A Clinically Generalizable Artificial Intelligence for Automated Anaerobic Threshold Assessment from Cardiopulmonary Exercise Tests

Multi-Center Standardization and Expert-Level Accuracy

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

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

November 3, 2025

Abstract

Background: Anaerobic threshold (AT) from cardiopulmonary exercise testing (CPET) guides risk stratification, perioperative triage, and rehabilitation. In practice, manual AT determination is subjective, time-consuming, and inconsistently standardized across centers/devices, limiting access and scale.
Methods: We assembled a large, multi-center, vendor-diverse CPET cohort (12,829 examinations) with cross-center/device harmonization. Expert consensus AT labels were obtained via two independent readers with blinded adjudication. We developed a transformer-based model (CPET-former) for multi-channel CPET time-series, introduced center-aware FiLM to model site/device effects, and used GroupDRO to improve robustness to unseen centers. Generalization was evaluated by a pooled, stratified 80%/10%/10% train/validation/test split, and by testing on two external centers to simulate unseen centers. A blinded reader study compared AI with clinicians across seniority, and decision utility was assessed at VO_2@AT thresholds (xx mL/kg/min).
Findings: CPET-former achieved expert-level accuracy with perfect reproducibility (ICC=xx). FiLM improved performance across known centers, and GroupDRO reduced worst-center error on external centers. In the reader study, AI was non-inferior to senior experts and outperformed junior/intermediate readers. Threshold analyses indicated favorable agreement and positive decision utility over clinically relevant ranges.
Interpretation: A clinically generalizable, auditable AI enables objective and consistent AT assessment across centers and devices. By aligning with expert performance and supporting decision thresholds, the framework can standardize CPET interpretation and reduce clinical workload.

1 Introduction

Cardiopulmonary exercise testing (CPET) informs risk stratification, perioperative triage, and rehabilitation planning [1]. Among CPET metrics, anaerobic threshold (AT) is widely used (e.g., VO_2@AT near xx mL/kg/min) to guide decisions in heart failure and major surgery [2, 3]. Yet, visual AT determination (e.g., V-slope) [4] is subjective, time-consuming, and inconsistently standardized, yielding modest inter-/intra-observer agreement [5] and constraining real-world scale.

Existing automated methods (curve-fitting or limited ML) are sensitive to protocol/device variability and rarely validated across unseen centers [7, 8]. To expand equitable access to CPET-informed care, a clinically generalizable, auditable, and reproducible AT solution is needed.

We report a multi-center framework that unifies heterogeneous CPET data across vendors and delivers expert-level, reproducible AT assessment. Our contributions are: (i) a large, vendor-diverse cohort across three hospitals; (ii) cross-vendor signal harmonization; (iii) CPET-former, a transformer tailored

*Corresponding author: email@address.com

to CPET time-series with center-aware FiLM for known-center generalization and GroupDRO for unseen-center robustness [12]; (iv) a blinded reader study spanning seniority; and (v) decision-utility analyses aligned to clinical thresholds. We hypothesize non-inferiority to senior experts with robust generalization and operational readiness.

2 Methods

2.1 Design and Reference Standard

We conducted a multi-center, retrospective diagnostic accuracy study with a prospective-simulated blinded reader study. Adults undergoing ramp-protocol CPET who met prespecified effort/quality criteria were included; incomplete or technically invalid files were excluded. Institutional review boards at [Zhongshan], [Shanxi], and [Xuhui] approved the study with consent waived. The reference standard for AT was established by two independent readers with blinded adjudication of disagreements; readers were blinded to AI outputs and to each other. The primary endpoint was MAE of VO_2@AT; secondary endpoints included RMSE, R^2, Bland–Altman, intraclass correlation coefficient (ICC), and calibration (slope/intercept). A clinically acceptable error band (e.g., ± 1.0 mL/kg/min) was prespecified for interpretability. Subgroup analyses were prespecified by center, device, sex, age, and protocol duration.

2.2 Data and Harmonization

We curated 12,829 ramp-protocol examinations across three hospitals (Shanxi, Xuhui, Zhongshan) using two vendors (Ganshorn, COSMED). Two additional centers (punan, rizhao) were held out as external test sets to evaluate generalization to unseen centers.

Data sources and standard. We defined a CPET data standard and device-specific conversion protocols to harmonize heterogeneous exports across vendors. The standard covers breath-by-breath/second-by-second signals, summary endpoints, and metadata. Based on this, we built the dataset with 69 feature channels, 4 AT targets, and 10 metadata fields. Names and units are normalized for ventilation and gas exchange (e.g., VE [L/min], VO_2 [mL/min], VO_2/kg [mL/kg/min], VCO_2 [mL/min], RER), hemodynamics (HR [1/min], VO_2/HR [mL/beat]), gas tensions ($PetO_2$/$PetCO_2$ [mmHg]), workload (Power_Load [W], RPM [r/min]), and ECG-derived features (ST/S amplitudes [mV]). Time is represented consistently via `Time`/`Phase_Time` (mm:ss) and `Time_Relative` (s); `Load_Phase` delineates exercise stages. A concise overview is provided in Table 1, with full specifications in the Supplementary Material.

Splits and evaluation. The primary dataset is the pooled mix of Zhongshan/Xuhui/Shanxi. We performed stratified patient-level splits into train/validation/test (e.g., 80%/10%/10%), preserving center/device proportions and preventing subject-level leakage; splits were seeded and frozen. For rapid model exploration, we additionally drew a per-center 30% subsample of the primary dataset (used only for fast iteration, with consistent stratification). Generalization was assessed by these pooled splits and by two explicit external-center test sets (punan and rizhao). For all external-center evaluations, standardization parameters were fit only on the primary-dataset training split and applied unchanged to the external sets. Results are reported in physical units (see Table 2).

Preprocessing and robustness. We adopted a conservative principle: preserve the original data distribution and avoid over-optimization unless a field exhibits extreme outliers. First, in the Xuhui dataset we filtered outliers in VE/VO_2 and VE/VCO_2 that arose when VO_2 or VCO_2 momentarily dropped to near-zero due to device or procedural glitches, inflating the ratios far beyond physiologic ranges. We set the maxima observed in Zhongshan and Shanxi as upper bounds and removed Xuhui samples exceeding those limits. Second, because some source datasets used 10s aggregation during clinical review to suppress spiky artifacts, we aligned breath-by-breath tests by aggregating to 10s windows; within each window, feature columns were filtered by the IQR rule and then averaged. Finally, features were standardized with z-score parameters fit on the training set and applied unchanged to validation/test; targets were standardized as needed. AT targets followed the standard nomenclature

Table 1: Concise overview of the CPET data standard used in this study. Groups list representative variables and typical units; the full specification appears in the Supplement.

Group	Representative variables	Units (rep.)
Ventilation	VE, VT, Bf, BR_pct, Ti/Te/Ttot, Ti_Ttot_Ratio, VD/VT, VT/Ti	L/min; L; 1/min; s; ratio
Gas exchange	VO_2, VO_2/kg, VCO_2, VCO_2/kg, RER, VE/VO_2, VE/VCO_2	mL/min; mL/kg/min; ratio
Hemodynamics	HR, VO_2/HR, SpO_2, BP_Syst, BP_Diast, HRR, CO	1/min; mL/beat; %; mmHg; L/min
Gas tensions	PaO_2, $PaCO_2$, $PetO_2$, $PetCO_2$	mmHg
Workload/protocol	Power_Load, RPM, Load_Phase	W; r/min; category
Energy expenditure	EE_Total_kcal, EE_kcal_h, Fat/CHO/PRO (kcal/h; kg-normalized; %)	kcal/h; kcal/kg/h; %
ECG features	ST_I–ST_V6; S_I–S_V6	mV
AT targets	VO2_kg_at_AT, HR_at_AT, Time_at_AT, RER_at_AT	mL/kg/min; 1/min; mm:ss; ratio
Metadata	Subject demographics, center, device, protocol	–

(`VO2_kg_at_AT`, `HR_at_AT`, `Time_at_AT`, `RER_at_AT`). Further implementation details and full variable listings are provided in the Supplement.

Table 2: Dataset composition, splits, and external-center evaluations.

Split	Centers	N exams	Usage	Standardization
Primary (train)	Zhongshan/Xuhui/Shanxi	[N_train]	Model training	Fit on primary-train
Primary (val)	Zhongshan/Xuhui/Shanxi	[N_val]	Model selection	Apply primary-train
Primary (test)	Zhongshan/Xuhui/Shanxi	[N_test]	Internal evaluation	Apply primary-train
30% per-center subset	Zhongshan/Xuhui/Shanxi	[N_30pct]	Rapid exploration only	Apply primary-train
External: Punan	Punan	[N_punan]	Unseen-center test	Apply primary-train
External: Rizhao	Rizhao	[N_rizhao]	Unseen-center test	Apply primary-train

2.3 Modeling and Evaluation

We developed *CPET-former*, a transformer-based model tailored to multi-channel CPET time-series that encodes breath-by-breath signals and aggregates sequence information to predict AT endpoints. To address multi-center generalization on known centers, we introduce center-aware Feature-wise Linear Modulation (FiLM): learned center embeddings explicitly modulate the backbone to capture site/device effects. For robustness to unseen centers, we adopt GroupDRO with centers as groups to optimize worst-center risk under distribution shift. Evaluation used MAE, RMSE, R^2, Bland–Altman, ICC, and calibration; decision utility was assessed at VO_2@AT thresholds (xx mL/kg/min) via thresh-

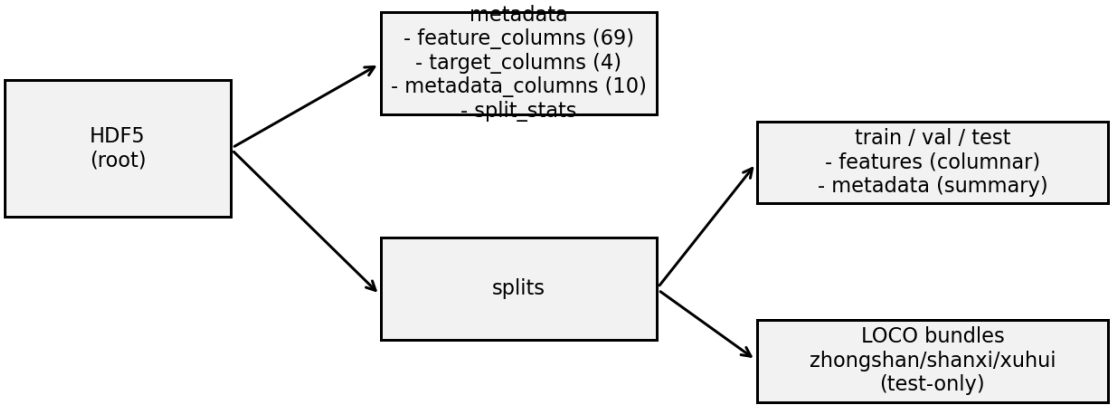

Figure 1: Overview of data sources, harmonization, and analysis splits.

old agreement, net reclassification improvement (NRI), and decision curve analysis (DCA). Implementation details, hyperparameters, interpretability analyses, and full statistical outputs are provided in the Supplement.

2.4 Reader Study

We conducted a blinded reader study to test non-inferiority of AI to senior experts for VO_2@AT estimation. Cases ([N]) were stratified by center, device, protocol duration, and difficulty; readers were grouped by seniority (junior/intermediate/senior: [J/I/S]) and received standardized training. Readers used a multi-panel interface (V-slope; VE/VO_2; VE/VCO_2; RER) and were blinded to the reference standard, to AI outputs, and to each other; a subset was re-read after a washout period ([$\geq$2 weeks]) for intra-reader reliability. The primary endpoint was MAE of VO_2@AT versus the reference standard with a prespecified non-inferiority margin δ (e.g., xx mL/kg/min); secondary endpoints included RMSE, R^2, Bland–Altman, ICC(2,1), calibration, and threshold utility at 11/14 mL/kg/min (agreement, NRI, DCA). Ethics approval and consent waiver were obtained ([IRB refs]). Full procedural details are provided in Section **Reader Study**.

3 Results

3.1 Study Cohort

The cohort comprised [N] patients (age [X] $\pm$ [Y] years; [Z%] female) across three centers and two devices. Protocols were predominantly ramp [(X%)] with median duration [T] minutes.

Study flow diagram placeholder

Insert flow chart: screened $\rightarrow$ included $\rightarrow$ analysis sets (pooled train/val/test 80%/10%/10%; two external centers as test).

Figure 2: Study flow and analysis splits.

Table 3: Baseline characteristics by center.

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

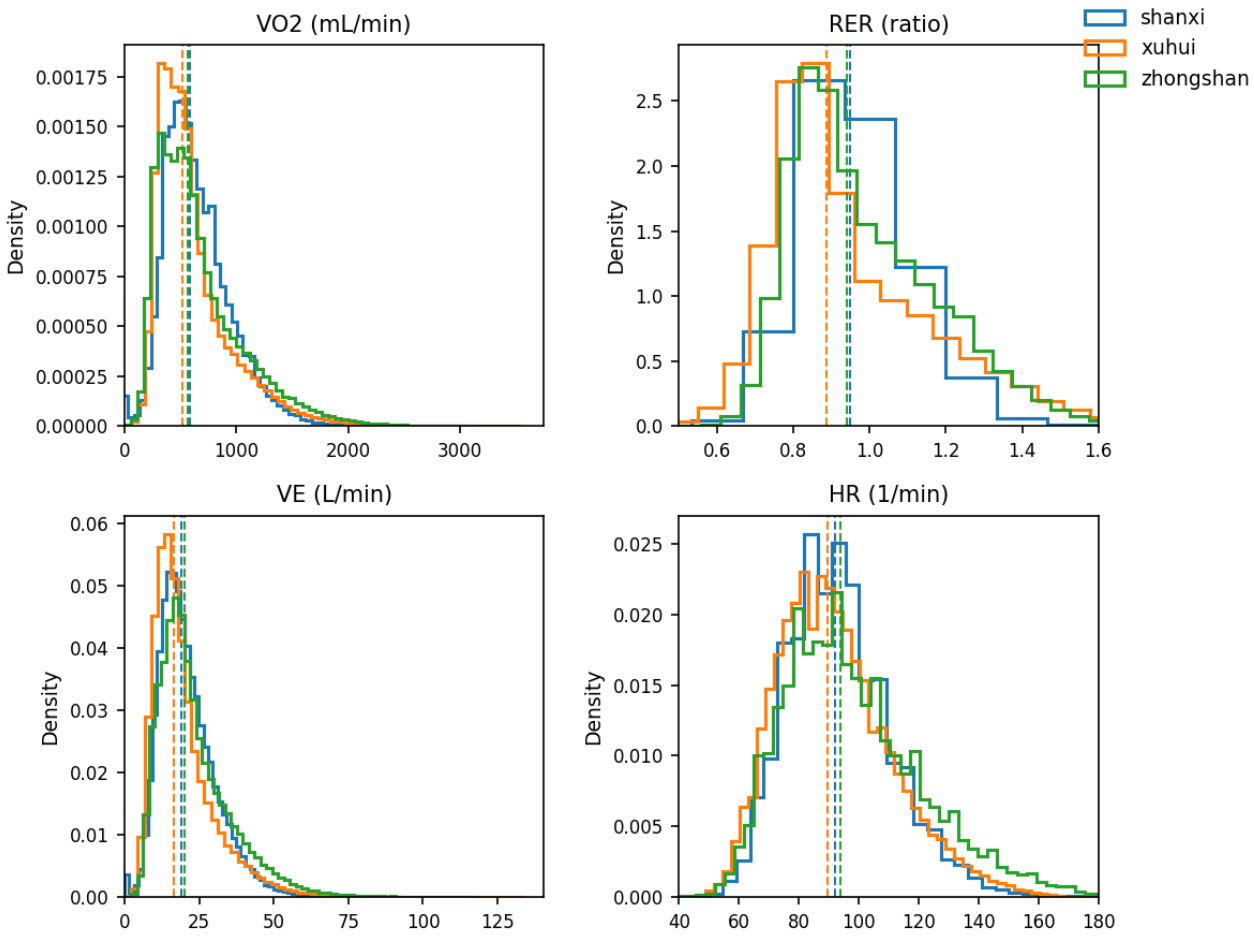

Figure 3: Distribution snapshots (by center) for VO_2, RER, VE, and HR in physical units; dashed lines indicate centre medians.

3.2 Clinical Threshold Agreement and Decision Utility

We evaluated agreement at VO_2@AT thresholds commonly used for risk stratification (e.g., 11 and 14 mL/kg/min). Outcome-oriented summaries included: (i) concordance rates and confusion matrices by threshold; (ii) net reclassification improvement (NRI) versus the reference standard; and (iii) decision curve analysis (DCA), which demonstrated positive net benefit across a wide threshold range.

Decision curve analysis (placeholder)

Insert DCA: net benefit versus threshold probability; compare AI-assisted vs. standard. Highlight clinically relevant regions.

Figure 4: Decision curve analysis for AI-assisted AT-based decision-making.

Feature set. The model consumes a comprehensive panel of ventilation, gas exchange, hemodynamic, and effort signals (e.g., VE, VO_2, VCO_2, RER, HR, Power_Load, VT, Bf), alongside a small set of derived ratios. The full feature list with units and descriptions is provided in Supplementary Tables S1A–S1H.

3.3 Model Performance and Generalization

We compared CPET-former variants against Ridge/SVR/RF/LightGBM baselines. Metrics included MAE, RMSE, and R^2 with 95% CIs; agreement was assessed via Bland–Altman and ICC.

On the pooled, stratified hold-out split, CPET-former (ERM) outperformed classical ML baselines. Centre-aware FiLM further improved performance across known centers, and on external centers, GroupDRO reduced worst-centre error and narrowed inter-centre variability (Table 4; Figure 5).

Table 4: Model performance on the pooled hold-out and external centers (mean [95% CI]).

Model	Setting	MAE	RMSE	R^2
Linear/SVR/RF/LightGBM	Pooled hold-out	[Y]	[Y]	[Y]
CPET-former (ERM)	Pooled hold-out	[X]	[X]	[X]
CPET-former (FiLM)	Pooled hold-out	[X]	[X]	[X]
CPET-former (GroupDRO)	Pooled hold-out	[X]	[X]	[X]
CPET-former (ERM)	External centers	[Y]	[Y]	[Y]
CPET-former (GroupDRO)	External centers	[X]	[X]	[X]

External-center generalization placeholder

Insert center-wise MAE/RMSE for ERM vs GroupDRO;
highlight improvement at worst/unseen center.

Figure 5: Performance on external centers by site.

3.4 Reader Study: AI vs. Clinicians

In the blinded reader study ([N] cases), AI performance matched senior experts (MAE [X] vs. [Y]; $p =$ n.s.) and exceeded junior/intermediate readers (MAE [Z]; $p < 0.001$). Inter-reader ICC was [0.80] (95% CI [0.75–0.84]), whereas AI predictions were perfectly reproducible (ICC = 1.00).

Reader study boxplots placeholder

Insert boxplots of MAE by reader seniority vs AI.

Figure 6: Reader study accuracy comparison.

Bland–Altman plots placeholder

Insert Bland–Altman plots: (a) AI vs. consensus; (b) Senior vs. consensus.

Figure 7: Agreement analyses against expert consensus.

4 Discussion

Principal findings. We collected a large multi-center, vendor-diverse CPET cohort and developed CPET-former, a transformer-based model for multi-channel CPET time-series. Center-aware FiLM improved performance across known centers by capturing site/device effects, while GroupDRO reduced worst-center error on external centers, addressing a key barrier to clinical adoption. In a blinded reader study, AI achieved senior-expert accuracy and perfect reproducibility, overcoming inherent subjectivity in manual interpretation.

Relation to prior work. Prior AI-CPET studies are typically single-center with limited validation. Transformers capture long-range temporal dependencies [9], aligning with the physiological progression of exercise. GroupDRO [12] minimizes worst-group risk, offering principled domain robustness in multi-center settings.

Strengths. (i) Scale and diversity across centers/devices; (ii) rigorous consensus ground truth; (iii) External-center validation approximating real-world practice; (iv) head-to-head comparison with clinicians; (v) interpretability analyses and QC of error cases.

Limitations. Retrospective design; limited number of centers/vendors; absence of prospective, point-of-care evaluation; demographics predominantly [region]. Future work should broaden geography and device coverage, assess clinical adoption in service delivery settings, and quantify downstream clinical impact with uncertainty-aware safety monitoring.

Clinical implications and future work. In routine clinical practice (e.g., within reporting systems or device software), this approach can standardize CPET interpretation and reduce workload. Extending to multi-task outputs (e.g., VT1/VT2, peak VO_2) and incorporating uncertainty quantification with conservative referral policies will support safe use and broader clinical adoption.

5 Conclusion

We present a generalizable AI framework for automated AT assessment that performs at senior-expert level with perfect reproducibility and robust cross-center generalization. By combining a transformer backbone with center-aware FiLM and GroupDRO, the approach addresses both known- and unseen-center variability. This enables standardized, scalable CPET interpretation in diverse clinical environments, supports standardized diagnostic pathways, and reduces clinical workload.

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]