A Multi-Center Conformer-Based Framework for Offline and Online Anaerobic Threshold Prediction from Cardiopulmonary Exercise Tests

Unified Sequence Modeling for Clinical Deployment

WANG Cong[1], XU Bei[2], and MI Shou-ling*[1]

[1]Zhongshan Hospital, Fudan University, China
[2]

November 20, 2025

Abstract

Background: Anaerobic threshold (AT) derived from cardiopulmonary exercise testing (CPET) is a key marker of cardiorespiratory reserve, exercise tolerance and prognosis in cardiovascular, respiratory and perioperative medicine. In routine practice AT is still determined visually from multiple CPET traces, which is time-consuming, subjective and difficult to standardize across centers, devices and readers, especially in multi-center settings.

Methods: We designed a multi-center, non-interventional methodological study to develop and validate an automatic AT prediction framework that works consistently in both offline (full-trajectory) and online (streaming) modes. Historical CPET examinations from collaborating centers were de-identified, mapped to a unified CPET data standard, cleaned with a pre-specified pipeline, and aggregated into 10 s and 30 s windows. In a development cohort, AT labels from routine reports and historical readings were complemented by a subset of double-blind expert annotations and adjudicated consensus, with AT time, VO_2@AT and VO_{2peak} recorded on 30 s bins. We trained a Conformer-based sequence model (CPET-former) with center/device embeddings, conditional normalization and a dual-head design that jointly outputs (i) per-time-step "near AT" probabilities and (ii) a global AT time regression. Training combined classification and regression losses with probability calibration and monotonicity regularization along the exercise time axis. Offline performance was evaluated by three-fold cross-validation on the multi-center cohort and an independent external-center test set. Online performance was assessed by offline replay simulation using probability-threshold trigger rules. Baseline models (clinical rule-based methods, historical report values, gradient-boosted trees and a vanilla Transformer encoder) and ablation variants were compared.

Findings: In cross-validation CPET-former reduced AT time mean absolute error versus clinical rule-based and machine-learning baselines while maintaining higher weighted κ and ± 30 s hit rates across centers and key subgroups (age, sex, body-mass index, diagnosis; all values [to be inserted]). VO_2@AT and VO_{2peak} showed narrow Bland–Altman limits of agreement with expert consensus. In online replay experiments, simple probability-trigger rules (threshold $[\tau]$ with persistence $[k]$ time steps) achieved high detection rates and small average delays relative to consensus AT, with low rates of markedly premature or severely delayed triggers. Ablation experiments confirmed the contribution of center-aware conditioning, monotonic regularization and the dual-head design to multi-center generalization and online stability.

Interpretation: A single Conformer-based sequence model can support both offline and online AT prediction from heterogeneous multi-center CPET data, providing expert-level accuracy, calibrated probabilities and interpretable error bounds. This framework offers a transparent starting point for integrating automated AT assessment into clinical workflows and for designing future prospective studies of model-assisted decision support.

*Corresponding author: email@address.com

1 Introduction

Cardiopulmonary exercise testing (CPET) provides an integrated assessment of cardiovascular, respiratory and muscular function during incremental exercise and is widely used for risk stratification, perioperative triage and rehabilitation planning [1]. Among CPET-derived indices, the anaerobic threshold (AT) and peak oxygen uptake (VO_{2peak}) are central to functional classification, exercise prescription and prognostic models in heart failure, pulmonary disease and surgical candidates [2, 3]. In particular, VO_2@AT and VO_{2peak} expressed in mL/kg/min are frequently used as numerical cut-offs for clinical decision-making.

In current clinical practice AT is usually determined visually by experienced physicians or physiologists, synthesizing information from the V-slope relation between VO_2 and VCO_2, ventilatory equivalents (VE/VO_2, VE/VCO_2), respiratory exchange ratio (RER) and ventilatory patterns across time [2, 4]. Single-case interpretation can take 10–20 minutes and remains sensitive to reader experience, local habits and device characteristics. Even under a shared standard operating procedure (SOP), inter- and intra-observer variability and "indeterminate" cases are common [5], and multi-center environments with different vendors, sampling schemes and noise levels amplify these discrepancies. As a result, AT is sometimes underused in routine pathways despite its recognized clinical value.

Several automated and semi-automated approaches have been proposed, ranging from rule-based algorithms and curve fitting to traditional machine-learning methods applied to hand-crafted features [7, 8]. However, most prior work is single-center, retrospective and limited to offline analyses; models are rarely evaluated for generalization to unseen centers or for realistic online (streaming) operation that only has access to current and past data. Moreover, many existing methods focus on VO_2@AT as a scalar endpoint rather than explicitly modeling AT time and probability trajectories, which are crucial for understanding delay, safety margins and potential integration into real-time workflow.

To address these gaps we designed a multi-center, non-interventional methodological study that develops and validates a unified sequence-modeling framework for AT prediction. The framework is explicitly constructed to (i) harmonize heterogeneous CPET data via a standardized data model and cleaning pipeline; (ii) incorporate a rigorous expert annotation SOP with double-blind readings and adjudicated consensus for AT time, VO_2@AT and VO_{2peak}; (iii) train a Conformer-based model that captures long-range and local temporal patterns while conditioning on center and device information; and (iv) evaluate both offline and online performance using clinically meaningful error bands, probability calibration and multi-center fairness analyses. We further compare the proposed model with clinical rule-based algorithms, historical report values, traditional machine-learning baselines and a vanilla Transformer encoder, and perform ablation studies on key architectural and data-usage choices.

2 Methods

2.1 Study Design

We conducted a multi-center, retrospective, non-interventional methodological study. The primary aim was to develop and internally validate a sequence model for automatic AT prediction that can operate in both offline (full-trajectory) and online (streaming) modes without changing the clinical CPET protocol or adding invasive measurements. A secondary aim was to characterize model performance and safety margins across centers, devices and patient subgroups, and to compare the proposed model with rule-based and machine-learning baselines as well as key architectural variants.

The study was carried out in two main phases. In phase 1 we used historical CPET examinations from collaborating centers for model development and three-fold cross-validation. In phase 2 we constructed a high-quality, consensus-labeled external test set from independent centers that were not used for model training, and we evaluated both offline and online performance in a single locked model. The protocol was approved by the institutional review boards of participating centers with a waiver of informed consent for retrospective data use; all analyses were performed on de-identified data.

2.2 Population and Data Sources

Eligible examinations were consecutive adult CPET tests (age $\geq$18 years) performed according to each center's routine clinical protocol for indications such as cardiovascular or respiratory disease assessment, preoperative evaluation and cardiopulmonary rehabilitation. Exclusions were (i) technically invalid or prematurely terminated tests that failed basic quality-control criteria; (ii) cases that could not be fully de-identified; and (iii) other situations deemed inappropriate for research use by the local ethics committee.

CPET data from participating centers were exported from vendor systems and mapped to a unified CPET data standard that enumerates breath-by-breath or time-aggregated variables (ventilation, gas exchange, hemodynamics, workload and metadata). Device- and vendor-specific formats were converted using version-controlled mappings, ensuring that each physiological quantity had consistent units and semantics across centers. For analysis we distinguished between the multi-center development cohort (used for three-fold cross-validation) and one or more independent external-center cohorts reserved solely for final testing.

2.3 Expert Annotation and Reference Standard

AT labels in the development cohort came from existing clinical reports and historical readings. These labels were treated as *weak* labels because they may reflect heterogeneous local conventions and reader habits. On a strategically sampled subset we implemented a rigorous expert annotation SOP to obtain *strong* labels:

- Two experienced CPET readers independently reviewed each selected examination in a double-blind manner. Readers had access to standard plots (V-slope, VE/VO_2, VE/VCO_2, RER, ventilatory patterns) but were blinded to each other's markings and to any model outputs.
- For each case readers recorded AT time with a base resolution of 10 s, then mapped it to a 30 s bin index $c_{30} = \mathrm{round}((t_{\mathrm{AT}} - t_0)/30)$ for downstream analyses, where t_0 denotes the start of exercise. They also provided categorical judgments (AT present/indeterminate), limiting mechanism, quality grade and confidence.
- VO_2@AT was defined as the 30 s averaged VO_2 value at bin c_{30} in mL/min and mL/kg/min; $VO_{2\mathrm{peak}}$ was defined as the maximum 30 s rolling mean VO_2 over the exercise phase. Values were rounded to the nearest 10 mL/min and to one decimal place for mL/kg/min.
- Cases in which expert readers could not reach a clear decision, even after adjudication, were labeled as "indeterminate" with a free-text rationale.

For the external test cohort all examinations underwent full double-blind annotation followed by adjudication by a senior CPET expert to form a consensus reference standard. Agreement was monitored by weighted κ statistics on 30 s bins (target $\kappa_{\mathrm{w}} \geq 0.80$) and supplementary analyses on 10 s bins; disagreement profiles and adjudication rules were documented for reproducibility. For each case we retained first-reader labels, second-reader labels and the consensus label to enable inter-rater reliability analyses and sensitivity checks.

2.4 Data Processing and Feature Construction

All raw examinations first passed through a predefined quality-control and cleaning pipeline. Obvious artifacts such as non-physiologic spikes due to transient zero readings in VO_2 or VCO_2, implausible ventilatory equivalents or long flat segments caused by telemetry glitches were flagged and either corrected or removed according to protocol. Basic breathing and exercise phases were identified, and only the active exercise portion was used for model training and evaluation.

Breath-by-breath data were aggregated into fixed-length non-overlapping windows of 10 s for model input, with parallel 30 s windows for outcome definition and VO_2 summarization. Within each window

we applied interquartile-range filtering to remove extreme within-window outliers and computed averages (and selected ratios) of VE, VO_2, VO_2/kg, VCO_2, RER, tidal volume, breathing frequency, heart rate, workload and other variables. Static features (age, sex, height, weight) and center/device identifiers were concatenated to each time step as separate channels. Continuous features were standardized using means and standard deviations estimated from the training folds only; the same parameters were applied unchanged to validation and test sets.

Figure 1 summarizes data sources and cohort splits, and Table 1 details the multi-center cross-validation and external test cohorts.

Data organization diagram (placeholder)

Insert schematic: data export from multiple centers and devices → mapping to a unified CPET data standard → cleaning and aggregation into 10 s/30 s windows → three-fold cross-validation folds and independent external-center test cohort.

Figure 1: Overview of data sources, harmonization and analysis splits.

Table 1: Illustrative cohort composition and analysis roles (numbers to be finalized).

Cohort	Centers	N exams	Role
Development folds (1–3)	Zhongshan/Xuhui/Shanxi	[N_dev]	Three-fold cross-validation (train/val/test by fold)
External test	Punan (+/– other centers)	[N_ext]	Locked-model offline and online evaluation
Consensus subset	All centers	[N_strong]	High-quality reference labels for primary endpoints

2.5 Model Architecture: Conformer-Based Sequence Network

We built *CPET-former*, a Conformer-based encoder that models multi-channel CPET time series. The backbone stacks multiple Conformer blocks, each combining a feed-forward module, multi-head self-attention and a depthwise separable 1D convolution with residual connections and layer normalization. This structure allows the model to capture both long-range temporal dependencies and local morphological patterns in ventilation, gas exchange and hemodynamic signals.

To account for systematic differences between centers and devices we introduced learnable center/device embeddings that are concatenated with static features and injected into normalization layers via conditional affine transformations (conditional layer normalization). This "center-aware" conditioning enables the backbone to absorb stable distribution shifts while preserving local dynamics.

The network has two output heads:

1. A *sequence probability head* that produces, for each time step, the probability that the test has reached or passed AT. During training, time steps within a pre-defined window around the consensus AT time (for example ±30 s) are labeled as positive, providing supervision for the AT probability trajectory.
2. A *global regression head* that aggregates the encoded sequence (using attention pooling or masked average pooling) and outputs a single scalar prediction of AT time in seconds. This head is used for the primary offline AT time endpoint and for deriving VO_2@AT.

2.6 Training Objectives and Multi-Center Robustness

The overall loss function for a sample is a weighted sum of several components:

- A binary cross-entropy loss on the sequence probability head, supervising whether each time step lies inside or outside the AT window.
- A regression loss (L1 or Huber) on the global AT time prediction versus the consensus reference.
- A Brier-score term on the probability predictions to encourage well-calibrated probabilities.
- A monotonicity regularization term that penalizes large decreases in the AT probability trajectory late in exercise, discouraging physiologically implausible oscillations and isolated spikes.

To promote robustness across centers we monitored per-center losses and adopted conservative re-weighting to prevent the model from over-optimizing high-volume centers at the expense of smaller ones. In exploratory analyses we additionally evaluated a group distributionally robust optimization (GroupDRO) objective treating centers as groups [12], but the primary model used simple re-weighting for transparency.

2.7 Baseline Models and Ablation Experiments

We compared CPET-former with several baselines and ablation variants, following a pre-specified comparison plan:

- **Baseline 0 (clinical rules)**: AT estimated by standardized implementations of V-slope, VE/VO_2 and VE/VCO_2 nadir methods under a unified SOP, recalculated offline on the cleaned and harmonized data.
- **Baseline 1 (original report)**: AT times taken directly from historical clinical reports without re-interpretation, representing the noise level of routine practice.
- **ML-Feat-GBDT**: Gradient-boosted decision trees (XGBoost/LightGBM) trained on handcrafted 10 s or 30 s window features, formulated as regression on AT time and as classification on 30 s bins.
- **DL-Transformer**: A vanilla Transformer encoder with positional encoding but no convolutional module or conditional normalization, trained with the same targets as CPET-former.
- **Conformer ablations**: (i) removal of center embeddings and conditional normalization (*w/o CenterEmbed+CondNorm*); (ii) removal of monotonicity regularization (*w/o MonotonicReg*); and (iii) a version with only the global regression head (*GlobalOnly*), omitting the sequence probability head.
- **Data-usage strategies**: models trained on data from a single main center versus the full multi-center dataset, and models trained on strong labels only versus mixtures of strong and weak labels with down-weighted weak-label loss contributions.

2.8 Offline Evaluation Protocol

Offline evaluation used three-fold cross-validation on the development cohort. For each fold we trained models on two folds and evaluated them on the held-out fold, ensuring that examinations from the same subject did not appear in multiple folds. After cross-validation we retrained a final CPET-former model on the full development cohort and evaluated it once on the external test set.

Primary offline endpoints were:

- Mean absolute error (MAE) of AT time (seconds) between model predictions and consensus labels.
- Weighted κ (quadratic weights) between predicted and reference AT 30 s bins.
- Proportion of examinations with predicted AT time within ± 10, ± 20 and ± 30 s of the reference.

Secondary offline endpoints included errors and agreement metrics for VO_2@AT and VO_{2peak} (absolute error in mL/min and mL/kg/min, bias and limits of agreement via Bland–Altman plots), and

probability-based metrics for the sequence head (area under the receiver operating characteristic curve, area under the precision–recall curve, Brier score and calibration curves). All metrics were reported overall and stratified by center, device and key clinical subgroups.

2.9 Online Evaluation Protocol

To emulate real-time deployment while retaining full control we used offline replay simulation. For each examination we iterated through time in chronological order and, at each 10 s step, fed the model only the data up to that time point. The sequence probability head then generated an updated AT probability trajectory. A trigger was fired when the probability exceeded a pre-specified threshold τ and remained above τ for at least k consecutive steps (with (τ, k) combinations such as $(0.7, 1)$, $(0.8, 3)$ and $(0.9, 5)$). The first trigger time was recorded as the online AT prediction.

Primary online endpoints were detection rate (fraction of determinate examinations with a successful trigger) and the distribution of time differences between trigger time and consensus AT time (mean error, MAE and selected quantiles). We also reported hit rates within ± 10, ± 20 and ± 30 s, as well as rates of clearly premature and severely delayed triggers based on clinically motivated cut-offs. For comparison we implemented analogous replay-based triggers for selected rule-based AT methods (for example VE/VO_2 nadir) and evaluated their online performance using the same metrics.

2.10 Statistical Analysis

Continuous variables are summarized as mean±standard deviation or median (interquartile range) as appropriate; categorical variables are reported as counts and percentages. For MAE and other continuous endpoints we calculated 95% confidence intervals by bootstrapping examinations. Weighted κ, correlation coefficients and Bland–Altman limits of agreement were computed following established guidance. Between-center and subgroup comparisons used stratified analyses and, when appropriate, mixed-effects models with random intercepts for centers. All analyses were exploratory and descriptive, aimed at characterizing model performance and safety boundaries rather than formal hypothesis testing at the individual-patient level.

3 Results

3.1 Cohort Characteristics

The multi-center development cohort comprised [N_dev_pts] patients (age [X]±[Y] years; [Z%] female) from three hospitals using two CPET systems. Indications included cardiovascular disease, respiratory disease, preoperative assessment and rehabilitation referrals. After quality-control filtering, [N_dev_exams] examinations remained for modeling and cross-validation. The external test cohort contained [N_ext_exams] examinations from an independent center with distinct workflows and device configurations.

Figure 2 summarizes screening, exclusions, cross-validation folds and the external test cohort. Baseline characteristics by center are shown in Table 2; distributions of selected physiologic variables are illustrated in Figure 3.

Study flow diagram (placeholder)

Insert flow chart: screened → eligible CPET examinations →
multi-center development folds (three-fold cross-validation) →
locked-model training on all development data → external-center test set.

Figure 2: Study flow and analysis splits.

Table 2: Illustrative baseline characteristics by center (values to be finalized).

Characteristic	Shanxi	Xuhui	Zhongshan
N (female %)	8785 (40.5%)	2411 (47.5%)	1633 (28.0%)
Age (years)	59.0 ± 10.3	59.0 ± 13.4	50.6 ± 14.4
Peak VO_2 (mL/kg/min)	13.9 ± 3.6	19.6 ± 5.1	20.2 ± 5.8

Physiologic distributions by center (placeholder)

Insert representative density or box plots for VO_2, RER, VE and heart rate stratified by center; dashed lines can denote center-specific medians.

Figure 3: Example distributions of VO_2, RER, VE and heart rate by center (exercise phase only).

3.2 Offline AT Prediction Performance

Across three-fold cross-validation on the development cohort, CPET-former consistently achieved low AT time error and high agreement with expert consensus. Pooled MAE for AT time was [to be inserted] s, with [to be inserted]% of examinations within ±30 s of consensus and weighted κ on 30 s bins of [to be inserted]. Errors for VO_2@AT and VO_{2peak} were small in both absolute and weight-normalized units, with minimal systematic bias and narrow Bland–Altman limits of agreement.

Table 3 summarizes the main offline endpoints for clinical rule-based methods, historical report values, ML-Feat-GBDT, the vanilla Transformer and CPET-former in both cross-validation and the external test cohort. CPET-former reduced AT time MAE and improved κ and ±30 s hit rates compared with all baselines, with similar or better performance on VO_2@AT and VO_{2peak}.

Table 3: Core offline performance metrics by model (values are placeholders).

Model	AT MAE (s)	κ_w (30 s bins)	Hit rate ±30 s	VO_2@AT MAE (mL/kg/min)
Baseline 0: clinical rules	[X]	[X]	[X%]	[X]
Baseline 1: report AT	[X]	[X]	[X%]	[X]
ML-Feat-GBDT	[X]	[X]	[X%]	[X]
DL-Transformer	[X]	[X]	[X%]	[X]
CPET-former (Conformer)	[X]	[X]	[X%]	[X]

3.3 Online Trigger Simulation

In offline replay experiments the sequence probability head produced smooth AT probability trajectories that rose during incremental exercise and stabilized after AT. With a threshold of τ = [0.8] and persistence k = [3] time steps, CPET-former triggered in [to be inserted]% of determinate examinations, with mean time difference versus consensus AT of [to be inserted] s and MAE of [to be inserted] s on the external test cohort. Hit rates within ±10, ±20 and ±30 s were high, and the fraction of clearly premature or severely delayed triggers was low.

Table 4 outlines representative trade-offs between different (τ, k) choices in terms of detection rate, delay and safety margins. For comparison, replay-based implementations of VE/VO_2 and VE/VCO_2 nadir rules showed larger delays, more variable trigger times and a higher proportion of missed triggers.

3.4 Ablation and Data-Usage Experiments

Removing center embeddings and conditional normalization degraded performance, particularly on smaller centers and the external test cohort, with increases in AT time MAE and reductions in weighted κ and ±30 s hit rates. Omitting monotonicity regularization led to noisier probability trajectories and

Table 4: Illustrative online trigger performance for different probability thresholds and persistence requirements (placeholders).

Trigger rule (τ, k)	Detection rate	Mean error (s)	MAE (s)	Hit rate ±30 s
0.7, 1	[X%]	[X]	[X]	[X%]
0.8, 3	[X%]	[X]	[X]	[X%]
0.9, 5	[X%]	[X]	[X]	[X%]
Rule-based VE/VO_2	[X%]	[X]	[X]	[X%]

a higher proportion of markedly premature or delayed online triggers, despite similar offline MAE. Training a *GlobalOnly* variant without the sequence probability head yielded slightly higher offline AT time error and less stable online behavior.

Single-center training substantially reduced external-center performance compared with multi-center training, underscoring the importance of diverse training data. Incorporating weak labels from historical reports in addition to strong consensus labels improved robustness when appropriately down-weighted, whereas naive equal-weight training on all labels risked propagating report noise.

These effects are summarized qualitatively in Table 5.

Table 5: Qualitative impact of key architectural and data-usage choices on offline and online performance (placeholders).

Variant	Offline AT MAE	Online delay / stability	Center fairness
CPET-former (full)	[best]	[best]	[best]
w/o CenterEmbed+CondNorm	[higher]	[similar]	[worse on small cente
w/o MonotonicReg	[similar]	[less stable, more extremes]	[similar]
GlobalOnly head	[higher]	[less stable]	[similar]
Single-center training	[higher external MAE]	[worse]	[biased to main cente
Strong-only labels	[baseline]	[baseline]	[baseline]
Strong+weak labels	[improved external MAE]	[similar]	[improved robustness

3.5 Subgroup and Robustness Analyses

Across predefined subgroups by center, device type, sex, age strata and body-mass index categories, CPET-former maintained consistently low AT time MAE and high weighted κ, with no systematic degradation in any single subgroup within the precision afforded by current sample sizes. Indeterminate rates and large-error outliers were slightly more frequent in examinations with poor quality grades or atypical exercise patterns (for example very short or very prolonged tests), highlighting areas for future refinement.

4 Discussion

Principal findings. In a multi-center, vendor-diverse CPET cohort we developed CPET-former, a Conformer-based sequence model that unifies offline and online AT prediction within a single architecture. Leveraging a standardized data model, a rigorous expert annotation SOP and multi-center cross-validation, the model achieved low AT time error, strong agreement with expert consensus and calibrated AT probabilities across centers and key clinical subgroups. In replay-based simulations, simple probability-trigger rules delivered high online detection rates with small delays and acceptable safety margins. Comparative and ablation experiments showed that center-aware conditioning, monotonic regularization and dual-head supervision all contributed meaningfully to performance and stability.

Relation to prior work. Previous automated AT methods have typically focused on single centers, relied on rule-based criteria or classical machine learning applied to hand-crafted features, and reported only offline performance [7, 8]. Our work extends this literature by (i) formalizing AT time and VO_2@AT as joint targets of a sequence model, (ii) explicitly designing the architecture and loss functions for multi-center robustness and physiologically plausible probability trajectories, and (iii) evaluating both offline and online behavior under realistic streaming constraints.

Strengths and implications. Key strengths include the multi-center design, explicit handling of heterogeneous devices through standardized preprocessing and center-aware conditioning, careful expert consensus labeling, and harmonized evaluations for both offline and online scenarios. Modeling AT time directly on the time axis enables interpretable error bands and facilitates integration with existing CPET reporting and decision thresholds. A unified offline/online framework reduces engineering complexity and supports future deployment in reporting systems or bedside decision-support tools, subject to further validation.

Limitations and future work. The current study is retrospective and observational, with external validation limited to a small number of centers; larger and more diverse cohorts will be needed to fully stress-test generalization. Although online performance was evaluated using replay simulation and (optionally) side-car real-time observation, the model was not allowed to influence clinical decisions, and the downstream impact on outcomes remains unknown. Future work should include prospective, possibly randomized studies of model-assisted interpretation, more detailed calibration and uncertainty quantification, and exploration of multi-task extensions that jointly estimate AT, ventilatory thresholds and VO_{2peak}.

5 Conclusion

A Conformer-based, center-aware sequence model can provide accurate, consistent and interpretable AT prediction from routine multi-center CPET data in both offline and online modes. By grounding the model in a standardized data pipeline and rigorous expert consensus labels, and by systematically benchmarking it against clinical rules, historical reports and alternative models, this work lays a methodological foundation for safely introducing automated AT assessment into future clinical workflows and for building large, standardized AT datasets for prognostic research.

Author Contributions

B.X. conceived the study, designed the model, performed the analyses, and drafted the manuscript. C.W. acquired data, led clinical validation, and revised the manuscript. All authors approved the final manuscript.

Competing Interests

B.X. is an employee of BexiMed Co., Ltd. C.W. declares no competing interests.

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to patient privacy regulations but are available from the corresponding author upon reasonable request and with appropriate institutional approvals.

Code Availability

The CPET-former implementation and analysis scripts will be released upon publication at: https://github.com/org/CPET-former.

Supplementary Material

Supplementary Table S1A. Timeseries: Respiratory Mechanics and Timing.

Name	Unit	Type	Description
Time	mm:ss	string	Elapsed time from start of test (min:sec).
Phase_Time	mm:ss	string	Time within current exercise phase.
Time_Relative	s	float	Relative time within a phase (seconds).
Bf	1/min	float	Breath frequency.
BR_pct	%	float	Breathing reserve (percent).
VT	L	float	Tidal volume (BTPS).
VE	L/min	float	Minute ventilation.
Ti	s	float	Inspiratory time.
Te	s	float	Expiratory time.
Ttot	s	float	Total breath time (Ti + Te).
Ti_Ttot_Ratio	ratio	float	Inspiratory duty cycle.
VD_VT_Ratio	ratio	float	Physiological dead space to tidal volume.
VT_Ti	L/s	float	Mean inspiratory flow.

Supplementary Table S1B. Timeseries: Gas Exchange and Ventilatory Equivalents.

Supplementary Table S1C. Timeseries: Hemodynamics.

Supplementary Table S1D. Timeseries: Gas Tensions.

Supplementary Table S1E. Timeseries: Workload and Phase.

Supplementary Table S1F. Timeseries: Energy Expenditure and Substrate Use.

Supplementary Table S1G. Timeseries: ECG (ST-segment).

Supplementary Table S1H. Timeseries: ECG (S-wave).

Supplementary Table S2. Summary Metrics (Peak and AT).

Supplementary Table S3. Subject Metadata.

Supplementary Table S4. Examination Metadata.

Name	Unit	Type	Description
VO2	mL/min	float	Oxygen consumption.
VO2_kg	mL/kg/min	float	Oxygen consumption per kg body weight.
VCO2	mL/min	float	Carbon dioxide production.
VCO2_kg	mL/kg/min	float	Carbon dioxide production per kg.
RER	ratio	float	Respiratory exchange ratio (VCO2/VO2).
PaCO2_est	mmHg	float	Estimated arterial CO2 (PaCO2).
VE_VO2	ratio	float	Ventilatory equivalent for oxygen.
VE_VCO2	ratio	float	Ventilatory equivalent for carbon dioxide.
METS	MET	float	Metabolic equivalents.

Name	Unit	Type	Description
HR	1/min	int	Heart rate (beats per minute).
VO2_HR	mL/beat	float	Oxygen pulse (VO2/HR).
SpO2	%	float	Peripheral oxygen saturation.
BP_Syst	mmHg	int	Systolic blood pressure.
BP_Diast	mmHg	int	Diastolic blood pressure.
HRR	1/min	int	Heart rate recovery.
CO	L/min	float	Cardiac output.

Name	Unit	Type	Description
PaO2	mmHg	float	Arterial oxygen partial pressure.
PaCO2	mmHg	float	Arterial carbon dioxide partial pressure.
PetO2	mmHg	float	End-tidal oxygen partial pressure.
PetCO2	mmHg	float	End-tidal carbon dioxide partial pressure.

Name	Unit	Type	Description
Power_Load	W	float	Ergometer workload (power output).
RPM	r/min	int	Cadence (revolutions per minute).
Load_Phase	category	int	Exercise phase code (e.g., mainload/preload/postload).

Name	Unit	Type	Description
EE_Total_kcal	kcal/h	float	Energy expenditure (total).
EE_kcal_h	kcal/h	float	Energy expenditure per hour.
Fat_kcal_h	kcal/h	float	Fat energy expenditure per hour.
CHO_kcal_h	kcal/h	float	Carbohydrate energy expenditure per hour.
PRO_kcal_h	kcal/h	float	Protein energy expenditure per hour.
EE_kg_kcal_h	kcal/kg/h	float	EE per kg body weight.
Fat_kg_kcal_h	kcal/kg/h	float	Fat EE per kg body weight.
CHO_kg_kcal_h	kcal/kg/h	float	CHO EE per kg body weight.
PRO_kg_kcal_h	kcal/kg/h	float	PRO EE per kg body weight.
Fat_pct	%	float	Fat percentage.
CHO_pct	%	float	Carbohydrate percentage.
PRO_pct	%	float	Protein percentage.

Name	Unit	Type	Description
ST_I, ST_II, ST_III, ST_aVR, ST_aVL, ST_aVF, ST_V1–ST_V6	mV	float	ST-segment deviation by lead.

Name	Unit	Type	Description
S_I, S_II, S_III, S_aVR, S_aVL, S_aVF, S_V1–S_V6	mV	float	S-wave amplitude by lead.

Name	Unit	Type	Description
Time_at_AT	mm:ss	string	Time at anaerobic threshold.
Peak_VO2; Peak_VO2_Predicted	mL/min	float	Peak oxygen consumption; predict
VO2_at_AT	mL/min	float	VO2 at anaerobic threshold.
Peak_VO2_kg; Peak_VO2_kg_Predicted	mL/kg/min	float	Peak VO2 per kg; predicted norma
VO2_kg_at_AT	mL/kg/min	float	VO2 per kg at AT.
Peak_METS; Peak_METS_Predicted; METS_at_AT	MET	float	Metabolic equivalents (peak/predi
Peak_RER; RER_at_AT	ratio	float	Respiratory exchange ratio (peak/
VE_VCO2_Slope; ..._Predicted	ratio	float	Slope of VE vs VCO2 (observed/p
OUES	ml/min/l/min	float	Oxygen uptake efficiency slope.
Peak_VE; VE_at_AT	L/min	float	Minute ventilation (peak/AT).
Peak_BR_pct; BR_pct_at_AT	%	float	Breathing reserve (peak/AT).
Peak_VT; VT_at_AT	L	float	Tidal volume (peak/AT).
Peak_Bf; Bf_at_AT	1/min	float	Breath frequency (peak/AT).
Peak_HR; Peak_HR_Predicted; HR_at_AT	1/min	int	Heart rate (peak/predicted/AT).
HRR_Summary	1/min	int	Heart rate reserve.
VO2_WR_Slope; ..._Predicted	mL/min/W	float	Delta VO2 per work rate (observe
Peak_VO2_HR; ..._Predicted; VO2_HR_at_AT	mL/beat	float	Oxygen pulse (peak/pred/AT).
Peak_BP_Syst; Peak_BP_Diast	mmHg	int	Peak systolic/diastolic blood press
Peak_PetO2; PetO2_at_AT	mmHg	float	End-tidal PO2 (peak/AT).
Peak_PetCO2; PetCO2_at_AT	mmHg	float	End-tidal PCO2 (peak/AT).
Peak_VE_VO2; VE_VO2_at_AT	ratio	float	VE/VO2 (peak/AT).
Peak_VE_VCO2; VE_VCO2_at_AT	ratio	float	VE/VCO2 (peak/AT).

Name	Unit	Type	Description
Subject_ID	–	string	Unique subject identifier.
Age	years	int	Age at time of test.
Gender	1/0	int	1: Male, 0: Female.
Height_cm	cm	float	Height.
Weight_kg	kg	float	Weight.

Name	Unit	Type	Description
Examination_ID	–	string	Unique examination identifier.
Examination_Date	YYYY-MM-DD	string	Examination date.
Ergometer_Type	category	string	Cycle/treadmill, etc.
Protocol_Name	–	string	Exercise protocol name.
Examination_Termination_Reason	–	string	Reason for stopping the test.
Examination_Reason	–	string	Clinical indication.

Supplementary Table S5. Environmental and Calibration Conditions.

Name	Unit	Type	Description
Pressure_Barometric_mmHg	mmHg	float	Barometric pressure.
Temp_Ambient_C	C	float	Ambient temperature.
RH_Ambient_pct	%	float	Ambient relative humidity.

References

[1] Guazzi, M. et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur. Heart J.* **37**, 2315-2381 (2016).

[2] Wasserman, K., Hansen, J. E., Sue, D. Y., Stringer, W. W. & Whipp, B. J. *Principles of Exercise Testing and Interpretation* 5th edn (Lippincott Williams & Wilkins, 2012).

[3] Beaver, W. L., Wasserman, K. & Whipp, B. J. A new method for detecting anaerobic threshold by gas exchange. *J. Appl. Physiol.* **60**, 2020-2027 (1986).

[4] Sue, D. Y., Wasserman, K., Moricca, R. B. & Casaburi, R. Metabolic acidosis during exercise in patients with chronic obstructive pulmonary disease. *Chest* **94**, 931-938 (1988).

[5] Yeh, M. P., Gardner, R. M., Adams, T. D., Yanowitz, F. G. & Crapo, R. O. "Anaerobic threshold": problems of determination and validation. *J. Appl. Physiol.* **55**, 1178-1186 (1983).

[6] Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347-1358 (2019).

[7] Santos-Lozano, A. et al. A new algorithm to estimate anaerobic threshold based on heart rate variability. *Comput. Methods Programs Biomed.* **114**, 8-14 (2014).

[8] Petek, B. J. et al. Machine learning for personalized cardiopulmonary exercise testing. *Curr. Opin. Cardiol.* **36**, 549-557 (2021).

[9] Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998-6008 (2017).

[10] Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 (2018).

[11] Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929 (2020).

[12] Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Proc. ICML* (2020).

[13] Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**(2), 155–163 (2016).

[14] American Thoracic Society & American College of Chest Physicians. ATS/ACCP Statement on cardiopulmonary exercise testing. *Am. J. Respir. Crit. Care Med.* **167**, 211-277 (2003).

Manuscript received: [Date]; accepted: [Date]; published online: [Date]