A Multi-Center, Generalizable Deep Learning Framework for Automated Anaerobic Threshold Assessment from Cardiopulmonary Exercise Tests

A Diagnostic Accuracy and Reader Study

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

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

October 31, 2025

Abstract

Background: The anaerobic threshold (AT) from cardiopulmonary exercise testing (CPET) is a key prognostic marker for risk stratification, perioperative assessment, and rehabilitation. Manual AT determination (e.g., V-slope) is time-consuming, experience-dependent, and exhibits substantial inter- and intra-observer variability, constraining scalability and standardization.
Methods: We curated CPET-12k, a standardized multi-center dataset of [12,000] tests from three hospitals ([Zhongshan], [Shanxi], [Xuhui]) and two device vendors ([Ganshorn], [Cosmed]). Consensus ground truth was established via a two-reader protocol with blinded adjudication. We developed a transformer-based framework (*CPET-former*) and applied GroupDRO [12] for domain generalization. Generalization was assessed via mixed cross-validation and leave-one-center-out (LOCO). We further conducted a blinded reader study with [12] clinicians to compare AI against human readers across seniority.
Findings: CPET-former achieved strong internal accuracy for AT (MAE [X]; R^2 [X]). Under LOCO, GroupDRO improved performance at unseen centers (MAE [X] vs. [Y]; $p < 0.01$). In the reader study, AI matched senior experts (MAE [X] vs. [Y]; p =n.s.) and exceeded junior/intermediate readers (MAE [Z]; $p < 0.001$). Inter-reader reliability was moderate (ICC [0.78]), while AI predictions were perfectly reproducible (ICC $= 1.00$). A self-supervised variant achieved [˜95%] of full-data performance using [10%] labels.
Interpretation: We present a generalizable, objective, and highly consistent AI framework for automated AT assessment. The model demonstrates robust real-world performance, including at unseen centers, and achieves senior-expert-level accuracy with perfect reproducibility, enabling standardized CPET interpretation and reduced clinical burden.

1 Introduction

Cardiopulmonary exercise testing (CPET) integrates cardiovascular, pulmonary, and metabolic responses to exertion [1]. Among CPET-derived metrics, the anaerobic threshold (AT) is central to prognosis and decision-making in heart failure, perioperative risk, and rehabilitation [2, 3]. Conventional AT determination via visual inspection (e.g., V-slope) [4] is subjective and labor-intensive. Inter- and intra-observer agreement can be modest even among experienced clinicians [5], compromising reproducibility and constraining scale.

Traditional automated approaches (e.g., curve-fitting) are sensitive to noise and protocol variability and often fail to generalize across vendors and clinical settings. Prior machine learning studies [7, 8] are typically single-center and small-scale, with limited validation on unseen centers. Robust generalization and head-to-head comparison with clinicians remain underexplored.

*Corresponding author: email@address.com

We address this gap with a multi-center, generalizable AI framework for automated AT assessment. Our aims are to deliver: (i) a large, standardized dataset (CPET-12k) spanning three centers and two device vendors; (ii) a transformer-based model (CPET-former) tailored to CPET time-series; (iii) domain generalization via GroupDRO to improve worst-center performance; and (iv) a blinded reader study benchmarking AI against clinicians. Our hypothesis is that AI accuracy and consistency are non-inferior to senior experts.

2 Methods

2.1 Study Design and Ethics

We conducted a multi-center, retrospective diagnostic accuracy study and a prospective-simulated blinded reader study. Institutional review board approvals were obtained at [Zhongshan], [Shanxi], and [Xuhui] with consent waived.

2.2 Dataset and Standardization (CPET-12k)

CPET-12k comprises 12,829 ramp-protocol examinations from three hospitals (*shanxi*, *xuhui*, *zhongshan*) and two vendors (Ganshorn, COSMED). Adults ([>18] years) who completed maximal or symptom-limited CPET with acceptable quality per ATS/ACCP [14] were included; we excluded incomplete files, protocol deviations, and technical artifacts.

Acquisition pipeline. Raw vendor spreadsheets are converted to a unified CPET standard (v1.4) using a CLI (`cpetx-data extract`). Vendor-specific extractors (COSMED, Ganshorn) are driven by YAML mappings and a common schema: device columns and summary cells are mapped to standard names/units, then written to per-center HDF5 files (`cpet_data_source_<center>.h5`). The extractor records institute metadata and logs, and de-duplicates examinations by a stable signature (`Subject_ID`, `Examination_Date`, summary digest).

Standardization layer. The CPET schema enumerates time-series (breath-by-breath) fields, summary targets, and metadata types. Mappings apply unit conversions and normalization rules, for example: liters to mL/min for VO_2/VCO_2 (Ganshorn), hPa to mmHg for barometric pressure, categorical encoding for gender, normalization of date/time strings, and mapping of phase markers to `Load_Phase`. COSMED/Ganshorn adapters are specified in YAML and versioned with the schema, ensuring traceable harmonization across devices.

Cleaning and preprocessing. From per-center HDF5 sources, we construct analysis-ready frames via: (i) optional time aggregation when raw sampling exceeds the configured interval (default seconds-scale) with robust averaging; (ii) per-examination linear interpolation and forward/backward fill for numeric channels; (iii) optional filters (time window, test phases, exercise-first minutes, and physiologic ranges for HR/VO_2); (iv) target integrity checks that drop examinations missing non-AT targets while permitting right-censored `Time_at_AT` when used only as auxiliary supervision. These steps are deterministic given the same inputs and seeds.

Dataset generation and splits. Processed center-wise frames are mixed with ‘cpetx-data generate‘ into HDF5 datasets. We support: (a) *standard* splits that stratify examinations into train/val/test; (b) explicit center splits where train/val centers and a held-out test center are provided; and (c) leave-one-center-out (LOCO), producing one dataset per held-out center. Standardization uses a scaler fitted on the training split for user-specified feature columns and applied to val/test; scaler parameters (mean/scale) and the list of standardized features are stored in split metadata for inversion and audit.

On-disk layout. Final datasets contain two top-level groups: ‘metadata‘ and ‘splits‘. The ‘metadata‘ group lists column categories (feature/target/metadata), split statistics (per split and per center), examination-to-center mapping, and standardization parameters. The ‘splits‘ group holds ‘train‘, ‘val‘, and ‘test‘, each with columnar ‘features‘ and a ‘metadata‘ subgroup. Features include breath-by-breath signals such as ‘VE‘, ‘VO2‘, ‘VCO2‘, ‘RER‘, ‘HR‘, ‘Power_Load‘, ‘VT‘, ‘Bf‘, and derived ratios (total [69] channels). Targets include ‘VO2_kg_at_AT‘, ‘HR_at_AT‘, ‘Time_at_AT‘, and ‘RER_at_AT‘. Representative split sizes (examinations): train 10,262; val 1,282; test 1,285 (total 12,829). Under LOCO, we release three folds with held-out *shanxi* (8,785), *xuhui* (2,411), and *zhongshan* (1,633).

Releases. We provide the full mixed dataset and derived variants: 10% subsamples for ablations (‘$*_{s}mall‘), self-supervisedpretraining(‘*_{s}sl‘), explicittrain/valcentersplitswithanexternaltestset(‘*_{e}$ $xplicit\_split‘), andLOCOfolds(‘*_{l}oco‘)$.

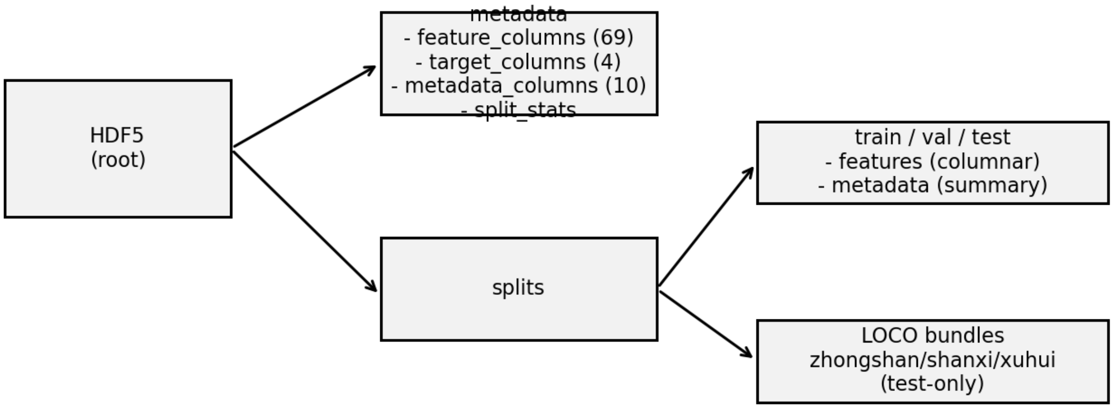

Figure 1: Data packaging and split structure for CPET-12k.

2.3 Reference Standard (Expert Consensus)

Ground truth AT was defined by a two-reader consensus process. Each case was independently annotated using V-slope corroborated by VE/VCO$_2$ nadir and ventilatory equivalents. Disagreements were resolved by blinded adjudication. Outlier and quality-control checks enforced physiological plausibility and protocol consistency.

2.4 Model: CPET-former

We evaluated a family of models for predicting the anaerobic threshold (AT) with VO$_2$@AT as the primary endpoint, and optionally auxiliary targets (e.g., Time@AT, HR@AT).

Classical ML baselines. We include Ridge, SVR, Random Forest, and LightGBM. Time-series are summarized into tabular features (statistical moments, temporal trends/slopes, domain features such as peaks/threshold crossings) together with demographics, and regressed to VO$_2$@AT. These serve as point-of-reference comparators.

Baseline (cpet_former). A transformer encoder over breath-by-breath sequences with sinusoidal positional encoding and multi-head self-attention. The backbone projects input channels (e.g., VE, VO$_2$, VCO$_2$, RER, HR, Power_Load, VT, Bf, time) to `d_model`, stacks L encoder blocks, and aggregates token features by masked mean pooling. A regression head outputs VO$_2$@AT (and optionally Time@AT via an auxiliary head/loss). This is our empirical risk minimization (ERM) baseline on mixed-center data.

Center-aware FiLM (cpet_former_center_film). To improve robustness on known centers, we inject centre embeddings via Feature-wise Linear Modulation: a learned centre vector adds a shallow bias before the encoder and generates per-layer scale/shift (FiLM) to modulate token features. This targets improved performance on the three-center mixed dataset without sacrificing overall accuracy.

GroupDRO variant (cpet_former_v1). To improve worst-centre performance under leave-one-center-out (LOCO), we adopt GroupDRO with centres as groups. Training uses per-sample MSE aggregated by a group-weighted objective with momentum-updated group losses and temperature-controlled softmax weighting, emphasizing the currently underperforming centre.

Self-supervised pretraining (cpet_former_v2). We pretrain the transformer backbone on unlabeled sequences via masked reconstruction (optionally with a contrastive projection head), then fine-tune a lightweight prediction head with few-shot labels. Backbones can be frozen or partially unfrozen during fine-tuning; pretrained weights are exportable/importable for reproducibility. This aims to retain strong accuracy with 5–10% labeled data in both mixed-CV and LOCO.

Multi-target variant. A multi-output head (shared backbone) predicts multiple CPET endpoints jointly, e.g., VO_2@AT, Time@AT, HR@AT, demonstrating that the architecture generalizes across related tasks. The trainer supports weighted auxiliary losses and head-wise diagnostics.

2.5 Training and Domain Generalization

Supervision uses MSE on VO_2@AT as the primary loss; when enabled, Time@AT and HR@AT receive auxiliary losses with fixed weights. Optimization uses AdamW with cosine schedule and warmup; regularization includes dropout, segment mixup, and mild time-warp. For domain robustness: (i) *FiLM* conditions the backbone on centre IDs for known-centre training; (ii) *GroupDRO* reweights centre-wise losses to minimize worst-centre risk under distribution shift [12]. For *SSL pretraining*, we mask valid timesteps and reconstruct inputs (with optional contrastive alignment between augmented views), then fine-tune the prediction head with frozen or partially unfrozen backbone.

2.6 Evaluation Protocols

We performed (i) stratified mixed k-fold cross-validation and (ii) leave-one-center-out (LOCO). Metrics included MAE, RMSE, R^2, and agreement analyses (Bland–Altman). Statistical testing used paired comparisons with Bonferroni correction. Calibration and failure modes were examined via error versus RER at AT and protocol duration.

2.7 Blinded Reader Study

[N] cases were sampled to span diverse protocols and demographics. Readers (junior, intermediate, and senior) independently estimated AT using standardized software without access to ground truth or each other's labels.
After a washout period, a subset was repeated to evaluate intra-reader reliability. AI predictions were generated once per case without manual tuning. Agreement was quantified via ICC [13] and Bland–Altman.

2.8 Interpretability and Quality Control

We probed attention maps around predicted AT and inspected high-error cases for physiologic plausibility. We report representative cases and failure modes to guide clinical integration.

3 Results

3.1 Study Cohort

The cohort comprised [N] patients (age [X] ± [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 → included → analysis sets (train/val/test; LOCO folds).

Figure 2: Study flow and analysis splits.

Table 1: 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

3.2 Model Performance and Generalization

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

Across mixed cross-validation, CPET-former outperformed traditional ML models (Table 3). Under LOCO, GroupDRO improved performance at unseen centers, reducing worst-center error and narrowing inter-center variability (Figure 3).

LOCO generalization placeholder

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

Figure 3: LOCO performance by center.

3.3 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).

3.4 Secondary Analyses

A self-supervised variant ([CPET-former-SSL]) achieved [˜95%] of full-supervision performance using [10%] labels (Figure 6); multi-task extensions are reported in the Supplement.

4 Discussion

Principal findings. We built a large, standardized multi-center CPET dataset and developed an AI framework that achieves accurate, objective, and reproducible AT assessment. GroupDRO markedly improved generalization to unseen centers under LOCO, addressing a key barrier to clinical deployment. 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.

Table 2: Feature Units and Descriptions (S3).

Feature	Unit	Description
Load_Phase	category	Exercise phase code (e.g., 0:Mainload, 1:Preload, 2:Postload).
Bf	1/min	Breath Frequency.
BR_pct	VT	L
Tidal Volume, the volume of air moved per breath (BTPS).		
VE	L/min	Minute Ventilation.
Ti	s	Inspiratory Time.
Te	s	Expiratory Time.
Ttot	s	Total Breath Time (Ti + Te).
Ti_Ttot_Ratio	ratio	Inspiratory duty cycle.
VD_VT_Ratio	ratio	Physiological Dead Space to Tidal Volume Ratio (VD/VT).
VT_Ti	L/s	Mean Inspiratory Flow.
VO2	mL/min	Oxygen consumption.
VO2_kg	mL/kg/min	Oxygen consumption per kilogram of body weight.
VCO2	mL/min	Carbon dioxide production.
VCO2_kg	mL/kg/min	Carbon dioxide production per kilogram of body weight.
RER	ratio	Respiratory Exchange Ratio (VCO2/VO2), also known as RQ.
PaCO2_est	mmHg	Estimated partial pressure of arterial CO2.
VE_VO2	ratio	Ventilatory equivalent for oxygen.
VE_VCO2	ratio	Ventilatory equivalent for carbon dioxide.
METS	MET	Metabolic equivalents.
HR	1/min	Heart rate in beats per minute.
VO2_HR	mL/beat	Oxygen Pulse (VO2/HR).
SpO2	BP_Syst	mmHg
Systolic Blood Pressure.		
BP_Diast	mmHg	Diastolic Blood Pressure.
HRR	1/min	Heart Rate Recovery.
CO	L/min	Cardiac Output.
PaO2	mmHg	Partial pressure of arterial O2.
PaCO2	mmHg	Partial pressure of arterial CO2.
PetO2	mmHg	End-tidal partial pressure of O2 (PETO2).
PetCO2	mmHg	End-tidal partial pressure of CO2 (PETCO2).
Power_Load	W	Workload or power output from the ergometer.
RPM	r/min	Revolutions Per Minute or cadence.
EE_Total_kcal	kcal/h	Energy expenditure.
EE_kcal_h	kcal/h	Energy expenditure per hour.
Fat_kcal_h	kcal/h	Fat energy expenditure.
CHO_kcal_h	kcal/h	Carbohydrate energy expenditure.
PRO_kcal_h	kcal/h	Protein energy expenditure.
EE_kg_kcal_h	kcal/kg/h	Energy expenditure per kilogram of body weight.
Fat_kg_kcal_h	kcal/kg/h	Fat energy expenditure per kilogram of body weight.
CHO_kg_kcal_h	kcal/kg/h	Carbohydrate energy expenditure per kilogram of body weight.
PRO_kg_kcal_h	kcal/kg/h	Protein energy expenditure per kilogram of body weight.
Fat_pct	CHO_pct	PRO_pct
ST_I	mV	ST-segment depression/elevation in lead I.
ST_II	mV	ST-segment depression/elevation in lead II.
ST_III	mV	ST-segment depression/elevation in lead III.
ST_aVR	mV	ST-segment depression/elevation in lead aVR.
ST_aVL	mV	ST-segment depression/elevation in lead aVL.
ST_aVF	mV	ST-segment depression/elevation in lead aVF.
ST_V1	mV	ST-segment depression/elevation in lead V1.
ST_V2	mV	ST-segment depression/elevation in lead V2.
ST_V3	mV	ST-segment depression/elevation in lead V3.
ST_V4	mV	ST-segment depression/elevation in lead V4.
ST_V5	mV	ST-segment depression/elevation in lead V5.
ST_V6	mV	ST-segment depression/elevation in lead V6.
S_I	mV	S-wave amplitude in lead I.
S_II	mV	S-wave amplitude in lead II.
S_III	mV	S-wave amplitude in lead III.
S_aVR	mV	S-wave amplitude in lead aVR.
S_aVL	mV	S-wave amplitude in lead aVL.
S_aVF	mV	S-wave amplitude in lead aVF.
S_V1	mV	S-wave amplitude in lead V1.
S_V2	mV	S-wave amplitude in lead V2.
S_V3	mV	S-wave amplitude in lead V3.
S_V4	mV	S-wave amplitude in lead V4.
S_V5	mV	S-wave amplitude in lead V5.
S_V6	mV	S-wave amplitude in lead V6.

Table 3: Model performance in mixed CV and LOCO (mean [95% CI]).

Model	Setting	MAE	RMSE	R^2
Linear/SVR/RF/LightGBM	Mixed-CV	[Y]	[Y]	[Y]
CPET-former (ERM)	Mixed-CV	[X]	[X]	[X]
CPET-former (GroupDRO)	Mixed-CV	[X]	[X]	[X]
CPET-former (ERM)	LOCO	[Y]	[Y]	[Y]
CPET-former (GroupDRO)	LOCO	[X]	[X]	[X]

Reader study boxplots placeholder

Insert boxplots of MAE by reader seniority vs AI.

Figure 4: Reader study accuracy comparison.

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

Limitations. Retrospective design; limited number of centers/vendors; lack of real-time (online) validation; demographics predominantly [region]. Future work should expand geography and device coverage, evaluate online inference, and examine downstream clinical impact.

Clinical implications and future work. Embedding the model into clinical workflows (EMR or device software) can standardize CPET interpretation and reduce workload. Extending to multi-task outputs (e.g., VT1/VT2, peak VO_2) and adding uncertainty quantification will broaden utility and support safe 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. This enables standardized, scalable CPET interpretation in diverse clinical environments.

Bland–Altman plots placeholder

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

Figure 5: Agreement analyses against expert consensus.

SSL learning curve placeholder

Insert learning curve: labeled fraction vs. MAE/R^2.

Figure 6: Label efficiency via self-supervision.

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 S1. CPET Data Specification (overview). The dataset contains 69 breath-by-breath feature channels grouped as follows: (i) Ventilation and gas exchange: VE, VO_2, VCO_2, VO_2/kg, VE/VO_2, VE/VCO_2, VO_2/HR; (ii) Respiratory timing and mechanics: Bf, VT, Ti, Te, Ttot, Ti/Ttot, VD/VT; (iii) Gas tensions: $PetO_2$, $PetCO_2$, PaO_2, $PaCO_2$, $PaCO_2$ (est.); (iv) Workload/protocol: Power_Load, RPM, Load_Phase; (v) Energy expenditure/substrate split: METS, EE (kcal/h; kg-normalized; total), CHO/Fat/PRO (kcal/h, kg-normalized, %). ECG-derived signals include HR, HRR, and ST/S amplitudes across standard leads (I/II/III, aVR/aVL/aVF, V1–V6). Units are harmonized to L/min (VE), mL/min (VO_2, VCO_2), mL/kg/min (VO_2/kg), bpm (HR), breaths/min (Bf), L (VT), W (Power_Load), and unitless ratios for RER and ventilatory equivalents.

Supplementary Table S2. Targets and metadata. Targets: VO_2/kg at AT (mL/kg/min), HR at AT (bpm), Time at AT (s), RER at AT (unitless). Key metadata: Examination_ID, Subject_ID, Time (timestamp), Gender, Age (years), Height_cm, Weight_kg, Source_Device (Ganshorn/COSMED), Institute_Name (Shanxi/Xuhui/Zhongshan).

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).

Table 4: Full list of 69 feature columns.

Feature	Feature	Feature
Load_Phase	BP_Syst	ST_II
Bf	BP_Diast	ST_III
BR_pct	HRR	ST_aVR
VT	CO	ST_aVL
VE	PaO2	ST_aVF
Ti	PaCO2	ST_V1
Te	PetO2	ST_V2
Ttot	PetCO2	ST_V3
Ti_Ttot_Ratio	Power_Load	ST_V4
VD_VT_Ratio	RPM	ST_V5
VT_Ti	EE_Total_kcal	ST_V6
VO2	EE_kcal_h	S_I
VO2_kg	Fat_kcal_h	S_II
VCO2	CHO_kcal_h	S_III
VCO2_kg	PRO_kcal_h	S_aVR
RER	EE_kg_kcal_h	S_aVL
PaCO2_est	Fat_kg_kcal_h	S_aVF
VE_VO2	CHO_kg_kcal_h	S_V1
VE_VCO2	PRO_kg_kcal_h	S_V2
METS	Fat_pct	S_V3
HR	CHO_pct	S_V4
VO2_HR	PRO_pct	S_V5
SpO2	ST_I	S_V6

[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]