40	Random forest	0.605	0.478–0.750	0.099
g: All predictors*	77	CatBoost	0.583	0.469–0.735	0.256
f: Clinical + functional*	37	Logistic + top-10 (in CV)	0.581	0.446–0.725	0.231
h: HRV, PCA	36	CatBoost	0.570	0.438–0.722	0.124
a: Demographics	4	Random forest	0.543	0.403–0.691	0.372
d: Clinical	1	**Baseline (prior)**	0.470	0.246–0.523	0.835

4. Discussion

Across 74 model-by-predictor-set combinations, nothing separated frail from non-frail. The methods span linear models, tree ensembles, and pre-trained transformers, and their convergence into a narrow band with overlapping intervals indicates a property of the data rather than of any estimator. TFMs conferred no advantage: the model whose design target best matches this cohort, TabPFN, ranked last and was significantly outperformed by CatBoost.

Two features of the cohort bound what was achievable. First, enrolment required EFS >= 4, so the comparison is between adjacent bands of a truncated scale; restriction of range attenuates association independently of sample size. Second, under the Hanley-McNeil approximation [17] this cohort supports 80% power only for a true AUC near 0.70,

and detecting an effect the size observed would require roughly 200 participants. The result is therefore an inconclusive null, not evidence against an HRV-frailty association.

The present estimate is lower than the AUC of 0.82 we reported earlier on these data [10]. That estimate came from a single 60/40 split; in this cohort, it has a standard deviation of 0.083 across splits and a central 95% range of 0.433-0.761, so 0.82 lies in its extreme upper tail. Repeated cross-validation reduces that standard deviation to 0.032, and uses every participant for both training and evaluation. On samples of this size, the evaluation design moves the reported number more than any modeling choice. That the physical-function block, which partly constitutes the frailty definition, also failed to reach significance bounds how much any independent signal could contribute here.

5. Limitations and Future Directions

The cohort is small (n = 67), single-source and narrow in age, and no external validation set was available; the narrow age range in particular explains why demographics carry so little information here. Future work should improve data informativeness by recruiting robust adults (EFS 0–3), modeling EFS as ordinal, and using the full multimodal protocol, including accelerometry and response-recovery dynamics. Any promising model should be externally validated with a pre-specified AUC-based analysis, rather than expanded through additional architectures that mainly increase false-positive risk.

Tabular synthetic data generation may also help expand effective training size, especially when external-cohort priors are learned, and synthesis is confined to training folds.

6. Conclusion

Neither heart rate variability, demographics, clinical history, nor physical function detected frailty above chance in this post-cardiac-surgery cohort, and tabular foundation models did not change that. The binding constraint is the dataset (size, label resolution, and range restriction), rather than the choice of algorithm.

Disclosures

S. Parvaneh is an employee of Edwards Lifesciences. The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of Edwards Lifesciences. This work used publicly available data, was conducted independently of the authors' employers, and the employers had no role in the study design, analysis, interpretation, or the decision to publish.

References

[1] M. A. Makary *et al.*, "Frailty as a Predictor of Surgical Outcomes in Older Patients," *The American College of Surgeons,* vol. 210, pp. 901–908, 2010.

[2] M. Pozzi *et al.*, "The frail patient undergoing cardiac surgery: lessons learned and future perspectives," *Frontiers in Cardiovascular Medicine,* vol. 10, 2023.

[3] S. Parvaneh *et al.*, "Regulation of cardiac autonomic nervous system control across frailty statuses: a systematic review," *Gerontology,* vol. 62, no. 1, pp. 3–15, 2015.

[4] M. Eskandari, S. Parvaneh, H. Ehsani, M. Fain, and N. Toosizadeh, "Frailty identification using heart rate dynamics: a deep learning approach," *IEEE journal of biomedical and health informatics,* vol. 26, no. 7, pp. 3409–3417, 2022.

[5] N. Toosizadeh *et al.*, "Frailty and heart response to physical activity," *Archives of gerontology and geriatrics,* vol. 93, p. 104323, 2021.

[6] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *Neuroimage,* vol. 180, pp. 68–77, 2018.

[7] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature,* vol. 637, no. 8045, pp. 319–326, 2025.

[8] J. Ma *et al.*, "TabDPT: Scaling tabular foundation models on real data," *Advances in Neural Information Processing Systems,* vol. 38, pp. 172692–172722, 2026.

[9] J. Qu, D. HolzmÃžller, G. Varoquaux, and M. L. Morvan, "TabICL: A tabular foundation model for in-context learning on large data," *arXiv preprint arXiv:2502.05564,* 2025.

[10] S. Parvaneh, S. Moharreri, N. Toosizadeh, and S. Rezaei, "Frailty Assessment using HRV During Physical Activity," *Computing in Cardiology,* vol. 52, 2025.

[11] A. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation,* vol. 101, no. 23, p. e215, 2000.

[12] D. Sokas *et al.*, "Wearable-based Signals during Physical Exercises from Patients with Frailty after Open-Heart Surgery (version 1.0.0)," *PhysioNet,* 2022.

[13] D. B. Rolfson, S. R. Majumdar, R. T. Tsuyuki, A. Tahir, and K. Rockwood, "Validity and reliability of the Edmonton Frail Scale," *Age and ageing,* vol. 35, no. 5, pp. 526–529, 2006.

[14] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE transactions on biomedical engineering,* no. 3, pp. 230–236, 1985.

[15] M. D. Costa, R. B. Davis, and A. L. Goldberger, "Heart rate fragmentation: a new approach to the analysis of cardiac interbeat interval dynamics," *Frontiers in physiology,* vol. 8, p. 255, 2017.

[16] L. P. Fried *et al.*, "Frailty in older adults: evidence for a phenotype," *The journals of gerontology series a: biological sciences and medical sciences,* vol. 56, no. 3, pp. M146–M157, 2001.

[17] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, no. 1, pp. 29–36, 1982.

Address for correspondence:
Saman Parvaneh
1 Edwards Way, Irvine, CA, USA
parvaneh@ieee.org