applied sciences

MDPI

Article

Explanatory Versus Predictive Modeling in Heart Rate Variability Research: Nonlinearity, Model Choice, and Machine Learning

Junichiro Hayano [1] and Emi Yuda [2,3,*]

[1] Graduate School of Medical Sciences, Nagoya City University, Nagoya 467-8601, Japan; hayano@nagoya-cu.ac.jp
[2] Innovation Center for Semiconductor and Digital Future, Mie University, Tsu 514-8507, Japan
[3] Department of Management Science and Technology, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan
* Correspondence: yuda@icsdf.mie-u.ac.jp

Abstract

Heart rate variability (HRV) changes nonlinearly across the lifespan, and such nonlinearity is often taken to require machine learning (ML). Distinguishing explanatory from predictive modeling, we tested three propositions using 24 h Holter recordings from a clinically referred population (ALLSTAR database), sampled uniformly across ages 0–102 years (n = 3854). Explanatory models expressed 12 HRV indices as functions of age and sex; predictive models estimated age, an error-free benchmark, from HRV under nested, age-stratified cross-validation. First, all indices showed significant nonlinear age dependence, but general linear models with spline terms explained it as well as random forest and gradient boosting (cross-validated R^2 differences ≤ 0.02) while providing parameters, confidence intervals, and tests. Second, gradient boosting reduced the mean absolute error of linear regression from 11.3 to 10.1 years (95% CI of the reduction in repeated analyses, 0.8–1.3 years), with gains largest in children and older adults, where all models had their largest errors and bias, and reduced to 0.3 years under the database age distribution. Third, permutation importance agreed only moderately with explanatory age dependence (Spearman ρ 0.48–0.73), differed between models and permutation schemes, and was allocated within correlated index groups in model-specific ways. In this clinically referred, cross-sectional dataset, nonlinearity did not imply that ML was required, and variable importance should not be read as physiological importance.

Keywords: heart rate variability; aging; explanatory modeling; predictive modeling; machine learning; variable importance; Holter electrocardiography; restricted cubic spline; autoregressive variability

Check for updates

Academic Editors: Stefan H. Bossmann and Jennifer Davis

Received: 5 October 2026
Revised: 17 October 2026
Accepted: 24 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Heart rate variability (HRV), the beat-to-beat fluctuation of the cardiac interval, has been used for more than four decades as a non-invasive window into autonomic cardiovascular control [1,2]. Spectral components of HRV have been linked to specific regulatory mechanisms, most notably the high-frequency (HF) component to vagally mediated respiratory sinus arrhythmia [3], and standards and normative values have been established for time-domain, frequency-domain, and nonlinear indices [2,4]. At the same time, the physiological interpretation of several indices remains contested; the ratio of low-frequency to high-frequency power (LF/HF), for example, does not accurately measure cardiac sympatho-vagal balance [5]. With the spread of wearable sensors, HRV metrics are now

applied far beyond the clinical laboratory [6], including respiratory-rate estimation and respiration-related analyses in which HRV is interpreted through explicit physiological models [7,8]. In all of these applications, HRV indices are interpreted through statistical models, and what can legitimately be concluded from a model depends on the purpose for which it was built.

Statistical models serve two distinct purposes [9–12]. Explanatory models aim to estimate how an outcome depends on its determinants; their estimands are interpretable parameters, and they are judged by the validity and precision of those estimates. Predictive models aim to minimize error on new observations; their estimand is the expected prediction error, and they are judged only by out-of-sample performance. The two goals can favor different models even when applied to the same data, yet they are frequently conflated in biomedical applications. In HRV research, three forms of this conflation are common. First, nonlinear relationships between HRV and its determinants are often taken to require flexible machine learning (ML) models, even when the goal is explanation. Second, the predictive accuracy of complex models is attributed to their complexity without comparison with well-specified linear models under equal conditions of tuning and validation. Third, the variable importance of a predictive model is read as a ranking of the physiological importance of its inputs.

These questions have become pressing as ML has entered the analysis of electrocardiographic and HRV data. Twenty-four-hour ambulatory (Holter) electrocardiography (ECG) remains the richest routine source of HRV data. Classical Holter-derived risk markers such as heart rate turbulence and microvolt T-wave alternans were developed on the basis of explicit physiological mechanisms and standardized measurement procedures [13,14], and large Holter databases make it possible to describe how autonomic indices vary with basic demographic factors such as age and sex [15]. In parallel, ML and deep learning have transformed ECG analysis. Beyond classical atrial fibrillation (AF) detectors based on RR-interval irregularity [16,17], convolutional neural networks now classify arrhythmias directly from ECG signals [18,19], achieve cardiologist-level performance on ambulatory recordings [20], identify patients with AF from ECGs recorded during sinus rhythm [21], and predict the short-term onset of AF [22]. Deep learning applied to 24 h Holter ECG has also been used to predict cardiac outcomes in patients with heart failure or myocardial infarction and to predict life-threatening ventricular arrhythmias [23,24]. These models are optimized for predictive accuracy, and their internal representations are generally not designed to be interpreted physiologically.

To address this opacity, explainable artificial intelligence (XAI) methods such as local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) [25,26] have been applied to ECG-based models; for example, for the diagnosis of AF from Holter waveforms and for the detection of specific arrhythmias [27–29]. Simpler measures of variable importance, such as permutation importance and standardized regression coefficients, are used for the same purpose. All of these measures describe how a fitted predictor uses its inputs; they do not by themselves estimate how a physiological quantity depends on its determinants. When the inputs are strongly correlated, as many HRV indices are, the importance attributed to each input also depends on how the model shares the common information among them, so that importance rankings may differ between models fitted to the same data.

Age provides a clear test case for these questions. Most HRV indices change nonlinearly across the lifespan: they rise during childhood and adolescence, decline through adulthood, and in some cases change again in old age [30,31]. This includes the autoregressive (AR) variability (ARV), which quantifies the stochastic component of HRV after removal of the predictable AR structure; in the Allostatic State Mapping by Ambulatory

ECG Repository (ALLSTAR) database, ARV was high at birth, decreased until young adulthood, and increased again in old age [31], and its estimation requires a robust AR order-selection procedure [32]. A common view holds that such nonlinearity makes linear models unsuitable both for explaining and for predicting age-related HRV changes. This view conflates linearity in the variables with linearity in the parameters. Nonlinear age effects can be represented within a general linear model through spline terms, and a multivariable linear predictor can combine indices whose age trajectories differ. Moreover, chronological age is recorded with negligible error and has well-established nonlinear associations with HRV, so that it can serve as a benchmark target for which the explanatory question (how does HRV depend on age?) and the predictive question (how accurately can age be recovered from HRV?) can be posed on the same data. How much is gained by ML for each of these purposes, and how its variable importance relates to the explanatory age dependence, has not been quantified in lifespan HRV data.

In this study, we used 24 h Holter ECG recordings from a clinically referred Holter population, the ALLSTAR database, sampled uniformly across ages 0 to 102 years in both sexes, and restricted the determinants to age and sex, which are reliably recorded in this database. We tested three propositions corresponding to the three forms of conflation described above. First, a nonlinear age dependence of HRV does not require ML for explanation: general linear models with spline terms should describe the age trajectories of 12 HRV indices as well as random forest and gradient boosting models evaluated on the same cross-validation folds, while also providing parameters, confidence intervals, and tests. Second, in predicting age from HRV, the gain from ML over linear regression is small when all models are tuned and evaluated by nested, age-stratified cross-validation. Third, the variable importance of a predictive model should not be read as the physiological importance of an HRV index; we compared permutation importance with the partial R^2 of age from the explanatory models, examined its stability across folds and models, and examined how it is distributed within groups of correlated indices, including whether it concentrates on indices that are not redundant with the others. Finally, we examined the robustness of these findings to quality control, the exclusion of recordings with paroxysmal AF, robust estimation, the omission of LF/HF, and the age distribution of the sample.

2. Materials and Methods

2.1. ALLSTAR Database

This study used the database of the Allostatic State Mapping by Ambulatory ECG Repository (ALLSTAR) project [31,33]. A 24 h Holter recording contains approximately 100,000 beats of ECG waveforms in an adult, yet in routine clinical practice only a fraction of this information, mainly that related to arrhythmia and transient myocardial ischemia, is used. The ALLSTAR project aims to establish methods for evaluating the impact of environmental factors on health and disease from Holter ECG data recorded throughout Japan, and thereby to contribute to predictive and preventive medicine and health promotion. To this end, the project has been collecting hundreds of thousands of Holter ECG recordings from all over Japan and constructing a time-series database that links ECG information (heart rate, HRV, heart rate dynamics, arrhythmia, transient myocardial ischemia, and ECG waveform variability) with temporal and environmental information and with basic medical information such as age and sex [15,33].

The database comprises Holter ECG recordings that were obtained for clinical purposes and referred for analysis to the Holter ECG analysis centers of Suzuken Co., Ltd. (Tokyo, Nagoya, and Sapporo, Japan) [31,33]. Recordings were made with Cardy series Holter recorders (Suzuken Co., Ltd., Nagoya, Japan), digitized at 125 Hz, and analyzed with a Holter ECG analyzer (Cardy Analyzer 05, Suzuken Co., Ltd.) by skilled medical

technologists, who manually corrected all errors of automated QRS detection and rhythm annotation [31]. The data are anonymized with a unique research number so that the source cannot be identified; the correspondence between names and research numbers is kept confidential by Suzuken Co., Ltd. and is not disclosed to anyone outside the company, including members of the ALLSTAR study group. Patients may request that their data not be used, and such data are excluded [33].

2.2. Study Sample and Quality Control

From the database, 3915 recordings (1927 men and 1988 women) aged 0 to 102 years were randomly extracted so that the age distribution was approximately uniform, with 30 to 40 recordings per year of age. The extraction was designed so that each individual contributed only one recording. The recordings were made between 2007 and 2014. Because information on body mass index, clinical diagnoses, and medication was not consistently available, and because sleep stages and AF episodes cannot currently be determined with validated reliability from these data, the analysis was restricted to age and sex as determinants.

For the primary analysis, recordings were required to meet the quality control (QC) criteria used in our previous study of the same database [31]: deceleration capacity (DC) > 4.5 ms and short-term scaling exponent of detrended fluctuation analysis (DFA α1) > 0.75. These criteria excluded 61 recordings (1.6%), leaving 3854 recordings (1899 men and 1955 women) for the primary analysis. These thresholds were taken from our previous study of the same database, in which they formed part of the criteria for a low-risk HRV profile [31]. Because DC and DFA α1 are also analyzed as outcomes and predictors, such thresholds can truncate genuine physiological variation, particularly at the extremes of age, where most exclusions occurred. We therefore compared indicators of recording quality that do not depend on the analyzed HRV values—the proportion of NN intervals among all beats (NN coverage) and the numbers of supraventricular and ventricular ectopic beats—between excluded and included recordings of the same age group, and we repeated the analyses with a quality criterion based only on NN coverage (Section 2.8). Two sensitivity datasets were defined: (i) all 3915 recordings without exclusion, and (ii) a strict dataset (n = 3556) that additionally required a proportion of normal-to-normal (NN) intervals of at least 80% of all beats, a standard deviation of NN intervals (SDNN) > 75 ms, and ln very-low-frequency (VLF) power > 5.75, approximating the full inclusion criteria of the previous study [31]. The non-Gaussianity index λ25s used in that study was not available in the present dataset.

2.3. HRV Indices

HRV was analyzed from the 24 h series of NN intervals according to published standards [2,4] and our previous methods [31]. The following 12 indices were used: mean heart rate; SDNN; natural-log-transformed power of the ultra-low-frequency (ULF, <0.0033 Hz), VLF (0.0033–0.04 Hz), low-frequency (LF, 0.04–0.15 Hz), and high-frequency (HF, 0.15–0.40 Hz) components; the LF/HF ratio; DC measured by phase-rectified signal averaging (PRSA); short-term (α1, 4–11 beats) and long-term (α2, >11 beats) scaling exponents obtained by DFA [34]; the spectral exponent β of the $1/f\beta$ power-law relationship; and ARV. ARV quantifies the random component of HRV as the fractional variance of the residual of an AR model fitted to the NN-interval series x(t):

$$x(t) = \sum_{i=1}^{p} a(i)\, x(t-i) + Z(t), \qquad \mathrm{ARV} = \frac{\sigma^2\,(Z(t))}{\sigma^2\,(x(t))} \tag{1}$$

where a(i) are the AR coefficients, Z(t) is the residual series, and the model order p was selected by the final prediction error criterion [31,32]. ARV was natural-log-transformed before analysis.

2.4. Analytical Framework

The study contrasts two analyses that use the same data but answer different questions (Table 1). The explanatory analysis follows the physiological direction, from age to HRV, and models each HRV index j as a function of age and sex:

$$\mathrm{HRV}_{ij} = f_j(\mathrm{Age}_i, \mathrm{Sex}_i) + \varepsilon_{ij}, \qquad j = 1, \ldots, 12 \tag{2}$$

Table 1. Analytical framework: explanatory versus predictive modeling of the age–HRV relationship.

Aspect	Explanatory Analysis	Predictive Analysis
Question	How does each HRV index depend on age and sex?	How accurately can age be recovered from HRV in new individuals?
Model form	$\mathrm{HRV_j} = \mathrm{f_j}(\mathrm{age}, \mathrm{sex}) + \varepsilon$, fitted separately for each of the 12 indices	$\mathrm{Age} = \mathrm{g}(\mathrm{HRV}_1, \ldots, \mathrm{HRV}_{12}, \mathrm{sex}) + \mathrm{e}$
Direction of variables	Age → HRV (physiological direction: age is a determinant of autonomic function)	HRV → age (inverse problem: age is inferred from its consequences)
Estimand	Age trajectories $\mathrm{f_j}$ and their parameters (turning points, sex differences)	Expected prediction error of g on new individuals
Models	General linear models with restricted cubic splines of age × sex (primary); linear age × sex (reference); RF and GBM (comparison)	Linear regression (reference); additive spline, RF, and GBM, tuned by nested CV
Evaluation criterion	Validity and precision of estimates: confidence intervals, tests, AIC; CV R^2 to check overfitting	Out-of-sample error only: test MAE, RMSE, and R^2 in nested CV
Interpretation of output	Coefficients and curves describe how HRV changes with age, with uncertainty	Predictions and variable importance describe how g uses its inputs, not how HRV depends on age
Role of age	Determinant	Benchmark target, chosen because it is recorded without error and has known nonlinear associations with HRV

AIC, Akaike information criterion; CV, cross-validation; GBM, gradient boosting machine; MAE, mean absolute error; RF, random forest; RMSE, root-mean-square error.

Its estimands are the age trajectories f_j and the parameters that describe them, and it is judged by the validity and precision of these estimates. Here, "explanatory" refers to modeling HRV as a function of prespecified determinants to estimate interpretable associations; it does not imply causal identification from the present cross-sectional observational data.

The predictive analysis reverses the direction and treats age as the quantity to be recovered from HRV, which is an inverse problem:

$$\mathrm{Age}_i = g(\mathrm{HRV}_{i1}, \ldots, \mathrm{HRV}_{i12}, \mathrm{Sex}_i) + e_i \tag{3}$$

Its estimand is the expected error of g on new individuals, and it is judged only by out-of-sample performance. Age was chosen as the target of the predictive analysis not because age prediction is clinically useful, but because chronological age is recorded with negligible error and has well-established nonlinear associations with HRV [30,31]. Clinical targets such as sleep stage or AF are affected by label uncertainty, which would confound a comparison of explanatory and predictive models with the quality of the reference standard. With age as the target, the two questions can be posed on exactly the same data, and differences between their answers can be attributed to the goals of

the models rather than to measurement error in the target. The three propositions of the study map onto this framework: proposition 1 concerns the choice of f_j, proposition 2 the choice of g, and proposition 3 the relationship between the variable importance of g and the explanatory estimates of f_j. The explanatory general linear models were compared with each other by the Akaike information criterion (AIC) and with random forest (RF) and gradient boosting machine (GBM) models by cross-validation (CV); the predictive models were evaluated by nested CV with the mean absolute error (MAE) and root-mean-square error (RMSE).

2.5. Explanatory Models

For each HRV index, the dependence on age and sex was modeled with a general linear model in which age entered through a restricted cubic spline with 5 degrees of freedom, allowing an age-by-sex interaction:

$$y_i = \beta_0 + \sum_{k=1}^{5} \beta_k s_k(\mathrm{Age}_i) + \gamma \,\mathrm{Sex}_i + \sum_{k=1}^{5} \delta_k s_k(\mathrm{Age}_i)\,\mathrm{Sex}_i + \varepsilon_i \quad (4)$$

where $s_k(\cdot)$ are the spline basis functions. This model is nonlinear in age but linear in its parameters; its degrees of freedom were specified in advance and not tuned. The basis was a natural cubic regression spline with boundary knots at 0 and 102 years and interior knots at the quintiles of age (20.4, 40.8, 61.2, and 81.6 years), centered so that the intercept remained identifiable. The linear reference model contained age, sex, and their interaction, so that the spline model differed from it only in the shape of the age terms. The coefficient of determination (R^2), the AIC, and the likelihood-ratio test for nonlinearity (8 degrees of freedom) all compared the spline model with this same reference. The age-by-sex interaction was tested by comparing the spline model with a model containing the spline terms and sex without interaction. The strength of the age dependence of each index was quantified by the partial R^2 of the age terms, defined as the proportional reduction in the residual sum of squares when the spline age terms and their interaction with sex were added to a model containing sex alone.

To test whether ML offers an advantage for explanation, RF and GBM models with age and sex as the only inputs were fitted to each HRV index. Their 10-fold CV R^2 was compared with that of the linear reference and spline models, computed on the same folds (Section 2.6). Their hyperparameters were selected with the same grids and inner CV as in the predictive analysis (Section 2.6), with the MAE of the HRV index as the inner-loop objective; age and sex entered untransformed. Unlike the spline models, the ML models do not provide parameters, confidence intervals, or tests for the age dependence; their fitted age curves were therefore compared with the spline curves graphically.

2.6. Predictive Models

Age was predicted from the 12 HRV indices and sex using four models: ordinary least-squares linear regression, an additive spline model (cubic B-spline expansion of each standardized predictor with a ridge penalty), RF, and histogram-based GBM. All models were evaluated by nested CV. The outer loop was a 10-fold CV stratified by 5-year age group, and the same outer folds were used for all analyses. Because each individual contributed a single recording, the folds were disjoint with respect to individuals. Within each outer training set, the hyperparameters of the additive spline, RF, and GBM models were selected by a 3-fold inner CV stratified by age group, minimizing the MAE. The grids were as follows: additive spline, 4, 6, or 8 knots and ridge penalty 0.01, 0.1, 1, 10, or 100; RF, 150 trees, 33% or 67% of predictors per split, and minimum leaf size 5 or 20; GBM, learning rate 0.1 with up to 300 iterations and early stopping (10% internal validation, 10 rounds without improvement), maximum 15 or 31 leaves, and minimum leaf size 20 or 50. Linear

regression has no hyperparameters. Standardization and the spline basis of the additive model were fitted within each training set as part of the model pipeline, with knots placed uniformly between the minimum and maximum of each standardized predictor in that training set, so that no preprocessing used information from the held-out fold. The tuned model was refitted to the whole outer training set and evaluated on the held-out fold. Performance was summarized by the MAE, RMSE, and R^2 on held-out data:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{Age}_i - \hat{\mathrm{Age}}_i\right|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Age}_i - \hat{\mathrm{Age}}_i\right)^2} \tag{5}$$

Training-set MAE was also computed to quantify the gap between in-sample and out-of-sample performance, and the mean prediction bias was examined as a function of age.

Because several hyperparameters selected in the primary analysis lay at the edges of the grids (Section 3.3), the predictive analysis was repeated with expanded grids: additive spline, 2, 3, 4, 6, or 8 knots and ridge penalty 0.001 to 100; RF, 33%, 67%, or 100% of predictors per split and minimum leaf size 1, 2, 5, or 20; and GBM, maximum 15, 31, 63, or 127 leaves and minimum leaf size 2, 5, 10, 20, 50, 100, or 200. This analysis was run with five repetitions of the nested CV, each with a different random assignment of the outer folds. To quantify the uncertainty of the differences in performance between models, 95% confidence intervals (CIs) of the paired differences in MAE and R^2 were obtained by bootstrapping (2000 resamples). In each bootstrap sample, one repetition of the nested CV was drawn at random and individuals were resampled with replacement within 5-year age groups; the difference between two models was then computed on the same resampled individuals. The resampling therefore accounted for the sharing of test cases between models and, through the random choice of repetition, for the variability of model fitting across fold assignments. Because no equivalence margin had been specified in advance, differences are reported with their CIs and their size relative to the MAE of linear regression rather than classified as negligible. The analysis code, including the spline specifications, the preprocessing pipelines, and the inner-loop objectives, is provided as Supplementary File S1.

2.7. Variable Importance Versus Explanatory Estimates

To examine proposition 3, the variable importance of the predictive models was compared with the strength of the age dependence estimated by the explanatory models. For each predictive model and outer fold, permutation importance was computed on the held-out fold as the increase in MAE when the values of one predictor were randomly permuted (20 permutations), so that importance reflects the contribution of a predictor to out-of-sample prediction. For linear regression, the absolute standardized coefficients were also recorded. Agreement between the importance ranking and the ranking of the partial R^2 of age across the 12 HRV indices was assessed with Spearman's rank correlation coefficient (ρ). Stability was assessed across folds, as the mean pairwise Spearman ρ between the importance vectors of the 10 folds, and across models, as the Spearman ρ between the mean importance vectors of different models.

To examine how importance is distributed among correlated predictors, the Pearson correlation matrix of the 12 HRV indices was computed, and groups of strongly correlated indices ($|r| \geq 0.7$) were identified by hierarchical clustering. The redundancy of each index was quantified as the R^2 of a linear regression of the index on the other 11 indices; $1 - R^2$ represents the information in an index that is not shared with the others. The association of mean importance with redundancy and with the partial R^2 of age was assessed with Spearman's ρ. Because LF/HF is a deterministic function of ln LF and ln HF (LF/HF = exp[ln LF − ln HF]), permuting one of these indices alone creates mathematically

inconsistent combinations. The importance analysis was therefore repeated (i) with models fitted without LF/HF; (ii) with a derivation-consistent permutation, in which LF/HF was recomputed after ln LF or ln HF had been permuted; (iii) with grouped permutations, in which the same random reordering of individuals was applied to all members of a group (ln LF, ln HF, and LF/HF, and each correlated group), so that their joint distribution was preserved; and (iv) with a conditional permutation, in which each predictor was replaced by its linear prediction from the other predictors, estimated in the training fold, plus a permuted residual, so that its relationship with the other predictors was approximately preserved. The share of importance of an index was defined as its marginal permutation score divided by the sum of the positive scores of all indices; because the permutation scores of correlated predictors are not additive, these shares express relative magnitudes and do not partition the predictive information.

2.8. Sensitivity and Robust Analyses

The explanatory and predictive analyses were repeated in the two sensitivity datasets. In addition, robust methods were applied to all 3915 recordings: Huber regression ($\varepsilon = 1.35$) with the same spline design, winsorization at the 1st and 99th percentiles followed by ordinary least squares, and tree-based ensembles. For the explanatory models, the deviation of each age curve from that of the primary analysis was expressed as the maximum absolute difference over ages 0 to 100 years in either sex, divided by the standard deviation (SD) of the index in the primary dataset. For the predictive models, only the training data were varied across datasets, and all models were evaluated on the same held-out quality-controlled recordings so that differences in performance could not arise from differences in the test set. Winsorization thresholds were derived from the training folds only. To assess sensitivity to a quality criterion that does not depend on the analyzed HRV values, the analyses were also repeated in recordings with NN coverage of at least 80% (n = 3633; 282 excluded, 22 of which had also failed the primary criteria). To examine how the conclusions extend to the excluded recordings, the predictive models trained on the QC-passed recordings of each outer training set were also applied to the excluded recordings of the corresponding held-out fold.

In a further sensitivity analysis, all recordings annotated as paroxysmal AF were excluded from the primary dataset (n = 3516; 1730 men and 1786 women), and the explanatory and predictive analyses were repeated (Supplementary Tables S1 and S2 and Figure S1). Because the LF/HF ratio does not index sympatho-vagal balance [5], the predictive analysis was also repeated with LF/HF omitted from the predictors (11 HRV indices and sex; Supplementary Table S4). The explanatory models do not require this analysis, because each is fitted to a single index.

2.9. Influence of the Age Distribution

Because the sample was drawn uniformly across ages, whereas the ALLSTAR database is dominated by older adults (men, 65 ± 17 years; women, 67 ± 17 years) [31], we examined how the age distribution affects the evaluation of the predictive models. Each recording i was assigned an importance weight

$$w_i = \frac{f_{\text{target}}(a_i \mid \text{sex}_i)}{\hat{f}_{\text{sample}}(a_i \mid \text{sex}_i)} \tag{6}$$

where a_i is the age of recording i, the target density is a normal density with the sex-specific mean and SD of the database truncated to 0–103 years, and the sample density is the empirical proportion of recordings of the same sex at that year of age in the sample (Supplementary Figure S2). The weights were normalized to a mean of 1; the effective

sample size was 2155. The target densities were fixed in advance from published statistics of the database [31] and did not depend on the study data. For evaluation, the weights of the held-out recordings were computed from these target densities and the age distribution of the whole sample; they only define the population to which the error refers and were not used for fitting or tuning. For retraining, the sample density was re-estimated within each outer training set from its recordings only, and the resulting weights were used both for fitting and for the weighted MAE that served as the objective of the inner CV, so that no information from the held-out fold entered the derivation of the training weights. Predictive performance was evaluated with weighted MAE, RMSE, and R^2, both for the models trained on the uniform sample and for the models retrained with the weights.

2.10. Software and Ethics

Analyses were performed in Python 3.11 with statsmodels 0.15 and scikit-learn 1.8. The analysis code is provided as Supplementary File S1. This study was performed according to a protocol approved by the Ethics Review Committee of Nagoya City University Graduate School of Medical Sciences and Nagoya City University Hospital (No. 709 and No. 60160209). In accordance with the Japanese ethical guidelines for medical and health research involving human subjects, the purpose of the research and the use of the information were disclosed publicly, and the research subjects were given the opportunity to refuse participation (opt-out) [31,33]. Generative artificial intelligence (Claude, Anthropic) was used to assist with writing the analysis code, preparing the figures, and drafting the text; all code, results, and text were reviewed and verified by the authors.

3. Results

3.1. Study Population

Of the 3915 recordings, 61 (1.6%) failed the quality criteria and 3854 (1899 men and 1955 women) were included in the primary analysis (Table 2). The mean age was 47.9 $\pm$ 27.4 years in men and 49.3 $\pm$ 28.3 years in women. The excluded recordings were concentrated at the extremes of age (11 at 0–9 years and 39 at $\geq$80 years), and only nine of them were annotated as paroxysmal AF. Recordings with a paroxysmal AF annotation that met the quality criteria (n = 338) were retained. Compared with included recordings of the same age group, the excluded recordings had lower NN coverage (median 87.3% vs. 97.2% at <10 years and 90.5% vs. 94.5% at $\geq$80 years), more often had NN coverage below 80% (45% vs. 5% at <10 years and 36% vs. 17% at $\geq$80 years), and at $\geq$80 years had more supraventricular ectopic beats (median 51 vs. 21 per recording) (Supplementary Table S9). Thus, many but not all exclusions were accompanied by independent evidence of reduced recording quality.

Table 2. Number of recordings by age group and sex in the primary analysis, and recordings excluded by quality control.

Age Group (Years)	Men	Women	Total	Excluded (Paroxysmal AF)
0–9	184	185	369	11 (3)
10–19	199	200	399	1 (1)
20–39	400	400	800	0 (0)
40–59	400	399	799	1 (0)
60–79	397	394	791	9 (0)
$\geq$80	319	377	696	39 (5)
Total	1899	1955	3854	61 (9)

Excluded, recordings with deceleration capacity (DC) $\leq$ 4.5 ms or DFA α1 $\leq$ 0.75; values in parentheses are recordings annotated as paroxysmal atrial fibrillation (AF).

3.2. Proposition 1: Explanation of the Nonlinear Age Dependence

All 12 HRV indices showed significant nonlinear dependence on age when the spline model was compared with the linear reference model containing age, sex, and their interaction ($p < 0.001$ for all; Table 3). The spline terms increased R^2 from 0.29 to 0.43 for heart rate, from 0.04 to 0.17 for SDNN, from 0.15 to 0.32 for DC, and from 0.10 to 0.26 for DFA α1, and reduced the AIC by 108 to 846. The partial R^2 of age ranged from 0.080 (LF/HF) to 0.424 (heart rate).

Table 3. Explanatory models of HRV indices as functions of age and sex (primary analysis, n = 3854).

HRV Index	R^2 Linear	R^2 Spline	ΔAIC	P Nonlinearity	P Age × Sex	Partial R^2 of Age	Sex: Men − Women, Age-Standardized (95% CI)
Heart rate (bpm)	0.290	0.428	−820	<0.001	<0.001	0.424	−2.261 (−2.848 to −1.674)
SDNN (ms)	0.042	0.165	−515	<0.001	<0.001	0.163	+3.076 (+0.756 to +5.395)
ln ULF (ms^2)	0.034	0.088	−208	<0.001	<0.001	0.088	−0.044 (−0.088 to +0.000)
ln VLF (ms^2)	0.226	0.312	−434	<0.001	<0.001	0.286	+0.257 (+0.221 to +0.294)
ln LF (ms^2)	0.339	0.401	−368	<0.001	0.002	0.387	+0.263 (+0.219 to +0.307)
ln HF (ms^2)	0.287	0.369	−458	<0.001	<0.001	0.368	−0.068 (−0.117 to −0.018)
LF/HF	0.105	0.164	−250	<0.001	<0.001	0.080	+0.880 (+0.791 to +0.968)
DC (ms)	0.154	0.324	−846	<0.001	<0.001	0.315	+0.336 (+0.250 to +0.422)
DFA α1	0.101	0.263	−746	<0.001	<0.001	0.187	+0.100 (+0.091 to +0.110)
DFA α2	0.415	0.434	−108	<0.001	<0.001	0.411	−0.022 (−0.025 to −0.019)
Spectral exponent β	0.189	0.293	−512	<0.001	0.379	0.293	−0.006 (−0.012 to +0.000)
ln ARV	0.189	0.302	−563	<0.001	0.088	0.302	+0.029 (−0.009 to +0.068)

R^2 linear, reference model with age, sex, and age × sex; R^2 spline, restricted cubic spline of age (5 degrees of freedom) interacting with sex; ΔAIC, AIC of the spline model minus that of the reference model; P nonlinearity, likelihood-ratio test of the spline model against the same reference model (8 degrees of freedom); P age × sex, likelihood-ratio test of the age-by-sex interaction within the spline model; partial R^2 of age, proportional reduction in residual variance by the spline age terms and their interaction with sex; sex, difference between men and women estimated from the interacting spline model and averaged over ages 0–100 years with equal weights (standardized to the uniform age distribution of the sample); because the contrast changes with age for several indices, age-specific contrasts are given in Supplementary Table S7 and Figure S3. ARV, autoregressive variability; CI, confidence interval; DC, deceleration capacity; DFA, detrended fluctuation analysis; HF, LF, VLF, ULF, high-, low-, very-low-, and ultra-low-frequency power; SDNN, standard deviation of normal-to-normal intervals.

The spline models revealed characteristic age trajectories (Figure 1A). Heart rate was highest in infancy, decreased steeply until about 20 years of age, and declined gradually thereafter. SDNN peaked at approximately 22 years and DC at 23 to 26 years, whereas LF and HF power peaked in adolescence (15 to 20 years) and reached a minimum at 74 to 83 years, after which HF power increased again. DFA α1 followed an inverted U-shape with a maximum at 47 to 51 years, and DFA α2 increased from adolescence until the ninth decade. The age-by-sex interaction was significant for 10 of the 12 indices (all except spectral exponent β and ARV). Because of these interactions, sex differences were estimated as age-specific contrasts from the interacting spline model (Supplementary Table S7 and Figure S3) and summarized as average contrasts standardized to the uniform age distribution of the sample (Table 3). Men had higher VLF and LF power, LF/HF, and DFA α1 than women from about 7–8 years of age onward, and women had higher DFA α2 from 10 years onward. For other indices, the direction of the sex difference changed with age: heart rate was higher in boys up to 7 years but higher in

women at 12–41 and 49–100 years; SDNN was higher in men at 9–44 years but higher in women at 60–75 years; HF power was higher in men at 18–24 years but higher in women at 43–85 years; ln ULF was higher in men at 19–37 years but higher in women at 52–87 years; and DC was higher in girls up to 6 years but higher in men from 15 years onward. For these indices, the age-standardized average contrast in Table 3 conceals opposite differences at different ages.

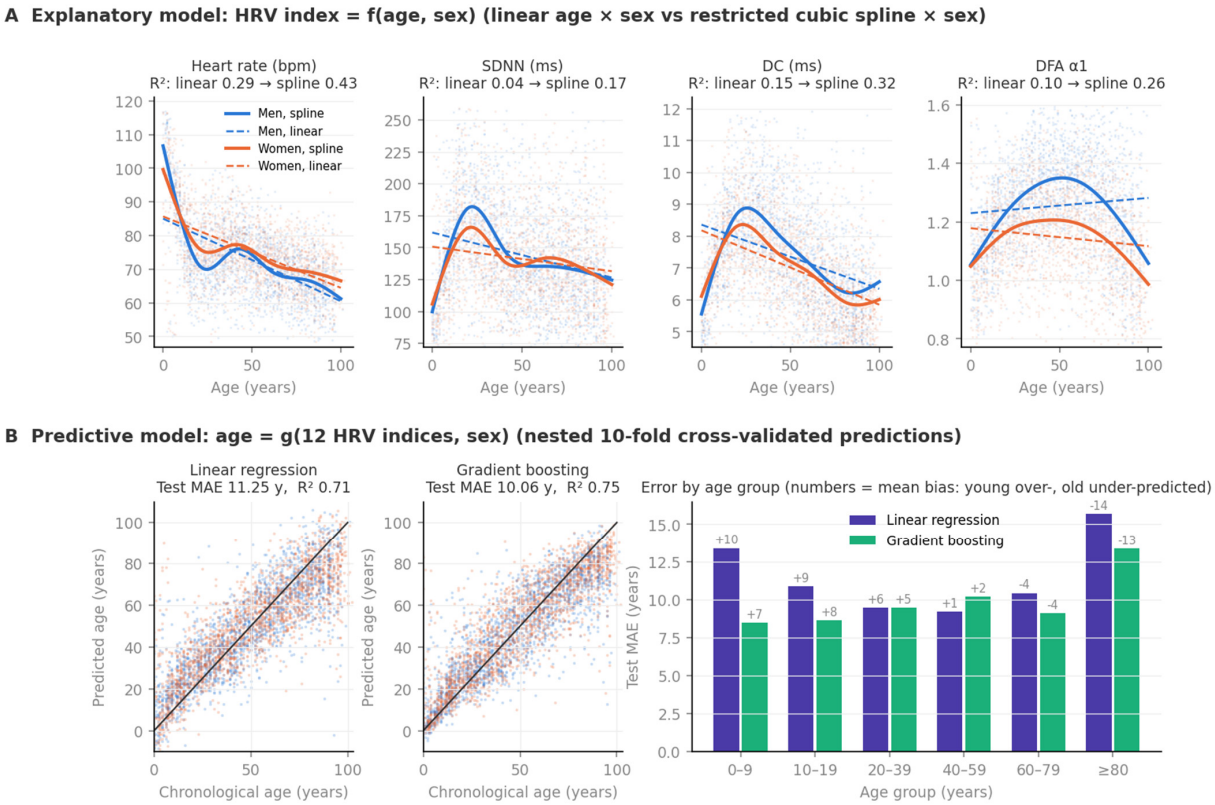


Figure 1. Explanatory and predictive models of age-related HRV changes (primary analysis, n = 3854). (**A**) Explanatory models: heart rate, SDNN, DC, and DFA α1 as functions of age and sex, fitted with the linear reference model (age, sex, and age × sex; dashed lines) and a restricted cubic spline of age interacting with sex (solid lines); dots are individual recordings (blue, men; orange, wom-en). (**B**) Predictive models: age estimated from 12 HRV indices and sex by linear regression and gradient boosting (nested 10-fold cross-validated predictions; dots are individual recordings, blue for men and orange for women as in (**A**); the diagonal is the line of identity), and test MAE by age group, with mean bias shown above each bar (purple, linear regression; green, gradient boosting).

ML did not materially improve the fit, while the spline model directly provided interpretable inferential quantities (Table 4; Figure 2). On the same cross-validation folds, the CV R^2 of the spline models was 0.083 to 0.429, whereas that of gradient boosting differed from it by −0.012 to +0.009 and that of random forest by −0.021 to +0.003. Gradient boosting exceeded the spline model for four indices (heart rate, SDNN, ln ULF, and ln VLF) by at most 0.009, and random forest exceeded it only for ln VLF (+0.003). The in-sample R^2 of the spline models exceeded their CV R^2 by at most 0.006, indicating no meaningful overfitting. The age curves fitted by the ML models followed the spline trajectories but showed local irregularities (Figure 2), and, unlike the spline models, they provided no parameters, confidence intervals, or tests for nonlinearity, turning points, or sex differences.

Table 4. Cross-validated R^2 of explanatory models with age and sex as inputs: general linear models versus machine learning.

HRV Index	Linear Age × Sex	Spline × Sex	Random Forest	Gradient Boosting	In-Sample − CV R^2, Spline
Heart rate (bpm)	0.288	0.425	0.425	0.429	0.003
SDNN (ms)	0.040	0.161	0.152	0.163	0.004
ln ULF (ms^2)	0.031	0.083	0.074	0.084	0.006
ln VLF (ms^2)	0.225	0.307	0.310	0.316	0.005
ln LF (ms^2)	0.337	0.397	0.385	0.392	0.004
ln HF (ms^2)	0.285	0.365	0.354	0.360	0.005
LF/HF	0.103	0.158	0.137	0.146	0.006
DC (ms)	0.153	0.319	0.307	0.313	0.005
DFA α1	0.100	0.257	0.241	0.249	0.005
DFA α2	0.414	0.429	0.422	0.427	0.004
Spectral exponent β	0.188	0.288	0.276	0.284	0.005
ln ARV	0.188	0.298	0.283	0.290	0.004

All models were evaluated on the same 10 outer folds stratified by age group; the hyperparameters of random forest and gradient boosting were tuned by inner CV. Only the general linear models provide parameters, confidence intervals, and tests for the age dependence (Table 3).

3.3. Proposition 2: Prediction of Age from HRV

Under nested CV, linear regression achieved a test MAE of 11.25 years and R^2 of 0.71 (Table 5). The additive spline model reduced the MAE to 10.87 years, random forest to 10.16 years, and gradient boosting to 10.06 years (R^2 = 0.75); the largest gain over linear regression was thus 1.19 years. The inner CV consistently selected 4 knots and a ridge penalty of 0.1 for the additive spline model, 67% of predictors per split and a minimum leaf size of 5 for random forest, and a minimum leaf size of 20 with 31 leaves (15 in one fold) for gradient boosting (Supplementary Table S5). Several of these values lay at the edges of the grids. With the expanded grids, the inner CV selected 2 knots (the lowest possible, i.e., a cubic polynomial per predictor) for the additive spline model, all predictors or 67% of predictors per split with a minimum leaf size of 2 for random forest, and interior values for gradient boosting (mostly 31 leaves and a minimum leaf size of 5–20) (Supplementary Table S10); the performance differences persisted (Tables 6 and 7). The tree-based ensembles fitted the training data more closely than the test data (training MAE 6.48 years for random forest and 7.04 years for gradient boosting), whereas the training MAE of linear regression (11.20 years) was nearly identical to its test MAE.

The advantage of the nonlinear models depended on age (Table 6; Figures 1B and 3). It was largest at the extremes of the age range: in children aged 0–9 years, the test MAE was 13.4 years for linear regression and 8.5 years for gradient boosting, and in adults aged $\geq$ 80 years it was 15.7 and 13.4 years, respectively. A smaller gain was also present at 60–79 years (10.4 vs. 9.1 years), whereas between 20 and 59 years the MAEs of all models were similar (9.2 to 10.2 years), and between approximately 30 and 60 years linear regression was as accurate as or more accurate than gradient boosting (Figure 3A). All models showed regression toward the mean, overestimating age in the young (mean bias +7.5 to +9.9 years at 0–9 years) and underestimating it in those aged $\geq$ 80 years ($-$12.5 to $-$14.1 years). Consequently, the youngest and oldest individuals were predicted with the largest errors by every model: the test MAE at 0–9 years and at $\geq$80 years exceeded that at 20–59 years by up to 4.2 and 6.5 years, respectively, and the errors were systematic rather than random, with predicted age pulled toward the middle of the age range at both ends (Figure 3B).

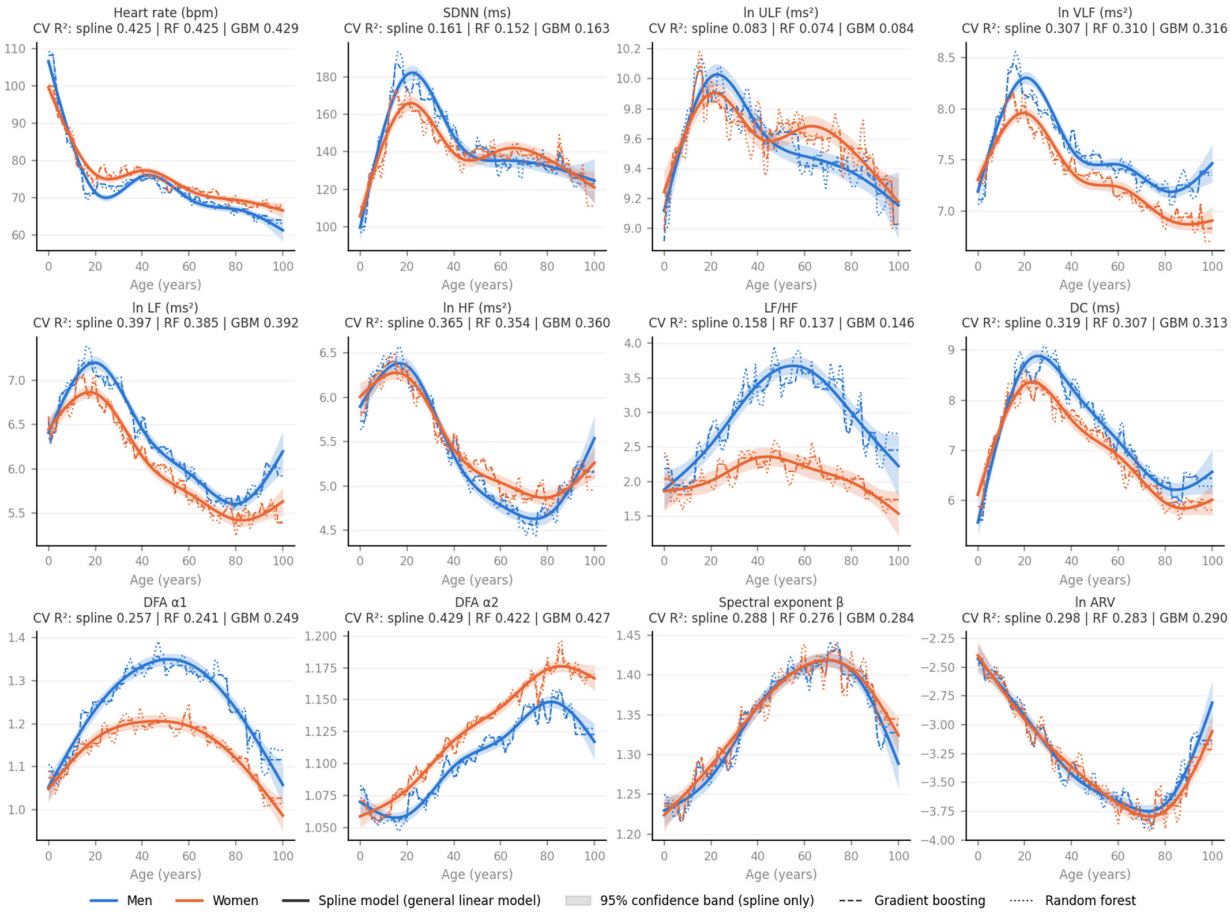


Figure 2. Explanatory models of the 12 HRV indices with age and sex as inputs: general linear models with restricted cubic splines (solid lines, with 95% confidence bands) versus gradient boosting (dashed lines) and random forest (dotted lines). Blue, men; orange, women. The cross-validated R^2 of the spline, RF, and GBM models is shown above each panel. The machine-learning models reproduce the spline trajectories with additional local irregularity, but provide no confidence bands, parameters, or tests. RF, random forest; GBM, gradient boosting.

Table 5. Performance of predictive models estimating age from 12 HRV indices and sex (primary analysis, nested 10-fold CV).

Model	Test MAE (Years)	Test RMSE (Years)	Test R^2	Training MAE (Years)
Linear regression	11.25	15.02	0.71	11.20
Additive spline	10.87	14.38	0.73	10.68
Random forest	10.16	13.96	0.75	6.48
Gradient boosting	10.06	13.83	0.75	7.04

Hyperparameters were tuned within each outer training set by a 3-fold inner CV stratified by age group. Training MAE is the mean over the 10 outer training sets.

Across five repetitions of the nested CV with expanded grids, the mean test MAE was 11.25 years for linear regression, 10.82 years for the additive spline model, 10.12 years for random forest, and 10.19 years for gradient boosting, varying by at most 0.11 years between repetitions (Table 7). Relative to linear regression, random forest reduced the MAE by 1.13 years (95% CI 0.91 to 1.36), gradient boosting by 1.06 years (95% CI 0.84 to 1.29), and the additive spline model by 0.43 years (95% CI 0.31 to 0.56), and the corresponding gains in R^2 were 0.041 (95% CI 0.030 to 0.053), 0.038 (0.026 to 0.050), and 0.024 (0.017 to 0.033). The gain of the ensembles over linear regression was therefore clearly different from zero

and amounted to about 10% of the MAE of linear regression; it was not an artifact of the restricted grids of the primary analysis. By age group, the reduction in MAE by gradient boosting was largest at 0–9 years (5.09 years; 95% CI 4.32 to 5.87) and ≥80 years (2.21 years; 95% CI 1.57 to 2.86), smaller but still present at 10–19 years (2.01 years; 95% CI 1.26 to 2.73) and 60–79 years (1.01 years; 95% CI 0.57 to 1.47), absent at 20–39 years (−0.15 years; 95% CI −0.64 to 0.39), and reversed at 40–59 years, where linear regression was more accurate (1.02 years in favor of linear regression; 95% CI 0.57 to 1.50).

Table 6. Differences in predictive performance between models with 95% confidence intervals (expanded hyperparameter grids, five repetitions of nested 10-fold CV).

Measure	Additive Spline	Random Forest	Gradient Boosting
Mean test MAE, years (range over repetitions)	10.82 (10.81–10.83)	10.12 (10.09–10.16)	10.19 (10.16–10.24)
Reduction in MAE, years (95% CI)	0.43 (0.31 to 0.56)	1.13 (0.91 to 1.36)	1.06 (0.84 to 1.29)
Relative reduction in MAE, %	3.8	10.0	9.4
Gain in R^2 (95% CI)	0.024 (0.017 to 0.033)	0.041 (0.030 to 0.053)	0.038 (0.026 to 0.050)
Reduction in MAE at 0–9 years (95% CI)	0.73 (0.27 to 1.17)	5.32 (4.50 to 6.12)	5.09 (4.32 to 5.87)
Reduction in MAE at 10–19 years (95% CI)	0.55 (0.20 to 0.91)	1.48 (0.80 to 2.12)	2.01 (1.26 to 2.73)
Reduction in MAE at 20–39 years (95% CI)	−0.04 (−0.29 to 0.20)	−0.03 (−0.50 to 0.45)	−0.15 (−0.64 to 0.39)
Reduction in MAE at 40–59 years (95% CI)	−0.05 (−0.25 to 0.16)	−0.42 (−0.81 to −0.01)	−1.02 (−1.50 to −0.57)
Reduction in MAE at 60–79 years (95% CI)	0.65 (0.43 to 0.87)	1.27 (0.83 to 1.74)	1.01 (0.57 to 1.47)
Reduction in MAE at ≥80 years (95% CI)	1.04 (0.69 to 1.41)	1.65 (1.00 to 2.28)	2.21 (1.57 to 2.86)

Reductions are the MAE of linear regression minus that of the model (positive values favor the model); gains in R^2 are the R^2 of the model minus that of linear regression. CIs were obtained by a paired bootstrap of individuals within 5-year age groups (2000 resamples), with a randomly selected repetition of the nested CV in each resample. The mean test MAE of linear regression was 11.25 years (range 11.24–11.25). CI, confidence interval.

Table 7. Test MAE (mean bias) of age prediction by age group (years).

Model	0–9	10–19	20–39	40–59	60–79	≥80
Linear regression	13.4 (+9.9)	10.9 (+9.2)	9.5 (+6.2)	9.2 (+0.9)	10.4 (−4.5)	15.7 (−13.6)
Additive spline	12.8 (+8.3)	10.4 (+8.2)	9.6 (+5.7)	9.3 (+1.4)	9.7 (−4.0)	14.7 (−12.9)
Random forest	8.6 (+8.3)	9.7 (+9.3)	9.3 (+5.9)	9.4 (+1.6)	8.9 (−4.2)	14.5 (−14.1)
Gradient boosting	8.5 (+7.5)	8.6 (+7.6)	9.5 (+5.2)	10.2 (+1.5)	9.1 (−3.7)	13.4 (−12.5)

Values are test MAE with mean bias (predicted minus chronological age) in parentheses.

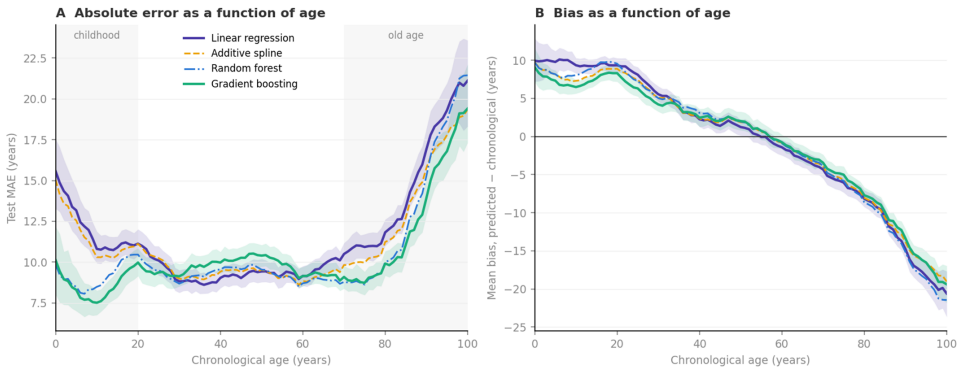


Figure 3. Prediction error as a function of chronological age (primary analysis, n = 3854; nested 10-fold cross-validated predictions). (**A**) Test MAE and (**B**) mean bias (predicted minus chronological age) for the four predictive models, smoothed with a moving window of ±5 years. Shaded bands are bootstrap 95% confidence intervals for linear regression and gradient boosting. The advantage of the nonlinear models is largest at the extremes of the age range (grey backgrounds), whereas in mid-adulthood linear regression is as accurate as or more accurate than the ensembles.

3.4. Proposition 3: Variable Importance Versus Explanatory Estimates

The importance rankings of the predictive models agreed only moderately with the ranking of the age dependence of the indices (Table 8; Figure 4B). Spearman's ρ between mean permutation importance and partial R^2 of age was 0.73 ($p = 0.007$) for linear regression, 0.53 ($p = 0.075$) for the additive spline model, 0.72 ($p = 0.008$) for random forest, and 0.48 ($p = 0.118$) for gradient boosting. Heart rate, which had both the strongest age dependence (partial R^2 0.424) and the lowest redundancy with the other indices (R^2 0.44), ranked first in every model and every fold and accounted for 25% to 44% of the total importance. Beyond heart rate, the rankings diverged from the explanatory estimates: the spectral exponent β, with the seventh-strongest age dependence (partial R^2 0.293), ranked 9th to 12th, whereas SDNN, with the tenth-strongest age dependence (partial R^2 0.163), ranked 3rd in linear regression and gradient boosting and 2nd in the additive spline model.

Table 8. Strength of age dependence (explanatory) versus permutation importance (predictive) of the 12 HRV indices.

HRV Index	Partial R^2 of Age (Rank)	R^2 with Other Indices	Linear Regression	Additive Spline	Random Forest	Gradient Boosting
Heart rate (bpm)	0.424 (1)	0.44	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)
DFA α2	0.411 (2)	0.83	4 (4–6)	3 (3–4)	2 (2–2)	2 (2–2)
ln LF (ms^2)	0.387 (3)	0.97	2 (2–3)	5 (5–6)	7 (5–7)	11 (9–12)
ln HF (ms^2)	0.368 (4)	0.98	5 (4–6)	7 (7–7)	3 (3–3)	4 (4–8)
DC (ms)	0.315 (5)	0.63	8 (8–9)	8 (8–9)	6 (4–7)	5 (4–7)
ln ARV	0.302 (6)	0.93	6 (4–6)	4 (3–4)	8 (8–8)	6 (4–7)
Spectral exponent β	0.293 (7)	0.88	9 (8–11)	10 (8–11)	11 (10–12)	12 (9–12)
ln VLF (ms^2)	0.286 (8)	0.87	10 (9–12)	12 (11–12)	4 (4–7)	8 (6–9)
DFA α1	0.187 (9)	0.89	7 (7–7)	6 (5–6)	9 (9–9)	7 (6–8)
SDNN (ms)	0.163 (10)	0.86	3 (2–3)	2 (2–2)	5 (4–7)	3 (3–3)
ln ULF (ms^2)	0.088 (11)	0.85	11 (9–12)	11 (9–12)	10 (10–12)	9 (8–11)
LF/HF	0.080 (12)	0.87	12 (10–12)	9 (8–11)	12 (11–12)	10 (9–12)

Importance columns show the rank of mean permutation importance among the 12 HRV indices, with the range of ranks across the 10 folds in parentheses. R^2 with other indices, R^2 of a linear regression of the index on the other 11 indices (redundancy). Importance values are given in Supplementary Table S6.

Within each model, the importance rankings were stable across folds (mean pairwise Spearman ρ 0.94 to 0.97), but they differed between models (ρ 0.59 between linear regression and gradient boosting, 0.53 between the additive spline model and random forest, and 0.82 between the two ensembles). The absolute standardized coefficients of linear regression ranked the indices almost identically to its permutation importance (ρ = 0.99). Thus, importance was a reproducible property of each fitted model rather than a model-independent property of the indices.

The correlation matrix revealed four groups of strongly correlated indices ($|r| \geq 0.7$): ln VLF–ln LF–ln HF, DFA α2–β–ln ARV, SDNN–ln ULF, and LF/HF–DFA α1 (Figure 4A). Within these groups, importance was concentrated on one member, and the member chosen depended on the model (Figure 4C). In the DFA α2–β–ln ARV group, DFA α2 received 82% of the group's importance in gradient boosting and 82% in random forest, whereas β received at most 0.30 years in any model despite a partial R^2 of 0.293. In the ln VLF–ln LF–ln HF group, linear regression assigned large individual importances to ln LF (7.22 years) and ln HF (5.04 years), but permuting the three indices jointly increased the MAE by only 2.26 years, less than one-fifth of the sum of the individual importances (12.45 years). In contrast, random forest distributed the group's importance among all three members. Across the 12 indices, importance was not significantly associated with non-redundancy

($1 - R^2$ with the other indices; Spearman ρ 0.01 to 0.47, all $p > 0.1$). These patterns were robust to the treatment of LF/HF and to the permutation scheme (Supplementary Table S8). In models fitted without LF/HF, Spearman's ρ between importance and partial R^2 was 0.45 to 0.64, heart rate ranked first in every model, and β ranked 9th to 11th. Derivation-consistent permutation, in which LF/HF was recomputed after ln LF or ln HF had been permuted, changed the individual importance of ln LF from 7.21 to 5.17 years in linear regression and from 2.23 to 4.49 years in the additive spline model. Conditional permutation reduced the importance of members of correlated groups to near zero (e.g., ln LF in linear regression, 0.19 years) and changed the rankings (Spearman's ρ with partial R^2, 0.38 to 0.56; between models, 0.48 to 0.98), whereas heart rate remained first and DFA α2 and SDNN remained among the four most important indices in every model. Thus, both the magnitude and the ranking of importance depended not only on the model but also on how the permutation was defined.

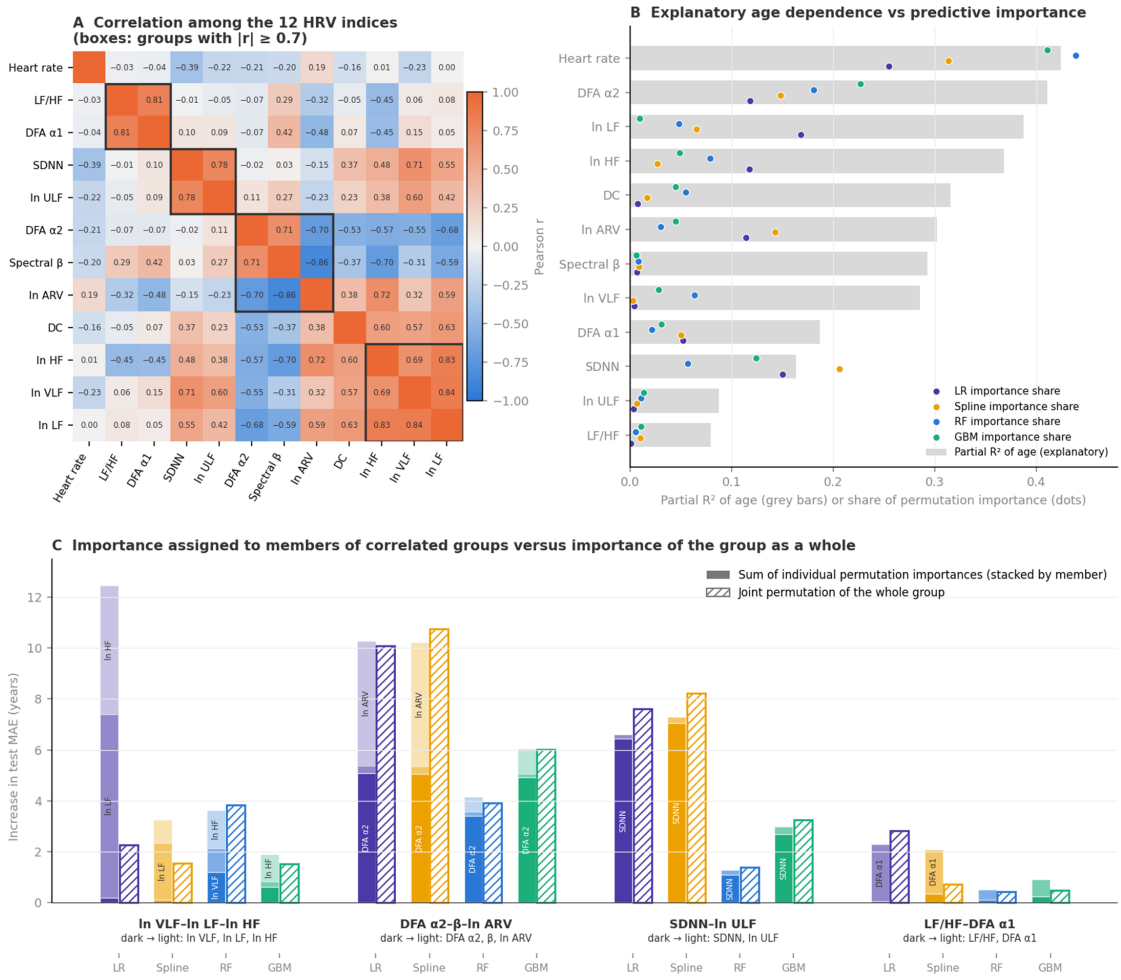


Figure 4. Variable importance of the predictive models versus explanatory estimates. (**A**) Pearson correlation matrix of the 12 HRV indices ordered by hierarchical clustering; boxes mark groups with $r \geq 0.7$. (**B**) Partial R^2 of age from the explanatory spline models (grey bars) and the share of each index in the sum of the positive marginal permutation scores in the four predictive models (dots; a relative magnitude, not an additive partition of predictive information), with indices ordered by partial R^2. (**C**) For each correlated group, the sum of the permutation importances of its members when permuted individually (solid bars, stacked by member) and the importance of the group when its members are permuted jointly (hatched bars). LR, linear regression; RF, random forest; GBM, gradient boosting machine.

3.5. Results of Sensitivity and Robustness Analyses

The explanatory models were robust to the handling of extreme values (Table 9; Figure 5B). When all recordings were included or when the variables were winsorized, the age curves deviated from those of the primary analysis by at most 0.15 SD, and the largest deviations occurred at the extremes of age. Including all recordings reduced the R^2 of the spline model for DC from 0.324 to 0.272, the largest change among the indices. The strict dataset produced larger deviations for ln VLF (0.35 SD), DFA α2 (0.34 SD), and ln ARV (0.29 SD); these criteria excluded 171 of 735 recordings (23%) from individuals aged ≥ 80 years. Huber regression deviated by up to 0.36 SD, mainly at the upper end of the age range, where it down-weighted observations in the region of greatest dispersion.

Table 9. Maximum deviation of the age curves from those of the primary analysis (SD units).

HRV Index	All Data	Strict Criteria	Winsorized	Huber Regression
Heart rate (bpm)	0.04	0.08	0.08	0.35
SDNN (ms)	0.03	0.13	0.03	0.20
ln ULF (ms^2)	0.07	0.11	0.09	0.13
ln VLF (ms^2)	0.03	0.35	0.03	0.15
ln LF (ms^2)	0.04	0.27	0.06	0.14
ln HF (ms^2)	0.07	0.23	0.05	0.07
LF/HF	0.04	0.07	0.05	0.33
DC (ms)	0.15	0.09	0.13	0.30
DFA α1	0.11	0.10	0.08	0.36
DFA α2	0.04	0.34	0.08	0.26
Spectral exponent β	0.07	0.28	0.14	0.36
ln ARV	0.10	0.29	0.07	0.18

Maximum absolute difference between the fitted curves over ages 0–100 years in either sex, divided by the SD of the index in the primary dataset. All data, n = 3915; strict criteria, n = 3556; winsorized and Huber regression, robust estimation on all data.

The predictive models were insensitive to the choice of training data (Table 10; Figure 5C). When evaluated on common QC-passed test recordings, the test MAE of each model differed by at most 0.13 years between training datasets, and robust estimators did not improve performance. When the strict dataset was evaluated on its own recordings, it appeared more accurate (linear regression, 10.70 years; gradient boosting, 9.76 years), but this apparent improvement disappeared when the test recordings were held constant (11.20 and 10.18 years), indicating that it reflected the exclusion of difficult cases rather than better models.

Excluding the 338 recordings annotated as paroxysmal AF (n = 3516) did not alter these findings (Supplementary Tables S1 and S2 and Figure S1): the spline terms remained significant for all indices ($p < 0.001$), the age curves deviated by at most 0.29 SD, and the test MAE ranged from 10.31 years (gradient boosting) to 11.22 years (linear regression). Omitting LF/HF from the predictors changed the test MAE by at most 0.06 years (Supplementary Table S4). With the quality criterion independent of the analyzed HRV values (NN coverage ≥ 80%, n = 3633), the spline terms remained significant for all indices ($p < 0.001$), the spline R^2 changed by at most 0.03, and the age curves deviated from those of the primary analysis by at most 0.05 SD between 10 and 79 years and by up to 0.40 SD at the upper boundary of the age range (Supplementary Table S9). The test MAE was 10.82, 10.47, 9.88, and 9.84 years for linear regression, the additive spline model, random forest, and gradient boosting, respectively, so that the gain of gradient boosting over linear regression (0.98 years) was similar to that in the primary analysis. When the models trained on QC-passed recordings were applied to the 61 excluded recordings, the MAE was 14.2 to 15.3 years, larger than in the included recordings because the excluded recordings were concentrated at the extremes of age, and the ensembles were not more accurate than

linear regression for these recordings (linear regression, 14.48 years; gradient boosting, 15.28 years). Including the excluded recordings in the test set changed the overall MAE of every model by less than 0.1 years (10.14 to 11.30 years).

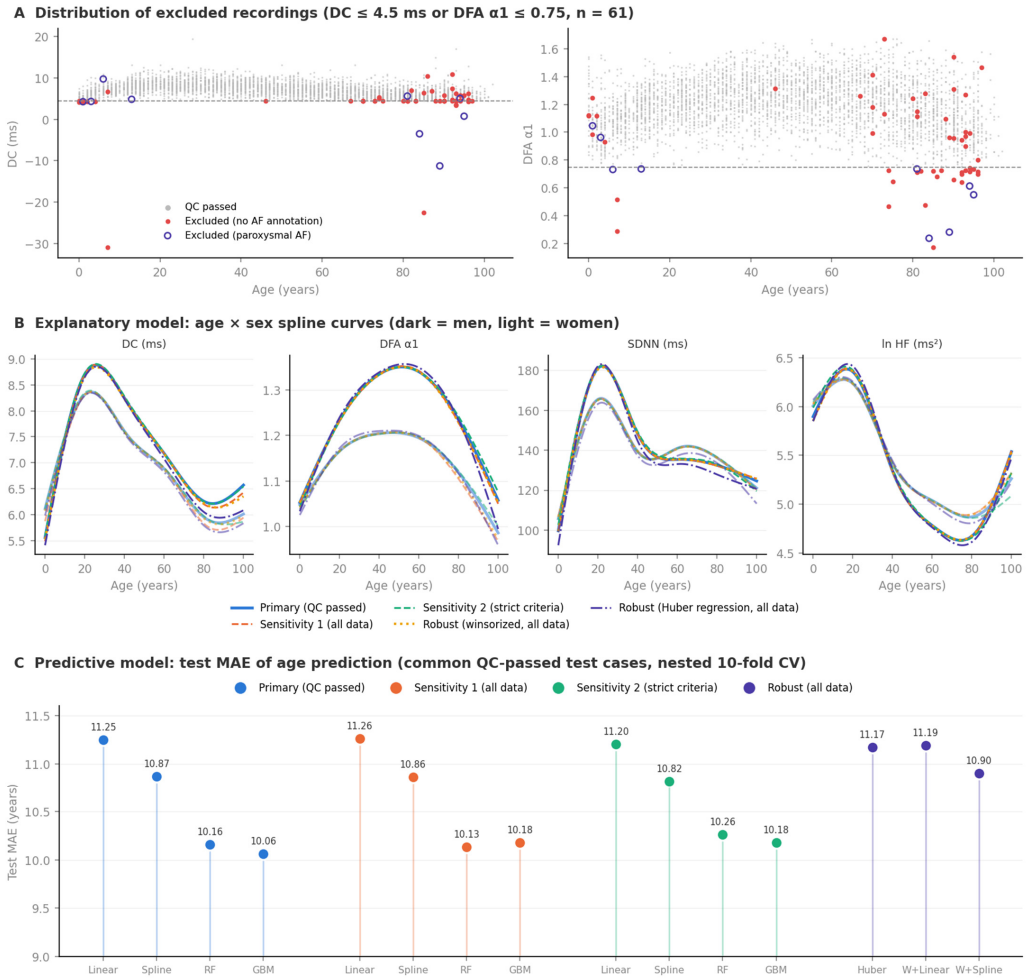


Figure 5. Sensitivity and robust analyses. (**A**) Distribution of the 61 recordings excluded by quality control (DC $\leq$ 4.5 ms or DFA α1 $\leq$ 0.75; dashed lines), distinguishing recordings in sinus rhythm (filled red) from those annotated as paroxysmal AF (open circles). (**B**) Age curves of DC, DFA α1, SDNN, and ln HF estimated in the primary analysis, in all data, in the strict dataset, and with winsorization and Huber regression on all data (dark lines, men; light lines, women). (**C**) Test MAE of age prediction for each training dataset and model, evaluated on common QC-passed test recordings (nested CV). GBM, gradient boosting; RF, random forest; W+, after winsorization.

Table 10. Test MAE (years) of age prediction evaluated on common quality-controlled test recordings for different training datasets (nested CV).

Training Data (n)	Linear	Additive Spline	Random Forest	Gradient Boosting
Primary, QC passed (3854)	11.25	10.87	10.16	10.06
All data (3915)	11.26	10.86	10.13	10.18
Strict criteria (3556)	11.20	10.82	10.26	10.18
All data, robust (3915)	11.17 [a]	10.90 [b]	—	—

In each outer fold, only the training data were varied; all models were tested on the same held-out QC-passed recordings. [a] Huber regression (winsorized linear regression: 11.19). [b] Additive spline after winsorization at the 1st and 99th percentiles. Random forest and gradient boosting are insensitive to monotone transformations of the predictors and were not refitted with robust losses.

3.6. Effects of Age-Distribution Reweighting on Model Performance

Reweighting the sample to the age distribution of the source database reduced the variance of age (weighted mean 64.7 years, SD 16.2 years) and markedly reduced R^2 while leaving the MAE similar (Table 11). For the models trained on the uniform sample, the weighted MAE was 9.93 to 10.80 years and R^2 was 0.25 to 0.35. When the models were retrained with weights estimated within each training set, the MAE decreased to 8.93 to 9.26 years and R^2 was 0.47 to 0.51. Under this distribution, the gain of the best models (additive spline and gradient boosting, 8.93 years) over linear regression (9.26 years) was 0.33 years, and random forest (9.19 years) was not more accurate than the additive spline model. Reweighting also reduced the variance in the HRV indices explained by age and the gain from the spline terms (Supplementary Table S3), because the weighted population contains few children, in whom the age trajectories are most strongly nonlinear.

Table 11. Predictive performance under the age distribution of the source database (importance-weighted evaluation, primary analysis, nested CV).

Model	Uniform: MAE	Uniform: R^2	Reweighted: MAE	Reweighted: R^2	Retrained: MAE	Retrained: R^2
Linear regression	11.25	0.71	10.80	0.25	9.26	0.47
Additive spline	10.87	0.73	10.29	0.34	8.93	0.51
Random forest	10.16	0.75	9.93	0.35	9.19	0.48
Gradient boosting	10.06	0.75	10.11	0.33	8.93	0.51

Uniform, unweighted evaluation of models trained on the uniform sample (Table 5); Reweighted, the same predictions evaluated with importance weights matching the sex-specific age distribution of the database (men, 65 ± 17 years; women, 67 ± 17 years); Retrained, models trained with importance weights estimated within each outer training set (also used as the inner-CV objective) and evaluated with the evaluation weights. Effective sample size, 2155. MAE in years.

4. Discussion

4.1. Principal Findings

Using 3854 quality-controlled 24 h Holter recordings from individuals aged 0 to 102 years, this study compared explanatory models of HRV as a function of age and sex with predictive models of age as a function of HRV on the same data, using age as an error-free benchmark target (Section 2.4). The three propositions were supported. First, the age dependence of all 12 HRV indices was strongly nonlinear, but general linear models with spline terms explained it as well as random forest and gradient boosting on cross-validation (differences in CV R^2 of −0.021 to +0.009), while also providing parameters, confidence intervals, and tests. Second, in predicting age, nested and tuned ML reduced the MAE of linear regression by at most 1.2 years (11.25 vs. 10.06 years; 1.1 years, 95% CI 0.8 to 1.3, in repeated analyses with expanded grids), with the gain largest in children and adults aged 80 years or older and smaller at 60 to 79 years, and by only 0.3 years under the age distribution of the source database. Third, the permutation importance of the predictive models agreed only moderately with the explanatory age dependence of the indices (Spearman ρ 0.48 to 0.73), differed between models and between permutation schemes although it was stable within each model, and was allocated within groups of correlated indices in model-specific ways. These conclusions were robust to quality-control exclusions, the exclusion of paroxysmal AF, robust estimation, the omission of LF/HF, the hyperparameter grids, a quality criterion independent of the HRV values, and the age distribution of the sample. They describe a clinically referred, cross-sectional Holter population and should be generalized with the caution discussed in Section 4.6.

4.2. Proposition 1: A Nonlinear Age Dependence Does Not Require ML for Explanation

The question of whether linear models can explain nonlinear age-related changes in HRV is resolved by distinguishing linearity in the parameters from linearity in the variables. Our explanatory models are general linear models in which age enters through a spline basis and interacts with sex. The analysis of variance with 5-year age strata used in our earlier study of the same database [31] is itself a linear model of this type, and the spline models reproduced its fit with fewer parameters while providing continuous estimates of turning points. The direct comparison with ML confirmed that nothing was lost by this choice: with age and sex as the only inputs, random forest and gradient boosting did not explain more of the variance in any index than the spline models beyond 0.01 in CV R^2, and the in-sample and cross-validated R^2 of the spline models were nearly identical. The ML age curves followed the spline trajectories with additional local irregularities that are unlikely to be physiological. More importantly, the two approaches differ in what they return. The spline models yield coefficients with confidence intervals, formal tests of nonlinearity and of age-by-sex interaction, and age-specific sex differences, which are the quantities required for physiological inference. The ML models return only a fitted function, from which such quantities can be obtained, if at all, only by additional resampling procedures without established inferential properties. For explanation, the general linear model is therefore not merely adequate but preferable.

The estimated trajectories were physiologically coherent. SDNN peaked at approximately 22 years, DC at 23 to 26 years, and HF and LF power in adolescence (15 to 20 years), whereas $\alpha 1$ followed an inverted U-shape with a maximum at 47 to 51 years and the long-term scaling exponent $\alpha 2$ continued to increase into the ninth decade. Men had higher LF power, LF/HF, and $\alpha 1$ from childhood onward, and women had higher $\alpha 2$, whereas the sex differences in heart rate, SDNN, HF power, and DC changed direction with age; for example, women had higher HF power only from about 43 years and men had higher DC only from about 15 years (Supplementary Table S7 and Figure S3). The adult differences are consistent with our previous report [31] and with published normative ranges [2,4], but a single age-adjusted sex difference would have concealed these reversals. These estimates should be interpreted as associations with age rather than as isolated measures of autonomic mechanisms. The LF/HF ratio does not index sympatho-vagal balance [5] and was reported only as a conventional descriptor. HF power reflects respiratory sinus arrhythmia, which depends on breathing frequency and pattern as well as on vagal tone [3]; age-related changes in respiratory rate and in the inhalation-to-exhalation ratio can therefore contribute to the observed HF trajectories [7,8], as can age-dependent differences in behavior during recording, such as sleep posture [15]. The explanatory framework makes such alternative explanations explicit and testable, because each can be introduced as an additional term with an interpretable coefficient.

4.3. Proposition 2: The Gain from ML in Prediction Is Small and Age-Specific

Estimating age from any single HRV index is an ill-posed inverse problem when the index changes non-monotonically; $\alpha 1$, for example, takes similar values in infancy and in the tenth decade. Combining indices with different trajectories resolves much of this ambiguity: heart rate declines almost monotonically, $\alpha 2$ increases until old age, and $\alpha 1$ follows an inverted U-shape, so that a weighted linear combination can separate age groups that no single index distinguishes. This explains why a linear predictor reached $R^2 = 0.71$ despite the strongly nonlinear univariate relationships. The residual advantage of tree-based ensembles was largest at the extremes of the age range (Figure 3): to childhood, when heart rate and spectral power change steeply and in opposite directions within a few years, and to old age, when several indices, such as HF and LF power, $\alpha 1$, and ARV, reverse

their direction of change. In mid-adulthood, where the trajectories are closer to monotonic, the linear model was as accurate as or more accurate than the ensembles. The gain was not an artifact of tuning: it persisted when the grids were expanded and the nested CV was repeated with different fold assignments (gradient boosting, 1.06 years; 95% CI 0.84 to 1.29; Table 7). We describe this gain as small in a relative sense—about 10% of the MAE of linear regression (a 10.6% reduction, from 11.25 to 10.06 years, in the primary analysis), with a residual error of about 10 years in every model—rather than as negligible; because no equivalence margin was specified in advance, whether a gain of this size matters depends on the intended application.

The predictive analysis also illustrates why explanatory and predictive models must be evaluated differently. Even after tuning, the tree-based ensembles fitted the training data with an MAE of 6.5 to 7.0 years but achieved about 10.1 years on held-out data; judged in-sample, as explanatory models commonly are, they would appear considerably more accurate than they are, because their training MAE was about 30% lower than their test MAE. Conversely, the explanatory spline models showed almost no gap between in-sample and cross-validated fit. All predictors also showed systematic regression toward the mean, overestimating the age of children by 7 to 10 years and underestimating that of individuals aged 80 years or older by 12 to 14 years. This bias is inherent to predicting a bounded target from noisy features, and it must be corrected before any difference between HRV-predicted and chronological age is interpreted as an index of autonomic aging. The age dependence of the error is therefore a central result rather than a detail: the youngest and oldest individuals, for whom the ensembles offered the largest gains, were also those predicted least accurately and with the largest systematic bias by every model, so that an overall MAE describes neither end of the age range.

Chronological age is known, and predicting it is not itself a clinical goal. In this study, age served as a benchmark target because it is measured without error and is strongly related to HRV, which allows the gain from model complexity to be quantified without the uncertainty of a clinical label (Section 2.4). The potential clinical interest of HRV-based age prediction lies instead in the difference between predicted and chronological age, which could serve as an index of accelerated or delayed autonomic aging, analogous to ECG-derived age estimated by deep learning. Our results indicate three prerequisites for such use: the age-dependent bias must be corrected without reintroducing chronological age into the index; the age gap must be interpreted relative to the age-specific error, which was largest in children and older adults; and its association with clinical outcomes must be demonstrated in prospective cohorts with external validation. Until these conditions are met, HRV-predicted age should be regarded as a methodological benchmark rather than a clinical biomarker.

The magnitude of R^2 also depends on the design of the sample. Because our sample was drawn uniformly across ages 0 to 102 years, the variance of age was larger than in the source database, which inflates R^2 even when the absolute error is unchanged. When the evaluation was reweighted to the age distribution of the database, R^2 fell to 0.25 to 0.35 for models trained on the uniform sample and to 0.47 to 0.51 for models retrained with the weights, whereas the MAE remained similar or decreased (8.9 to 10.8 years). Under this adult-dominated distribution, the gain of the best model over linear regression shrank to 0.3 years. R^2 values of age prediction should therefore not be compared across studies without regard to the age distribution of the samples, and MAE is the more transportable measure. These results complement recent deep-learning studies of ECG and Holter data [20–24], which have reported substantial gains from complex models for arrhythmia detection and outcome prediction. For age, most of the predictable information appears to be captured by a small set of conventional HRV indices, and the added value of model com-

plexity is modest and age-specific. Simple, well-specified linear models should therefore be reported as baselines when the predictive performance of ML models for HRV-based outcomes is evaluated.

4.4. Proposition 3: Variable Importance Is Not Physiological Importance

The comparison between permutation importance and the explanatory estimates shows why the variable importance of a predictive model should not be read as the physiological importance of its inputs. Importance and age dependence coincided only for heart rate, which was at the same time the index most strongly dependent on age and the least redundant with the others; in this case, the two criteria cannot be distinguished. For the remaining indices, importance answered a different question. Permutation importance measures how much a particular fitted function relies on a predictor, given the other predictors in the model, whereas the partial R^2 of age measures how strongly the index itself changes with age. Within groups of correlated indices, the predictive model needs only one member to carry the shared age information, and which member it uses is determined by the model rather than by physiology. Thus, DFA $\alpha 2$ carried most of the information shared with β and ARV, so that β, whose age dependence was substantial (partial R^2 0.29), received almost no importance; conversely, SDNN ranked high in several models despite a weak age dependence, presumably because it carries information not contained in the spectral indices that is useful for discriminating the extremes of the age range. Permutation importance is therefore a model-dependent quantity: when HRV indices are strongly correlated, as they typically are, the importance of an individual index depends on which other indices are in the model and on how the model uses them, so that the same index can appear essential in one model and redundant in another.

Two further observations reinforce this interpretation. First, importance was reproducible across folds within each model ($\rho \geq 0.94$) but not across models (ρ 0.53 to 0.88). Stability across resamples is therefore not evidence that an importance ranking reflects a property of the physiology; it may simply reflect a stable property of the fitted model. Second, for linear regression, the individual importances of ln LF and ln HF together exceeded 12 years of MAE, whereas permuting the whole ln VLF–ln LF–ln HF group increased the MAE by only 2.3 years. When predictors are strongly correlated, a linear model can rely on contrasts between them, and permuting one member alone creates combinations of values that do not occur in the data, so that individual importance overstates the information carried by the group. LF/HF illustrates the extreme case, because it is an exact function of ln LF and ln HF. When LF/HF was recomputed after permutation or omitted from the models, the importance of ln LF and ln HF changed substantially, and conditional permutation, which approximately preserves the relationships among predictors, reduced the importance of members of correlated groups to near zero and changed the rankings (Section 3.4). The shares in Figure 4B should therefore be read as relative magnitudes of marginal permutation scores, not as an additive partition of predictive information. Contrary to our expectation, importance was not significantly associated with non-redundancy across the 12 indices; rather than concentrating systematically on non-redundant indices, importance was allocated among correlated indices in model-specific ways.

We did not compute SHAP values, and our results do not directly evaluate them. SHAP values are, however, also defined relative to a fitted model [26], and they too must distribute credit among correlated inputs according to the model and the chosen reference distribution. Similar caution is therefore warranted when XAI methods such as LIME and SHAP [25,26] are applied to ECG-based models for arrhythmia detection [27–29]: their attributions describe how a predictor uses its inputs and are valuable for auditing predictive models, but they are not substitutes for explanatory models when the goal is physiological

inference. The explanatory spline models directly estimate the quantity of scientific interest, namely the expected value of each HRV index as a function of age and sex, together with its uncertainty.

4.5. Data Quality and Sensitivity Analyses

Only 61 recordings (1.6%) failed the quality criteria of DC > 4.5 ms and $\alpha 1 > 0.75$. Of these, only nine carried an annotation of paroxysmal AF; many of the remainder showed frequent supraventricular ectopy, and they were concentrated in children under 10 years and adults aged 80 years or older. Because DC and $\alpha 1$ are themselves analyzed, these thresholds could in principle remove genuine extreme values at the ends of the age range. Several observations argue against this as an explanation of our findings: the excluded recordings more often showed independent evidence of reduced quality (lower NN coverage and, in older adults, more supraventricular ectopic beats); a quality criterion based only on NN coverage, which does not depend on the HRV values, led to the same explanatory and predictive conclusions; and including the excluded recordings in the test set changed the MAE of every model by less than 0.1 years. Nevertheless, not every exclusion was accompanied by such evidence, and the comparisons of predictive models apply strictly to recordings that meet the quality criteria; for the excluded recordings, all models predicted age poorly and the ensembles offered no advantage. Ectopic beats perturb the RR-interval series in ways that are physiologically informative in some contexts, such as heart rate turbulence [13], but that distort indices such as DC and $\alpha 1$. The performance of automated AF detectors, whether based on RR-interval irregularity [16,17] or on deep learning [20], varies in long-term recordings, and rhythm labels in large clinical databases should be re-evaluated against current standards before they are used as outcomes. We note that stricter exclusion criteria based on thresholds for HRV indices themselves (SDNN and VLF power) removed 23% of individuals aged 80 years or older and altered the age curves at the upper end of the age range. Such criteria amount to selection on the dependent variable, and we therefore restricted the primary quality control to criteria reflecting measurement integrity. Recordings with a paroxysmal AF annotation that met the quality criteria (n = 338) were retained in the primary analysis because the annotation indicates that AF occurred at some time during the recording, not that the HRV indices were computed from AF; only NN intervals between sinus beats were analyzed. Excluding these recordings did not change the conclusions (Supplementary Tables S1 and S2 and Figure S1), nor did omitting LF/HF from the predictors (Supplementary Table S4). Similarly, AR-based indices depend on the robustness of model-order selection, which we have examined separately [32].

4.6. Limitations and Future Directions

This study has several limitations. First, and most importantly, the ALLSTAR database is not a population sample. Holter monitoring is ordered for a clinical reason, such as palpitations, syncope, suspected arrhythmia, or ischemia, and information on body mass index, medication, and comorbidity is not available. Although the recordings were selected to meet quality criteria, the age trajectories reported here describe a clinically referred Holter population across the lifespan and should not be interpreted as normative values for healthy individuals. In particular, the trajectories and the age-specific sex differences are not lifespan reference values, and the predictive performance reported here should not be taken as a reference for age estimation in other populations. Because the indications for Holter monitoring, the prevalence of cardiovascular disease, and medication use differ from those in healthy individuals, the absolute levels of HRV, the shape of the trajectories at the extremes of age, and the accuracy of age prediction may differ in healthy populations. The methodological conclusions—that nonlinearity can be modeled within the general

linear model, that the gain from ML must be evaluated against a well-specified linear baseline across the age range, and that importance depends on the model—do not rely on the representativeness of the sample, but their confirmation in population-based cohorts of healthy individuals is required. The proportion and type of clinical indications are also likely to differ between children and older adults, so that part of the apparent age dependence may reflect age-dependent referral patterns. Second, the design is cross-sectional, so that age effects cannot be separated from cohort effects. Third, the non-Gaussianity index λ25s used in our previous inclusion criteria was not available, and the proportion of NN intervals was used as a surrogate for the sinus-rhythm criterion. Fourth, predictive performance was evaluated by internal cross-validation within a single database; external validation, including in the shorter recordings typical of wearable devices [6], is required. Holter and remote ECG monitoring technologies are evolving rapidly, with patch-type, wearable, and implantable devices increasingly used for long-term monitoring [35], and the transferability of the present models to recordings from such devices remains to be examined. Fifth, the ML models should be regarded as the tested implementations within the grids examined; expanding the grids did not change the results, but other algorithms or more extensive searches could perform slightly differently. Sixth, the importance analysis was based on 12 indices, which limits the power of rank correlations, and on permutation importance; although derivation-consistent, grouped, and conditional permutation led to the same qualitative conclusions, other attribution methods, such as SHAP, could allocate importance among correlated indices differently, which would itself support our conclusion that importance depends on the method and the model. Finally, we deliberately restricted the determinants to age and sex. Sleep stages cannot currently be determined reliably from Holter data, and whether accelerometer signals from the recorders can support sleep staging must first be validated against polysomnography. Analyses involving AF likewise require a re-evaluation of the reliability of the AF detection software. Once these ground-truth problems are resolved, sleep stage and AF would provide further test cases for the explanatory–predictive distinction examined here.

5. Conclusions

Across 3854 quality-controlled 24 h Holter ECG recordings covering ages 0 to 102 years from a clinically referred Holter population, the three propositions of the study were supported. First, although the age dependence of all 12 HRV indices was strongly nonlinear, ML was not needed to explain it for the present explanatory question: general linear models with spline terms described the age trajectories as well as random forest and gradient boosting on cross-validation, and they alone provided parameters, confidence intervals, and tests. Second, in predicting age from HRV, nonlinear ML reduced the MAE of linear regression by only about one year (95% CI 0.8 to 1.3 years in repeated analyses), with the gain largest at the extremes of the age range, where all models had their largest errors and systematic bias, and largely absent under the age distribution of the source database. Third, the variable importance of the predictive models agreed only moderately with the explanatory age dependence, differed between models and permutation schemes, and was allocated within groups of correlated indices in model-specific ways; it should therefore not be read as the physiological importance of an HRV index. In this clinically referred, cross-sectional dataset, nonlinearity did not imply that ML was required for either explanation or prediction. Explanatory and predictive modeling should be distinguished explicitly in HRV research, simple, well-specified linear models should serve as the reference for both purposes, and variable importance should be interpreted as a property of the fitted model and of the way it is computed. HRV-predicted age should not be interpreted clinically until its age-dependent bias has been corrected and its prognostic value established. The

conceptual distinction between explanatory and predictive modeling is also relevant to the broader use of artificial intelligence in biomedical research.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162110592/s1: Table S1: Explanatory models of HRV indices after exclusion of recordings annotated as paroxysmal atrial fibrillation (AF); Table S2: Predictive models of age after exclusion of recordings annotated as paroxysmal AF; Table S3: Explanatory models under the age distribution of the source database Explanatory models under the age distribution of the source database; Table S4: Predictive models of age without the LF/HF ratio; Table S5: Hyperparameters selected by the inner cross-validation in the 10 outer folds; Table S6: Permutation importance of the HRV indices in the predictive models and absolute standardized coefficients of linear regression; Table S7: Age-specific sex differences (men − women) in the HRV indices estimated from the interacting spline models; Table S8: Robustness of permutation importance to the treatment of LF/HF and to the permutation scheme; Table S9: Characteristics of the recordings excluded by the quality criteria and analyses with a quality criterion independent of the HRV values; Table S10: Predictive performance and selected hyperparameters with expanded grids; Figure S1: Age curves of the 12 HRV indices estimated by restricted cubic spline models in the primary analysis and after exclusion of recordings annotated as paroxysmal AF; Figure S2: Age distribution of the study sample after quality control and the target distribution used for importance weighting; Figure S3: Age-specific sex differences (men − women) in the 12 HRV indices estimated from the interacting spline models; File S1: Analysis code.

Author Contributions: Conceptualization, E.Y.; methodology, E.Y.; software, E.Y.; validation, E.Y.; formal analysis, E.Y.; investigation, E.Y.; resources, J.H.; data curation, E.Y.; writing—original draft preparation, E.Y.; writing—review and editing, E.Y. and J.H.; visualization, E.Y.; project administration, E.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Ethics Review Committee of Nagoya City University Graduate School of Medical Sciences and Nagoya City University Hospital (No. 709 and No. 60160209).

Informed Consent Statement: Patient consent was waived because the study used anonymized data recorded for clinical purposes; information about the research was disclosed publicly, and patients were given the opportunity to opt out.

Data Availability Statement: The data presented in this study are available from the authors upon reasonable request (hayano@acm.org), subject to approval by the ALLSTAR project. The analysis code is provided as Supplementary File S1.

Acknowledgments: In the preparation of this manuscript, the authors received support from DeepL (paid version; DeepL SE, Cologne, Germany) as a tool to assist with English-language writing. Claude (version Opus 5.5; Anthropic, San Francisco, CA, USA) was used in the creation of the figures. The authors have reviewed and edited the generated content and accept full responsibility for the contents of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AF	atrial fibrillation
AIC	Akaike information criterion
ALLSTAR	Allostatic State Mapping by Ambulatory ECG Repository
AR	autoregressive
ARV	autoregressive variability

CI	confidence interval
CV	cross-validation
DC	deceleration capacity
DFA	detrended fluctuation analysis
ECG	electrocardiography
GBM	gradient boosting machine
HF	high frequency
HRV	heart rate variability
LF	low frequency
LIME	local interpretable model-agnostic explanations
MAE	mean absolute error
ML	machine learning
NN	normal-to-normal
PAF	paroxysmal atrial fibrillation
PRSA	phase-rectified signal averaging
QC	quality control
RF	random forest
RMSE	root-mean-square error
SDNN	standard deviation of normal-to-normal intervals
SHAP	Shapley additive explanations
ULF	ultra-low frequency
VLF	very low frequency
XAI	explainable artificial intelligence

References

1. Akselrod, S.; Gordon, D.; Ubel, F.A.; Shannon, D.C.; Berger, A.C.; Cohen, R.J. Power Spectrum Analysis of Heart Rate Fluctuation: A Quantitative Probe of Beat-to-Beat Cardiovascular Control. *Science* **1981**, *213*, 220–222. [CrossRef] [PubMed]
2. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart Rate Variability: Standards of Measurement, Physiological Interpretation and Clinical Use. *Circulation* **1996**, *93*, 1043–1065. [CrossRef]
3. Berntson, G.G.; Cacioppo, J.T.; Quigley, K.S. Respiratory Sinus Arrhythmia: Autonomic Origins, Physiological Mechanisms, and Psychophysiological Implications. *Psychophysiology* **1993**, *30*, 183–196. [CrossRef] [PubMed]
4. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef] [PubMed]
5. Billman, G.E. The LF/HF Ratio Does Not Accurately Measure Cardiac Sympatho-Vagal Balance. *Front. Physiol.* **2013**, *4*, 26. [CrossRef] [PubMed]
6. Yuda, E. Applications of Heart Rate Variability Metrics in Wearable Sensor Technologies: A Comprehensive Review. *Electronics* **2026**, *15*, 1707. [CrossRef]
7. Yuda, E.; Morikawa, N.; Hayano, J. HRV-Based Respiratory-Rate Estimation in Older Adults: Error Modeling and Pilot Validation for Wearable Sleep Monitoring. *Clocks Sleep* **2026**, *8*, 54. [CrossRef] [PubMed]
8. Yuda, E.; Yoshida, Y. Synchronization of Inhalation/Exhalation Ratio and Heart Rate Variability During Spontaneous Breathing. *Electronics* **2025**, *14*, 1903. [CrossRef]
9. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
10. Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
11. Shmueli, G.; Koppius, O.R. Predictive Analytics in Information Systems Research. *MIS Q.* **2011**, *35*, 553–572. [CrossRef]
12. Shmueli, G. To Explain, to Predict, or to Describe: Figuring out the Study Goal [Commentary on "On the Uses and Abuses of Regression Models" by Carlin and Moreno-Betancur]. *Stat. Med.* **2025**, *44*, e10307. [CrossRef] [PubMed]
13. Bauer, A.; Malik, M.; Schmidt, G.; Barthel, P.; Bonnemeier, H.; Cygankiewicz, I.; Guzik, P.; Lombardi, F.; Müller, A.; Oto, A.; et al. Heart Rate Turbulence: Standards of Measurement, Physiological Interpretation, and Clinical Use: International Society for Holter and Noninvasive Electrophysiology Consensus. *J. Am. Coll. Cardiol.* **2008**, *52*, 1353–1365. [CrossRef] [PubMed]
14. Verrier, R.L.; Klingenheben, T.; Malik, M.; El-Sherif, N.; Exner, D.V.; Hohnloser, S.H.; Ikeda, T.; Martínez, J.P.; Narayan, S.M.; Nearing, B.D.; et al. Microvolt T-Wave Alternans: Physiological Basis, Methods of Measurement, and Clinical Utility—Consensus Guideline by International Society for Holter and Noninvasive Electrocardiology. *J. Am. Coll. Cardiol.* **2011**, *58*, 1309–1324. [CrossRef] [PubMed]

15. Yuda, E.; Hayano, J. Sleep Posture and Autonomic Nervous System Activity Across Age and Sex in a Clinical Cohort: Analysis of a Nationwide Ambulatory ECG Database. *Sensors* **2025**, *25*, 5982. [CrossRef] [PubMed]
16. Tateno, K.; Glass, L. Automatic Detection of Atrial Fibrillation Using the Coefficient of Variation and Density Histograms of RR and ΔRR Intervals. *Med. Biol. Eng. Comput.* **2001**, *39*, 664–671. [CrossRef] [PubMed]
17. Petrėnas, A.; Marozas, V.; Sörnmo, L. Low-Complexity Detection of Atrial Fibrillation in Continuous Long-Term Monitoring. *Comput. Biol. Med.* **2015**, *65*, 184–191. [CrossRef] [PubMed]
18. Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 664–675. [CrossRef] [PubMed]
19. Acharya, U.R.; Oh, S.L.; Hagiwara, Y.; Tan, J.H.; Adam, M.; Gertych, A.; Tan, R.S. A Deep Convolutional Neural Network Model to Classify Heartbeats. *Comput. Biol. Med.* **2017**, *89*, 389–396. [CrossRef] [PubMed]
20. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nat. Med.* **2019**, *25*, 65–69. [CrossRef] [PubMed]
21. Attia, Z.I.; Noseworthy, P.A.; Lopez-Jimenez, F.; Asirvatham, S.J.; Deshmukh, A.J.; Gersh, B.J.; Carter, R.E.; Yao, X.; Rabinstein, A.A.; Erickson, B.J.; et al. An Artificial Intelligence-Enabled ECG Algorithm for the Identification of Patients with Atrial Fibrillation During Sinus Rhythm. *Lancet* **2019**, *394*, 861–867. [CrossRef] [PubMed]
22. Grégoire, J.M.; Gilon, C.; Marelli, F.; Bersini, H.; Groben, L.; Nguyen, T.; Deruyter, B.; Godart, P.; Carlier, S. Short-Term Atrial Fibrillation Onset Prediction Using Machine Learning. *Eur. Heart J. Digit. Health* **2025**, *6*, 1159–1168. [CrossRef] [PubMed]
23. Kim, J.Y.; Kim, K.G.; Joo, S.; Chang, M.; Kim, J.; Park, K.M.; On, Y.K.; Kim, J.S.; Lee, Y.S.; Park, S.J.; et al. Deep Learning-Based Risk Assessment and Prediction of Cardiac Outcomes Using Single-Lead 24-Hour Holter-ECG in Patients with Heart Failure or Myocardial Infarction. *J. Clin. Med.* **2025**, *14*, 7209. [CrossRef] [PubMed]
24. Wu, Y.; Chen, Y.; Zhang, B.; Chen, X.; Peng, C.; Jiang, J.; Chen, J.; Jian, C.; Wu, G. Deep Learning-Based Early Prediction of Life-Threatening Ventricular Arrhythmias Using Long-Term Holter ECG Signals. *Front. Cardiovasc. Med.* **2026**, *13*, 1714589. [CrossRef] [PubMed]
25. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
26. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
27. Taniguchi, H.; Takata, T.; Takechi, M.; Furukawa, A.; Iwasawa, J.; Kawamura, A.; Taniguchi, T.; Tamura, Y. Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms. *Int. Heart J.* **2021**, *62*, 534–539. [CrossRef] [PubMed]
28. Kilic, M.E.; Tufekcioglu, O.A.; Yilancioglu, Y.R.; Turan, O.E.; Ozcan, E.E. Explainable AI for Specific Arrhythmia Detection: SHAP-Based Insights from Multi-Lead ECG Data. *J. Med. Biol. Eng.* **2025**, *45*, 314–324. [CrossRef]
29. Kovalchuk, O.; Barmak, O.; Radiuk, P.; Klymenko, L.; Krak, I. Towards Transparent AI in Medicine: ECG-Based Arrhythmia Detection with Explainable Deep Learning. *Technologies* **2025**, *13*, 34. [CrossRef]
30. Umetani, K.; Singer, D.H.; McCraty, R.; Atkinson, M. Twenty-Four Hour Time Domain Heart Rate Variability and Heart Rate: Relations to Age and Gender Over Nine Decades. *J. Am. Coll. Cardiol.* **1998**, *31*, 593–601. [CrossRef] [PubMed]
31. Hayano, J.; Ohashi, K.; Yoshida, Y.; Yuda, E.; Nakamura, T.; Kiyono, K.; Yamamoto, Y. Increase in Random Component of Heart Rate Variability Coinciding with Developmental and Degenerative Stages of Life. *Physiol. Meas.* **2018**, *39*, 054004. [CrossRef] [PubMed]
32. Yuda, E.; Kaneko, I.; Hirahara, D.; Hayano, J. Robustness of AIC-Based AR Order Selection in HRV Analysis. *Electronics* **2026**, *15*, 1319. [CrossRef]
33. ALLSTAR Project. Available online: https://allstar.jpn.org/ (accessed on 29 September 2026).
34. Peng, C.-K.; Havlin, S.; Stanley, H.E.; Goldberger, A.L. Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series. *Chaos* **1995**, *5*, 82–87. [CrossRef] [PubMed]
35. Preda, A.; Falco, R.; Tognola, C.; Carbonaro, M.; Vargiu, S.; Gallazzi, M.; Baroni, M.; Gigli, L.; Varrenti, M.; Colombo, G.; et al. Contemporary Advances in Cardiac Remote Monitoring: A Comprehensive, Updated Mini-Review. *Medicina* **2024**, *60*, 819. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.