Research Square

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

Development and internal evaluation of a model combining ECG waveforms with machine reports to identify recorded right ventricular systolic dysfunction in adults

Junze Li[1]
The First Affiliated Hospital with Nanjing Medical University
Yongfeng Shao[1]
shaoyongfeng@njmu.edu.cn

The First Affiliated Hospital with Nanjing Medical University

Research Article

Keywords: electrocardiography, right ventricular function, machine learning, incremental prediction, internal validation, calibration

Posted Date: November 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11143146/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

ECG waveform models can identify structural and functional cardiac abnormalities. Their incremental value beyond jointly modeled machine measurements and interpretation text requires evaluation using the same patients and reference labels.

Methods

We linked MIMIC-IV-ECG v1.0 with MIMIC-IV-Echo v1.0.1 from a single hospital and selected one index ECG–echocardiography pair per eligible adult. The primary training cohort comprised 25,315 patients, including 3,165 endpoint-positive patients. The primary endpoint was recorded global or unspecified reduction in right ventricular systolic function. A selected machine-feature model (B; XGBoost) was compared with the same configuration refitted with an additional convolutional-network waveform score (BW). Development, calibration and threshold selection were separated from internal testing. The primary comparison was test specificity at model-specific thresholds selected for a 90% validation sensitivity target. Actual test sensitivities were reported separately. Nominal 95% confidence intervals were obtained from 2,000 paired patient-bootstrap resamples of the frozen systems.

Results

The primary test cohort included 5,505 patients, of whom 703 were endpoint-positive. The area under the receiver operating characteristic curve (AUROC) increased from 0.782 (95% CI 0.764–0.798) to 0.850 (95% CI 0.835–0.865), a difference of 0.069 (95% CI 0.057–0.080). Specificity increased from 40.19% to 49.77%, a difference of 9.58 percentage points (95% CI 8.40–10.79), corresponding to 460 fewer false positives (95% CI 403–518). Actual sensitivity was 92.18% versus 91.89%, a difference of − 0.28 percentage points (95% CI − 2.29–1.63). Sensitivity was lower in the operationally complete-measurement subset, and a secondary severity-label scenario showed a sensitivity disadvantage.

Conclusions

Adding a waveform score improved discrimination and reduced false positives within this frozen internal evaluation. The findings do not establish sensitivity noninferiority, clinical utility or external applicability.

Registration:

OSF grs6q; stage-specific registration.

Background

Automated ECG systems provide numerical measurements and interpretation statements derived from the recorded waveform. [1, 2] These outputs offer an accessible reference representation when assessing whether a waveform model adds predictive information. A comparison against measurements alone or a few selected diagnostic statements may leave information in the routine machine output unused. An incremental evaluation should therefore compare models using the same patients and reference labels, with the machine measurements and interpretation text jointly represented.

Right ventricular systolic function is routinely assessed by echocardiography. [3] ECG models have been developed to identify structural heart disease, including right ventricular abnormalities defined by cardiac magnetic resonance measurements or qualitative echocardiographic labels. [4, 5, 6] In retrospective data, a model trained against a structured interpretation predicts the recorded finding. It does not automatically reproduce an independently adjudicated physiological measurement.

Several studies inform the choice of comparator. Rogers et al. compared waveforms with automated measurements and physician-interpretation keywords for left ventricular wall-motion abnormalities, while Joshi et al. compared engineered ECG features with a waveform network for that target. [7, 8] Do et al. evaluated waveform models, engineered features and covariate fusion on EchoNext, including a right ventricular endpoint. [9] These studies support evaluating the contribution of different ECG representations. They do not establish the increment over the particular combination of machine measurements, missingness indicators and machine-generated statements examined here. Moreover, improvement in ranking metrics does not specify how false positives and sensitivity change at a threshold selected before testing.

We developed and internally evaluated a machine-feature model and a waveform-augmented model for identifying recorded right ventricular systolic dysfunction in adults with linked ECG and echocardiography. The question was whether a waveform score improved prediction beyond jointly modeled machine measurements and interpretation statements. The primary comparison was the paired change in specificity at model-specific thresholds selected for a 90% validation sensitivity target. We also evaluated actual test sensitivity, discrimination, probability error, calibration and prespecified sensitivity analyses. This was a retrospective comparison of information representations, without a defined deployment workflow or a prespecified minimum clinically useful gain.

Methods

Study design and data sources

This retrospective prediction-model study used MIMIC-IV-ECG v1.0, MIMIC-IV-Echo v1.0.1 and the MIMIC-IV Clinical v3.1 patient table for age eligibility and subsequent descriptive reporting. [1, 10, 11] MIMIC-IV originates from Beth Israel Deaconess Medical Center. [12] The released ECG resource covers

2008–2019 and the Echo structured measurements cover 2008–2022. [1, 10] These are source-resource coverage periods; actual calendar dates for the selected patients were not reconstructed from shifted timestamps. Data linkage used the shared deidentified patient identifier. The analysis addressed the classification of an adjacent echocardiographic record, not incident disease, subsequent mortality or a treatment effect. Clinical timestamps were treated as operational linkage data; agreement between released files was not assumed to verify the ECG device clock.

A stage-specific protocol was registered on OSF after inspection of public ECG metadata and selected waveform headers, with echocardiographic phenotype mapping and detailed model implementation still pending. [13] Subsequent dated local supplements specified labels, pairing, quality control, training and evaluation at their respective development stages. These documents are not collectively described as a fully specified outcome-blind preregistration. The model configuration was frozen before the first fit. Its use of L2 logistic regression candidates replaced the sparse-regularization constraint in the Stage 1 plan; the deviation was explicitly documented after training had begun and before formal test prediction. No contemporaneous rationale establishing that replacement as conservative was recorded. Supplementary Methods detail this chronology and the information available at each stage. Reporting was organized with reference to TRIPOD + AI. [14]

Index-pair construction and population

We selected resting transthoracic echocardiograms with internally consistent patient, examination-time and test-type metadata. Adult eligibility used the Clinical patient-table age anchor plus the difference between the examination year and anchor year. For top-coded ages, we used a conservative lower bound rather than treating the code as an exact age. Missing or invalid age information did not establish eligibility. Supplementary Methods specify the age rules.

For each patient, we selected the earliest eligible echocardiogram with an ECG recorded in the preceding 0–168 hours, inclusive. We paired it with the nearest preceding ECG, using record_list.ecg_time. Ties were resolved using the smaller ECG study_id or echo measurement_id, as applicable. The identifiers of the two examination types were not treated as interchangeable. Labels were joined only after index-pair selection.

Patient assignment to approximately 70% training, 15% validation and 15% test partitions used a deterministic patient-identifier hash without label stratification and was fixed before model development. The same assignment was maintained across endpoints and sensitivity analyses. Unknown primary labels were not changed to negative, and an index examination was not replaced by a later examination to obtain a usable label. Quality-control failure likewise did not trigger replacement of the selected index pair.

Study size and descriptive reporting

The analysis used all patients satisfying the frozen index-pair, age, technical and reference-label rules in the specified dataset releases. Cohort size was determined by data availability and eligibility, rather than recruitment to a calculated target. Feasibility counts were inspected before model development; no prospective power or precision target was used to determine the final sample size. The test cohort was used to estimate performance and uncertainty under the executed procedure, not to establish a prespecified clinically useful effect.

Demographic and data-availability summaries were prepared after the frozen test evaluation for manuscript reporting. They describe the existing cohort without changing eligibility, labels, partitions, fitted models or thresholds. Approximate age was summarized separately from top-coded age, and the source patient-table gender field was reported without inferring gender identity. These summaries were not used for model selection. No subgroup performance or fairness evaluation was performed.

Reference outcomes

The reference field was rv_function (description: Right Ventricle - Function) in resting TTE records, with an empty unit field. The primary label was positive for mild, moderate or severe global hypokinesis or reduced function without a severity grade. Normal, low-normal and hyperdynamic categories were negative under this definition. Regional-only findings, including apical free-wall or basal hypokinesis, were unknown rather than negative or globally abnormal. Unassessable, missing, unrecognized or metadata-incompatible values were also unknown.

Repeated identical categories were collapsed. When an examination contained several different categories, any regional-only, unassessable or unrecognized nonempty category made its strict label unknown. Otherwise, all retained categories had to agree on the primary binary label. The secondary label additionally required all categories to have known and agreeing severity mappings. Category mappings and validity rules are provided in Table S1 in Additional file 1.

The primary endpoint included documented reduction whose severity was not specified. The secondary endpoint was explicitly graded moderate or severe global reduction. A reduced-function record without a usable severity grade was excluded from the observed secondary-endpoint analysis. These outcomes were obtained from structured clinical records, without an independent blinded echocardiographic reread for this study. The analysis did not ascertain the identity or number of original readers, inter-reader agreement, or whether the original echocardiographic interpretation was blinded to ECG findings. They do not establish agreement with an independently adjudicated physiological gold standard. Quantitative right ventricular measurements were not substituted for missing qualitative labels.

Common waveform quality control

Technical checks required readable paired waveform files, consistent header identity, the expected 12 unique leads, 500-Hz sampling and 5,000 samples. The implemented reader supported format 16, one sample per frame and zero skew. Unsupported formats or nonzero skew were excluded as unsupported,

rather than interpreted as physiological corruption. Header gain, voltage unit and baseline handling had to permit conversion to millivolts. File length, digital missing values and finiteness of converted signals were checked. Available checksums were verified; an absent checksum was flagged rather than treated as a successful verification. A recording with all leads constant failed; a constant individual lead was flagged. Network or download problems were tracked as pending rather than assigned waveform failure. These checks establish compliance with the implemented input requirements, not comprehensive physiological signal quality.

Both models were evaluated in the same quality-control-passing population, including patients with operationally unavailable machine measurements. Complete-case restriction was not used for the primary comparison. This common inclusion rule avoids different model denominators but limits the analysis to ECGs accepted by the implemented waveform checks.

Machine measurements and interpretation text

The machine-feature input comprised nine released fields: rr_interval, p_onset, p_end, qrs_onset, qrs_end, t_end, p_axis, qrs_axis and t_axis. The documented units were milliseconds for the first six fields and degrees for the three axes. [2] These were treated as raw machine measurements under an operational usability codebook. We did not derive PR, QRS or QT intervals, QTc or heart rate, because coordinate conventions and all physiological validity rules were not independently established.

Five prespecified numerical codes (29999, 32767, 65534, 65535 and − 32768), blank or unparseable values, nonfinite values and nonpositive RR intervals were excluded from the continuous channels. Their field-specific missingness, code and reason indicators were retained. Continuous values were imputed with medians and standardized with means and population standard deviations estimated from the relevant training subset. All-unavailable continuous channels and constant features were removed using that subset alone. Operational unavailability was not interpreted as a specific rhythm or equipment mechanism.

All nonempty segments from report_0 through report_17 were eligible as whole-statement binary features. Normalization was limited to trimming, whitespace consolidation and lowercase conversion; negation and uncertainty were retained. A statement had to occur in at least five training patients to enter the vocabulary. Repeated statements within one ECG contributed only once. Preprocessing and vocabulary fitting were repeated within each internal training fold, and the final processor was fitted using the complete corresponding training partition. Validation and test data did not expand the vocabulary. Patient identifiers, dates, device identifiers, age, sex and comorbidities were not model inputs. Including all report segments did not imply that every rare statement was retained after the training-frequency filter.

Baseline selection and waveform augmentation

Six machine-feature configurations were compared in three patient-level internal folds: L2 logistic regression with C values of 0.01, 0.1 and 1, and XGBoost models with maximum depths of 2, 4 and 6. Selection used the lowest pooled out-of-fold, unweighted log-loss. A prespecified ordering resolved differences within 10^{-6} of the minimum. All candidate configurations had to complete; the procedure did not select among a reduced set of successful candidates. Configuration details and fold construction are given in Supplementary Methods. This bounded comparison does not establish that the baseline attained a performance ceiling.

Waveforms were decoded in millivolts from the accepted 12-lead, 500-Hz, 5,000-sample recordings. Fixed polyphase resampling reduced them to 250 Hz and 2,500 samples, followed by per-record, per-lead median centering. The float32 model input repeated median centering identically in training and inference. No amplitude normalization, additional filtering, augmentation or external pretraining was used. A fixed residual one-dimensional convolutional network with 1,022,145 parameters was trained for 25 epochs with each of three prespecified seeds. Supplementary Methods give its architecture, optimizer and training settings.

For every training patient, three CNNs trained without that patient's internal fold generated out-of-fold probabilities. The three probabilities were averaged, clipped to $[10^{-6}, 1 - 10^{-6}]$, transformed to a logit and standardized using the training out-of-fold distribution. BW added this single waveform score to all columns from the final B processor. It used the same selected B model family and hyperparameters and was refitted without a further augmentation-specific search. Three CNNs were then refitted on the complete training partition to generate the validation and test waveform score. Thus, the CNNs generating the training stacking input and final held-out stacking input had different training-set sizes. Models were trained separately for the primary and secondary endpoints.

Validation calibration and fixed working points

For each endpoint, branch and model, an unpenalized two-parameter logistic calibration map was fitted in the corresponding validation population. A usable mapping required finite coefficients and a positive slope. No alternative calibrator was substituted after a failure. In the same validation population, the largest observed positive-patient score threshold attaining at least the target sensitivity was selected, using the required positive count rounded upward. A prediction at or above the threshold was positive. The targets were 90% and 95% for both reference endpoints; the 90% target for the primary endpoint defined the main comparison.

All fitted models, processors, calibration maps and thresholds were frozen before the single authorized internal test evaluation. Thresholds were not adjusted to obtain matching sensitivities in the test sample. Test-label counts and data structure had previously been inspected for feasibility; the test cohort was held out from model selection, not claimed to have been wholly inaccessible throughout preparation.

Outcomes and statistical evaluation

The primary performance comparison was the paired difference in test specificity, BW minus B, at their respective 90% validation-target thresholds. Net fewer false positives was the corresponding count expression among the same reference-negative patients, not a separate independent finding. Actual sensitivity and its paired difference were reported, together with newly detected and lost reference-positive patients. We also calculated AUROC, average precision, Brier score and predictive values. Average precision was not replaced by trapezoidal integration of the precision–recall curve.

Uncertainty was estimated with 2,000 ordinary, nonstratified paired patient-bootstrap resamples. Each resample applied the same patient draw to B and BW while leaving the prediction systems, validation calibration and thresholds fixed. Nominal 95% percentile confidence intervals were reported without multiplicity adjustment. Under the implemented rule, an interval was withheld if any of its required resample statistics was undefined; undefined draws were not replaced. These intervals do not incorporate uncertainty from training, model selection, calibration-map estimation, threshold selection, label validity or transport to another population.

Calibration-in-the-large was estimated using an offset logistic model with slope fixed at one. A separate diagnostic model jointly estimated the calibration intercept and slope. Both diagnostics used frozen test probabilities and did not update predictions. The original calibration display was the specified logit-linear diagnostic curve with pointwise percentile intervals, not a flexible calibration smooth. A fixed-bin fallback was defined for failure but was not triggered. Ideal-value confidence-interval coverage was not interpreted as proof of adequate calibration.

Sensitivity analyses

S1 retained original index records with concordant timestamps between the index and measurement tables, excluding conflicts and indeterminate comparisons. S2 restricted the original ECG–echo interval to no more than 24 hours without rematching. Each was evaluated both with the fixed main system and with a system developed and calibrated within the restricted development population. S3 removed the p_end continuous channel and its associated indicators and repeated the frozen development procedure. The complete9 analysis applied the fixed main system to records meeting all nine operational measurement-usability rules.

The secondary L and H scenarios included primary-positive patients with unspecified severity in the otherwise observed secondary-label population. L assigned these patients to the non-moderate/severe category, while H assigned them to the moderate/severe category. The previously frozen secondary models, probabilities, calibration maps and thresholds were unchanged. These scenarios represent label assumptions rather than factual imputations, clinical adjudications or mathematical performance bounds. No clinical minimum useful gain, noninferiority margin or equivalence margin was specified.

Use of AI assisted tools

The authors did not use generative artificial intelligence or AI-assisted technologies in the writing process, and take full responsibility for the content of this article.

Results

Cohort and fitted systems

The TTE input comprised 179,928 examinations from 85,292 patients. Of these, 61,928 examinations met adult eligibility and had an ECG within the operational preceding window. Selection of one index pair per person yielded 39,890 patients. Their pre-quality-control pairing interval had a median of 18.72 hours (IQR 5.83–40.83); this distribution does not validate clock accuracy.

Waveform checks passed for 39,287 patients and failed for 603. After these checks, the primary label was known for 36,262 patients, including 4,550 positives; 3,025 remained unknown. The secondary label was known for 35,259 patients, including 1,625 positives. A total of 1,003 primary-positive patients had unspecified severity. Partition counts are reported in Table 1.

Table 1
Cohort and label availability by patient partition

Population after waveform checks	Training	Validation	Test	Total
All index patients	27460	5889	5938	39287
Known primary label	25315	5442	5505	36262
Primary positive	3165	682	703	4550
Primary negative	22150	4760	4802	31712
Primary unknown	2145	447	433	3025
Known secondary label	24628	5297	5334	35259
Secondary positive	1121	241	263	1625
Secondary negative	23507	5056	5071	33634
Primary positive, severity unspecified	687	145	171	1003

The primary-label-known population included 17023 patients (46.9%) recorded as female in the source gender field. The estimated age among non-topcoded patients was 67 years (IQR 56–78); 1869 patients (5.2%) had topcoded age anchors. Table 2 summarizes characteristics by partition. Additional file 1 reports the full descriptive table and the 3025 quality-control-passing patients with unknown primary labels.

Table 2
Patient characteristics and measurement availability by partition

Characteristic	Training (n = 25315)	Validation (n = 5442)	Test (n = 5505)	Overall (n = 36262)
Primary endpoint positive, n (%)	3165 (12.5)	682 (12.5)	703 (12.8)	4550 (12.5)
Female in source gender field, n (%)	11780 (46.5)	2579 (47.4)	2664 (48.4)	17023 (46.9)
Male in source gender field, n (%)	13535 (53.5)	2863 (52.6)	2841 (51.6)	19239 (53.1)
Estimated age among non-topcoded patients, years, median [IQR]	67.0 [56.0, 78.0]	67.0 [56.0, 78.0]	68.0 [56.0, 78.0]	67.0 [56.0, 78.0]
Age topcoded at anchor, n (%)	1307 (5.2)	274 (5.0)	288 (5.2)	1869 (5.2)
ECG-to-echocardiography interval, hours, median [IQR]	19.2 [6.1, 41.5]	18.8 [6.2, 41.7]	18.6 [5.7, 40.9]	19.0 [6.0, 41.4]
ECG within 24 hours before echocardiography, n (%)	15202 (60.1)	3324 (61.1)	3321 (60.3)	21847 (60.2)
Any prespecified sentinel in nine machine measurements, n (%)	8096 (32.0)	1717 (31.6)	1779 (32.3)	11592 (32.0)
All nine machine measurements operationally usable, n (%)	17219 (68.0)	3725 (68.4)	3726 (67.7)	24670 (68.0)

Values are counts (percentages) or medians [interquartile ranges]. Age is estimated at year level; patients with top-coded age anchors are excluded from age quantiles and reported separately. F and M are the source patient-table gender codes. Pairing intervals use the operational ECG timestamp and do not establish clock accuracy. Measurement availability follows the frozen nine-field rules. Descriptive summaries were prepared after test evaluation without changing the cohort or fitted models. Complete descriptive results and an unknown-label comparison are provided in Additional file 1, Section S10.

For the main primary and secondary endpoints, the baseline selection procedure chose XGB_D4. The final primary model used 731 machine-feature columns, and the augmented model used those columns plus the waveform score. The corresponding secondary models used 719 and 720 columns. Across eight planned branch-by-endpoint units, 256 model fits and 16 validation calibrations completed, with one recorded technical retry. The frozen evaluation completed all planned units without a calibration-map failure. Execution success is not itself evidence of predictive or clinical performance.

Main working-point comparison

At the primary endpoint's 90% validation-target working point, specificity increased from 40.19% with B to 49.77% with BW, a paired difference of 9.58 percentage points (95% CI 8.40–10.79). False positives

decreased from 2,872 to 2,412, yielding 460 fewer false positives (95% CI 403–518). This comprised 704 avoided false positives and 244 newly introduced false positives among the same reference-negative patients.

Actual test sensitivity was 92.18% for B and 91.89% for BW, a difference of − 0.28 percentage points (95% CI − 2.29–1.63). BW newly identified 22 reference-positive patients and no longer identified 24 who were identified by B. The sensitivity interval does not establish noninferiority or equivalence. Positive predictive value increased from 18.41% to 21.12%; most positive predictions still did not correspond to the primary reference label. Results for the 95% target and the secondary endpoint are shown in Table 3 and Fig. 1.

Table 3
Main test working-point comparisons

Endpoint	Validation target	B sens. % (95% CI)	BW sens. % (95% CI)	Sens. difference pp (95% CI)	Spec. difference pp (95% CI)	Net fewer false positives (95% CI)
Primary	90%	92.18 (90.30, 94.02)	91.89 (89.87, 93.78)	-0.28 (-2.29, 1.63)	9.58 (8.40, 10.79)	460 (403, 518)
Primary	95%	96.30 (94.93, 97.62)	95.59 (94.07, 96.96)	-0.71 (-2.07, 0.67)	12.04 (10.88, 13.18)	578 (523, 633)
Secondary	90%	93.92 (90.91, 96.67)	95.06 (92.31, 97.60)	1.14 (-1.87, 4.15)	16.60 (15.35, 17.95)	842 (779, 910)
Secondary	95%	97.34 (95.24, 98.92)	96.20 (93.67, 98.41)	-1.14 (-3.66, 1.19)	24.79 (23.45, 26.08)	1257 (1189, 1322)

The primary cohort comprised 5,505 patients, including 703 endpoint-positive patients; the secondary cohort comprised 5,334 patients, including 263 endpoint-positive patients. Sens., sensitivity; spec., specificity; pp, percentage points. Differences are BW minus B and are expressed in percentage points for rates. Net fewer false positives is B minus BW. All intervals are nominal 95% confidence intervals. Validation targets do not guarantee either equal or target-level test sensitivity.

Discrimination and probability error

Primary-endpoint AUROC increased from 0.782 (95% CI 0.764–0.798) to 0.850 (95% CI 0.835–0.865), a difference of 0.069 (95% CI 0.057–0.080). Average precision increased from 0.354 (95% CI 0.321–0.392) to 0.514 (95% CI 0.475–0.555). Brier score decreased from 0.097 to 0.082; the paired difference, calculated before rounding, was − 0.014 (95% CI − 0.017 to − 0.011). The secondary-endpoint AUROC increased from 0.823 to 0.894 (Table 4).

Table 4
Discrimination and probability error in the main test cohorts

Endpoint	Measure	B (95% CI)	BW (95% CI)	BW minus B (95% CI)
Primary	AUROC	0.782 (0.764, 0.798)	0.850 (0.835, 0.865)	0.069 (0.057, 0.080)
Primary	Average precision	0.354 (0.321, 0.392)	0.514 (0.475, 0.555)	0.160 (0.130, 0.189)
Primary	Brier score	0.097 (0.091, 0.102)	0.082 (0.078, 0.088)	-0.014 (-0.017, -0.011)
Secondary	AUROC	0.823 (0.798, 0.846)	0.894 (0.875, 0.911)	0.071 (0.054, 0.090)
Secondary	Average precision	0.214 (0.174, 0.262)	0.385 (0.325, 0.451)	0.170 (0.122, 0.222)
Secondary	Brier score	0.043 (0.038, 0.047)	0.037 (0.033, 0.041)	-0.006 (-0.007, -0.004)

AUROC and average precision use continuous scores and do not depend on a working-point threshold. Lower Brier scores indicate lower squared probability error, not proof of adequate calibration. Differences were calculated from unrounded estimates, so subtraction of displayed values may differ slightly.

Calibration and sensitivity analyses

In the primary cohort, calibration-in-the-large was 0.009 for B and − 0.002 for BW. Joint calibration slopes were 1.009 and 1.041, respectively (Table 5; Fig. 3). These summary diagnostics were close to their ideal values but cannot exclude nonlinear miscalibration. The secondary BW joint intercept was 0.263 (95% CI 0.023–0.527), while its calibration-in-the-large interval included zero; the two intercepts arose from different diagnostic models.

Table 5
Calibration diagnostics in the main test cohorts

Endpoint	Model	CITL (95% CI)	Joint intercept (95% CI)	Joint slope (95% CI)
Primary	B	0.009 (-0.076, 0.098)	0.024 (-0.129, 0.188)	1.009 (0.932, 1.093)
Primary	BW	-0.002 (-0.091, 0.089)	0.054 (-0.078, 0.188)	1.041 (0.972, 1.115)
Secondary	B	0.066 (-0.070, 0.194)	0.246 (-0.072, 0.578)	1.073 (0.961, 1.199)
Secondary	BW	0.066 (-0.072, 0.197)	0.263 (0.023, 0.527)	1.096 (0.998, 1.210)

CITL denotes calibration-in-the-large, estimated with slope fixed at one. The joint intercept and slope are from a separate regression. Diagnostic coefficients were not applied to update test predictions.

The primary-endpoint specificity difference remained positive in the prespecified analyses shown in Fig. 2, but the populations and development procedures differed. In the complete9 subset, 3,726 patients included 339 primary-positive patients. At the main system's 90% validation-target thresholds, actual sensitivity was only 86.43% for B and 86.14% for BW. Restricting to operationally usable measurements therefore did not preserve the main cohort's sensitivity level.

The secondary-label analyses added 171 primary-positive patients whose severity was unspecified, expanding the test population from 5,334 to 5,505. Under H, the positive count was 434. At the 95% validation-target thresholds, sensitivity was 95.16% for B and 91.94% for BW, a difference of − 3.23 percentage points (95% CI − 5.84 to − 0.90). This unfavorable result is retained alongside the specificity improvement. All 32 working-point rows and 32 model-by-case calibration summaries are included in the supplement, retaining their original analysis roles.

Discussion

In this frozen internal evaluation, adding an ECG waveform score to jointly modeled machine measurements and interpretation statements improved discrimination for recorded right ventricular systolic dysfunction. At the primary validation-selected working point, specificity increased by 9.58 percentage points and the paired false-positive count decreased by 460. Actual test sensitivity was slightly lower, with an interval spanning both decreases and increases. The finding is therefore an incremental prediction result under the executed procedure, rather than a demonstration of clinical benefit or preserved sensitivity.

Relation to prior ECG representation studies

Prior studies address closely related comparisons. Carbonati et al. evaluated MIMIC ECG models for qualitative global right ventricular hypokinesis, using age, sex and cardiologist-report findings in a reference model. [6] The interpretation source differs from the machine-generated statements used here. Do et al. tested waveform–covariate fusion on EchoNext, so combination of inputs is not itself the distinction. [9] Conventional machine-reported right-axis deviation and right ventricular hypertrophy have also been studied against right-heart echocardiographic phenotypes. [15] Our comparison does not establish a first use of fusion or a first ECG model for right ventricular dysfunction. Its focus is the increment over a defined machine-feature baseline that jointly uses available numerical fields, missingness information and interpretation statements. Both models shared the patients, labels, index rule and main feature representation, and the augmented model used the selected baseline configuration without a separate configuration search. Nevertheless, only six baseline configurations and one waveform architecture were evaluated. The results depend on these choices and do not establish an optimal baseline or an architecture-independent amount of additional waveform information.

Discrimination and the operating-point tradeoff

The AUROC and average-precision improvements describe changes in continuous ranking, whereas the specificity result describes a threshold-specific comparison. A validation sensitivity target determines how a threshold is selected, not the sensitivity that must subsequently occur in another sample. The observed loss of two net reference-positive identifications at the primary working point cannot be dismissed because its interval crosses zero. Equivalence and noninferiority would require a justified margin and an appropriate analysis specified for those aims. Likewise, the increase in positive predictive value to 21.12% does not demonstrate that the resulting referral or review workload would be acceptable.

No minimum clinically useful gain was fixed, and it would be inappropriate to define one from the observed effect and then claim it had been met. The study did not evaluate clinician behavior, downstream echocardiography use, treatment decisions, resource consumption or patient outcomes. Those questions would require an explicitly defined intended-use population, external evaluation and prospective assessment. The number of avoided false positives here is relative to a structured reference label in this cohort, not an observed reduction in clinical harm.

Reference labels and measurement availability

The endpoint was a recorded global or unspecified reduction in right ventricular systolic function, with regional-only findings handled separately by the codebook. It was not a quantitative composite of TAPSE, tissue Doppler velocity, fractional area change and pulmonary pressure. Label construction retained unknown values and did not replace the index examination to obtain a definite finding. This limits outcome-driven selection of examinations but does not remove selection associated with which patients underwent both tests or had sufficiently informative reports.

Missing severity was consequential: the H scenario produced a sensitivity disadvantage at the 95% target. This is an assumption analysis, not evidence that ungraded cases were truly moderate or severe. Similarly, the lower sensitivity in complete9 cannot establish why the result changed. Measurement completeness may reflect rhythm, equipment, documentation or other patient differences, and operationally usable numerical values are not necessarily physiologically validated. The prespecified missingness indicators preserve recorded patterns, but neither their inclusion nor the p_end analysis establishes or eliminates a particular device-related confounding pathway.

Limitations

The evaluation was an internal patient-level holdout from linked single-system data. Patients receiving both ECG and echocardiography are a selected clinical population; the findings do not estimate screening performance in unselected adults. External validation and subgroup performance evaluation were not performed. The descriptive table covers age, the recorded gender field, pairing interval and measurement availability; it does not characterize comorbidity, care setting or the clinical indication for either examination. The reference labels arose from routine structured records without an independent blinded reread, and incomplete labels could introduce selection and misclassification.

Timestamp concordance and a shorter pairing window do not validate ECG machine time. Changes in a patient's condition within the accepted interval could affect label correspondence. Waveform quality control imposed a common technical inclusion rule on both models, but performance in excluded or technically unsupported recordings was not estimated. Device identifiers and clinical covariates were not predictors, and excluding identifiers does not remove all device or care-setting signals from either input type.

The stacked waveform score used out-of-fold models trained on smaller subsets than the final CNNs, introducing a possible training-to-inference distribution difference. The study used three fixed CNN seeds without including training variability in the reported bootstrap intervals. Those intervals were conditional on fitted systems, validation calibration and thresholds; they also did not propagate label or transportability uncertainty. Secondary analyses had nominal intervals without multiplicity correction. Finally, implementation details were specified in stages, and the sparse-to-L2 baseline deviation was documented after training began. This chronology limits any claim of fully prospective specification and is disclosed rather than retrospectively removed.

Conclusions

Within a frozen internal test of patients with linked ECG and echocardiography, a waveform score improved prediction beyond machine measurements and interpretation text, with fewer false positives at a validation-selected working point. Actual sensitivity and unfavorable prespecified analyses qualify that finding. Clinical utility, sensitivity noninferiority and external applicability remain unestablished.

Abbreviations

AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; B, machine-feature baseline model; BW, waveform-augmented model; CI, confidence interval; CITL, calibration-in-the-large; CNN, convolutional neural network; ECG, electrocardiogram; IQR, interquartile range; MIMIC, Medical Information Mart for Intensive Care; OSF, Open Science Framework; QC, quality control; RV, right ventricle; TTE, transthoracic echocardiography.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The Ethics Committee of The First Affiliated Hospital with Nanjing Medical University determined that the present secondary analysis was exempt from ethics review; no reference number was issued. The source MIMIC resources were approved by the institutional review boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, with waiver of individual consent for the deidentified database.

[1, 10, 12] Access to credentialed data was obtained through PhysioNet and was subject to the applicable data-use agreement.

Consent for publication

Not applicable. The manuscript contains no identifiable individual participant information.

Availability of data and materials

The datasets analyzed in this study are available from PhysioNet as MIMIC-IV-ECG v1.0, MIMIC-IV-Echo v1.0.1 and MIMIC-IV Clinical v3.1. [1, 10, 11] The ECG release is open access; the Echo and Clinical resources require individual PhysioNet credentialing, the required training and agreement to their data-use terms. Patient-level records, linked labels and prediction arrays are not redistributed with this article. The analysis source code, configurations and integrity manifest are provided in Additional file 3 for editor and reviewer inspection. Copies of the registered Stage 1 documents and subsequent dated local specifications and amendments are provided in Additional file 4. The Stage 1 record is registered on OSF. [13] Its embargo does not make the local implementation supplements additional registrations. The archives contain no patient-level records, trained weights or individual predictions. The preserved code requires authorized inputs and local configuration; an independent end-to-end reproduction of the redistributed package has not been tested. No public repository deposition is claimed.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific funding.

Authors' contributions

JL processed the data and drafted the manuscript. YS guided the research direction and supervised the study. Both authors read and approved the final manuscript.

Acknowledgements

We acknowledge the contributors to the MIMIC resources and PhysioNet. [16]

Patient and public involvement

Patients and members of the public were not involved in the design, interpretation or dissemination of this secondary study.

References

1. Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0). PhysioNet. 2023. 10.13026/4nqg-sb35
2. Documentation MIMIC. machine_measurements: ECG machine measurements. Accessed 23 September 2026.
3. Mukherjee M, Rudski LG, Addetia K, Afilalo J, D'Alto M, Freed BH, et al. Guidelines for the Echocardiographic Assessment of the Right Heart in Adults and Special Considerations in Pulmonary Hypertension: Recommendations from the American Society of Echocardiography. J Am Soc Echocardiogr. 2025;38:141–86. 10.1016/j.echo.2025.01.006.
4. Poterucha TJ, Jing L, Ricart RP, Adjei-Mosi M, Finer J, Hartzel D, et al. Detecting structural heart disease from electrocardiograms using AI. Nature. 2025;644:221–30. 10.1038/s41586-025-09227-0.
5. Duong SQ, Vaid A, My VTH, Butler LR, Lampert J, Pass RH, et al. Quantitative Prediction of Right Ventricular Size and Function From the ECG. J Am Heart Assoc. 2024;13:e031671. 10.1161/JAHA.123.031671.
6. Carbonati T, Eslami P, Waks JW, Fiorina L, Chaudhari A, Henry C, et al. Deep neural networks detect regional wall motion abnormalities and preclinical cardiovascular disease from 12-lead ECGs. medRxiv [preprint]. 2024. 10.1101/2024.05.31.24308304.
7. Rogers AJ, Bhatia NK, Bandyopadhyay S, Tooley J, Ansari R, Thakkar V, et al. Identification of cardiac wall motion abnormalities in diverse populations by deep learning of the electrocardiogram. npj Digit Med. 2025;8:21. 10.1038/s41746-024-01407-y.
8. Joshi SM, Shaik HR, Rai Sharma S, Strong P, Srivatsa U, Ebong I, et al. Linking Electrocardiogram and Echocardiogram: Comparing Classical Machine Learning and Deep Learning Neural Networks for the Detection of Regional Wall Motion Abnormalities. IEEE Access. 2025;13:134519–28. 10.1109/ACCESS.2025.3592693.
9. Do DN, Do MN, Nguyen D, Le KTQ, Pham KD, Huynh HN et al. Domain-Adapted Fine-Tuning of ECG Foundation Models for Multi-Label Structural Heart Disease Screening. arXiv [preprint]. 2026;2604.23385v1. 10.48550/arXiv.2604.23385
10. Gow B, Pollard T, Greenbaum N, Moody B, Han A, Waks JW et al. MIMIC-IV-Echo: Echocardiogram Matched Subset (version 1.0.1). PhysioNet. 2026. 10.13026/307c-mr50
11. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R. MIMIC-IV (version 3.1). PhysioNet. 2024. 10.13026/kpb9-mt58

12. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. 10.1038/s41597-022-01899-x.
13. Li J. Incremental value of ECG waveforms for right ventricular systolic dysfunction: Stage 1 protocol registration (v0.1) [registration]. OSF; 2026. Accessed 23 September 2026. Embargoed registration.
14. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
15. Fuchigami Y. Machine-Reported Electrocardiographic Right Axis Deviation or Right Ventricular Hypertrophy and Echocardiographic Pulmonary Hypertension Phenotypes and Right-Heart Abnormalities. Echocardiography. 2026;43:e70539. 10.1111/echo.70539.
16. Pollard T, Moody BE, Lehman LWH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1:792–5. 10.1038/s44360-026-00096-z.

Figures

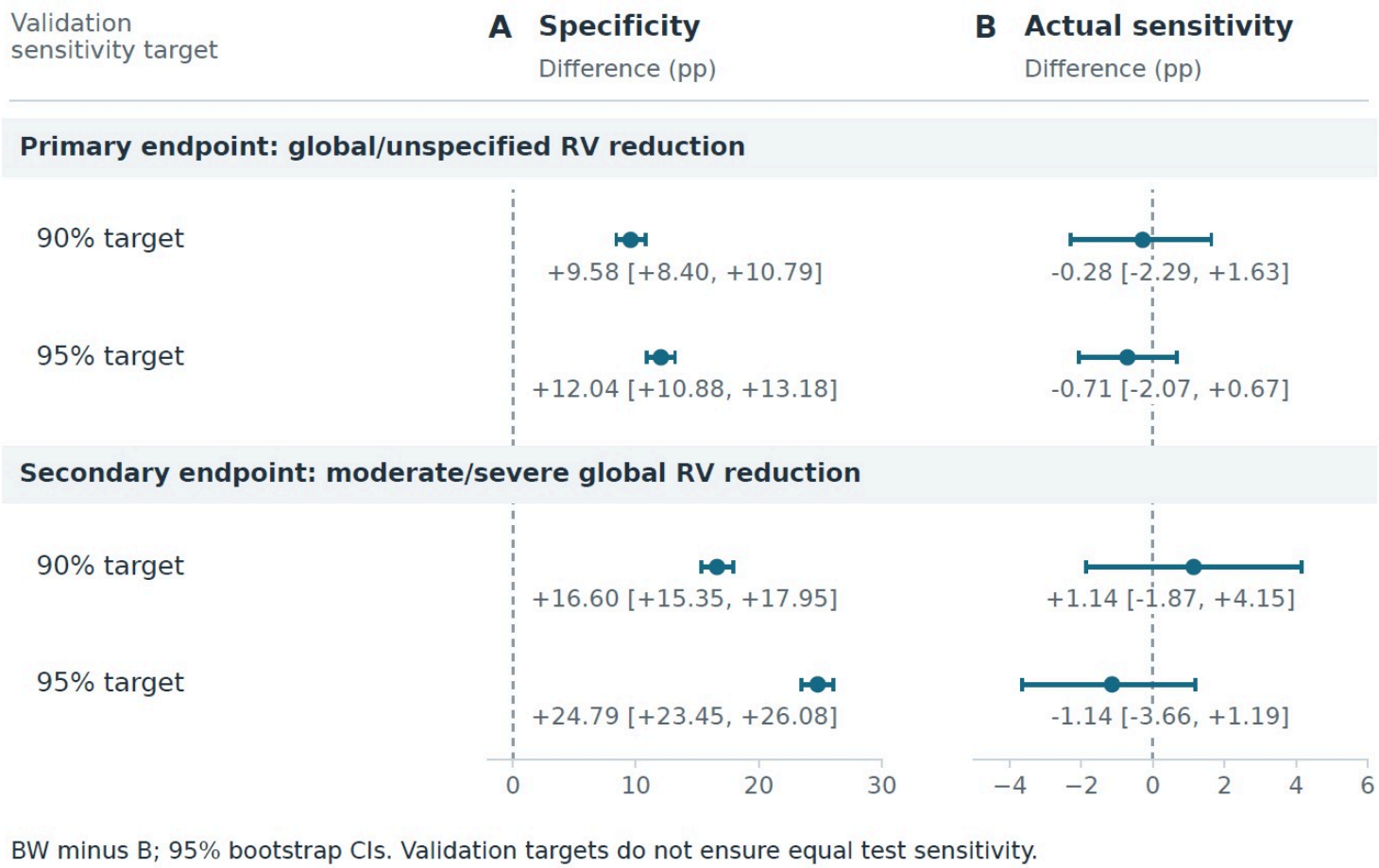


Figure 1

Working-point differences in the main test cohorts

Paired differences between the waveform-augmented model (BW) and the machine-feature model (B) are shown for test specificity (A) and actual test sensitivity (B). The primary endpoint is recorded global or unspecified reduction in right ventricular (RV) systolic function; records with regional-only abnormalities are not classified as positive under this definition. The secondary endpoint is explicitly graded moderate or severe global reduction. Each model's threshold was selected in its corresponding validation cohort to attain a sensitivity target of 90% or 95%, then fixed for testing. These targets do not constrain the models to equal sensitivity in the test cohort. Points denote BW minus B, and bars denote nominal 95% percentile confidence intervals from 2,000 nonstratified, paired patient-bootstrap resamples. Values are expressed in percentage points. Resampling holds the fitted models, validation calibration mappings, and validation-selected thresholds fixed; the intervals exclude uncertainty from their estimation. The zero reference lines denote no difference. The 90% target for the primary endpoint is the main working-point comparison; the other rows retain their prespecified secondary roles.

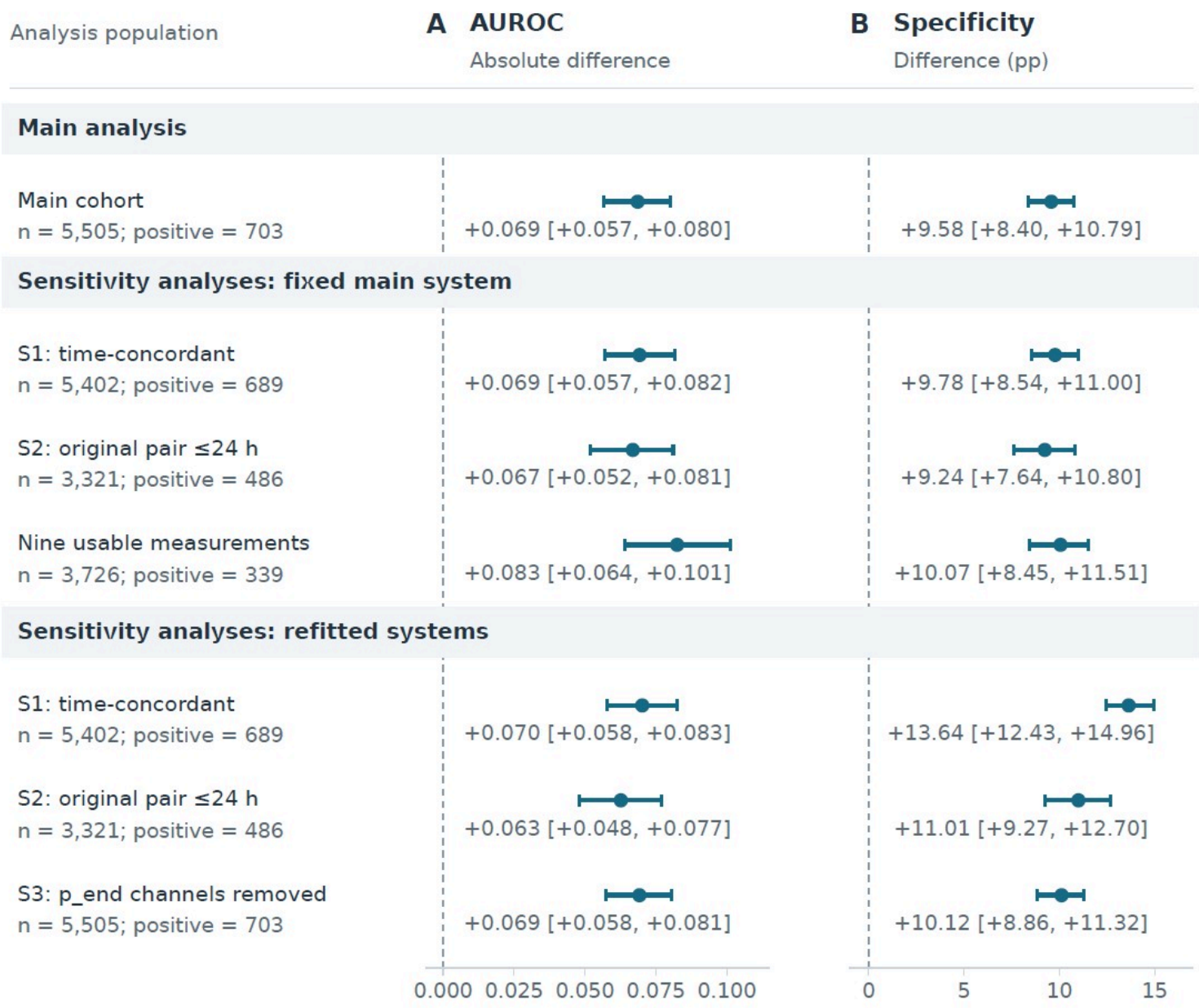


Figure 2

Prespecified primary-endpoint analyses

Paired differences in AUROC (A) and test specificity (B) compare BW with B across seven prespecified analyses of recorded global or unspecified reduction in RV systolic function. The specificity estimates use each system's threshold selected for a 90% validation sensitivity target; actual test sensitivity can differ between systems. AUROC is calculated from continuous scores and does not depend on that threshold. Row labels give the number of evaluated patients and patients positive for the primary endpoint. S1 retains original index records with concordant timestamp comparisons and excludes conflicts or unknown comparisons. S2 retains original ECG–echo pairs separated by no more than 24 hours. Fixed-main analyses apply the original main models, calibration mappings, and thresholds to the

restricted test populations. Refitted analyses use models, calibration mappings, and thresholds developed within the corresponding restricted populations. The nine-measurement analysis applies the fixed main system to records meeting all nine operational measurement-usability rules; operational usability is not a claim of physiological validity. S3 removes the p_end continuous input and all related indicators, then repeats the frozen development procedure. None of these analyses replaces the original index ECG or echo. Bars are nominal 95% percentile intervals from 2,000 paired patient-bootstrap resamples, conditional on the corresponding frozen prediction systems. Differences in AUROC are absolute, unitless changes; specificity differences are percentage points. The rows represent different analysis populations and development procedures, not a randomized comparison of those procedures or independent external replications.

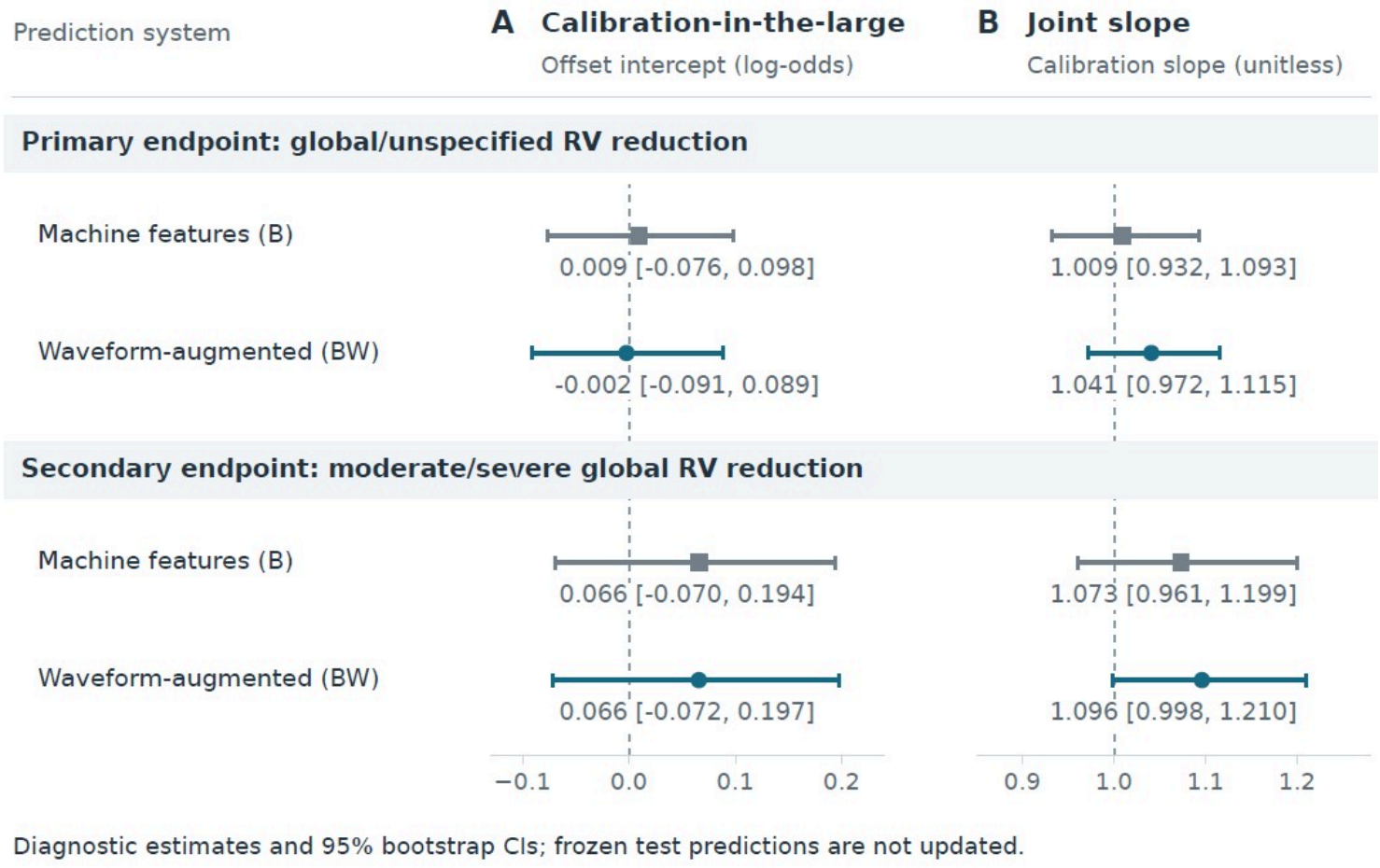


Figure 3

Calibration diagnostics of frozen predictions

Calibration-in-the-large (A) and the joint calibration slope (B) are shown separately for B and BW in the primary and secondary main test cohorts. Calibration-in-the-large is estimated as an intercept with the logit of the frozen probability entered as an offset and its coefficient fixed at one; its ideal reference is zero. The calibration slope comes from a separate logistic regression that jointly estimates an intercept and slope; its ideal reference is one. These are distinct diagnostic regressions, and the offset intercept is

not the intercept from the joint fit. Points and bars show the diagnostic estimates and nominal 95% percentile intervals from 2,000 patient-bootstrap resamples. Diagnostic regressions are re-estimated in each resample, but their coefficients are never applied to update the test predictions. Models and validation calibration mappings remain fixed. Intervals are shown under the frozen rule requiring the point estimate and every resample estimate to be defined. These summary diagnostics cannot exclude nonlinear miscalibration and do not establish adequate calibration on their own. The original calibration displays are retained separately as local appendices.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ABMCadditionalfile1v0.3.docx
- ABMCadditionalfile2TRIPODAIv0.3.docx
- Aadditionalfile3code20260923.zip
- Aadditionalfile4protocols20260923.zip