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) derived from cardiopulmonary exercise testing (CPET) is a key prognostic marker used for risk stratification, perioperative evaluation, and rehabilitation planning. Manual AT determination (e.g., V-slope) is time-consuming, experience-dependent, and exhibits substantial inter-observer variability, limiting scalability and standardization.
Methods: We assembled CPET-12k, a large, standardized, multi-center dataset comprising [12,000] tests from three hospitals ([Zhongshan], [Shanxi], [Xuhui]) and two device vendors ([Ganshorn], [Cosmed]). Expert consensus ground truth was established via a two-reader protocol with blinded adjudication. We developed a transformer-based framework (*CPET-former*) and incorporated domain generalization using GroupDRO [12]. Generalization was evaluated using mixed cross-validation and leave-one-center-out (LOCO). Finally, we conducted a blinded reader study with [12] clinicians across experience levels to compare AI performance with human readers.
Findings: Internally, CPET-former achieved strong accuracy for AT prediction (MAE [X] mL/kg/min; R^2 [X]). Under LOCO, GroupDRO substantially improved performance at unseen centers (MAE [X] vs. [Y]; $p < 0.01$). In the reader study, AI performance 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 only [10%] labeled data.
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) provides an integrated assessment of cardiovascular, pulmonary, and metabolic responses to exertion [1]. Among CPET-derived metrics, the anaerobic threshold (AT) is central to risk stratification, perioperative evaluation, and rehabilitation planning [2, 3]. However, conventional AT determination via visual inspection (e.g., V-slope) [4] is subjective and labor-intensive. Inter-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 efforts [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 contributions are: (i) a large, standardized dataset (CPET-12k) spanning three centers and two vendors; (ii) a transformer-based model (CPET-former) tailored to physiological time-series; (iii) domain generalization via GroupDRO to improve performance at unseen centers; (iv) rigorous evaluation using leave-one-center-out (LOCO) validation; and (v) a blinded reader study comparing AI performance with clinicians across seniority levels. We hypothesized that AI accuracy and consistency are non-inferior to, and potentially exceed, expert performance.

2 Methods

2.1 Study Design and Ethics

We performed a multi-center, retrospective diagnostic accuracy study with a prospective (simulated) blinded reader study. Institutional review board approvals were obtained from [Zhongshan, Shanxi, Xuhui] with waiver of informed consent for retrospective data use.

2.2 Dataset: CPET-12k and Standardization

We aggregated 12,829 cardiopulmonary exercise tests (CPETs) from three centers (shanxi: n=8,785; xuhui: n=2,411; zhongshan: n=1,633) using two device vendors (Ganshorn, COSMED). Inclusion criteria were age ¿ 18 years and a complete incremental CPET; exclusions were protocol deviations or poor data quality. A unified CPET data specification harmonized variable names, units, and derived variables. Time-series were downsampled to 10-second intervals (rows per examination: median 67; p90 106; max 162). The final matrix contains 82 variables per time step, including ventilatory gas exchange, workload/protocol markers, ECG-derived summaries, and static demographics/identifiers. Four AT-related targets were included: `VO2_kg_at_AT`, `HR_at_AT`, `Time_at_AT`, and `RER_at_AT`. Continuous variables (n=75) were standardized using training-split statistics (StandardScaler; train n_rows=741,799) and the same transform was applied to validation/test; identifiers/categorical fields and Time/Load_Phase were not standardized. We used an 80/10/10 examination-wise split (train/val/test = 10,262/1,282/1,285 examinations; 741,799/92,211/92,638 rows). Additionally, we constructed leave-one-center-out (LOCO) variants with each center held out for testing.

2.3 Reference Standard: Expert Consensus

Two independent clinicians (blinded to each other and AI) annotated AT using established CPET criteria (V-slope, ventilatory equivalents, end-tidal gas trends). Discrepancies exceeding [1.0 mL/kg/min] or qualitative disagreement triggered adjudication by a third senior expert to form the consensus ground truth. Reader identities, timestamps, and rationales were logged for audit.

2.4 Model: CPET-former and Domain Generalization

CPET-former is a transformer-based architecture designed for multivariate physiological sequences, ingesting synchronized breath-by-breath features and protocol markers. We trained a baseline empirical risk minimization (ERM) model and a GroupDRO variant [12], defining groups by center and device to mitigate worst-group errors and enhance cross-site robustness.

2.5 Evaluation Protocols

We conducted (i) mixed cross-validation (stratified by center/device) and (ii) leave-one-center-out (LOCO) validation to simulate deployment at unseen centers. Primary endpoints were MAE (mL/kg/min) and R^2 for AT prediction. Secondary endpoints included RMSE, Pearson r, and calibration slope/intercept.

2.6 Reader Study

We recruited [12] clinicians ([4] junior, [4] intermediate, [4] senior) to estimate AT on a stratified sample of [N] cases. Readers were blinded to AI and consensus labels and followed a standardized

reading protocol. Outcomes were reader MAE vs. consensus and inter-reader reliability (ICC).

2.7 Statistical Analysis

We report means with 95% CIs using nonparametric bootstrap ([10,000] replicates). Group comparisons used paired tests with Bonferroni correction where applicable. Inter-reader reliability used a two-way random-effects, absolute-agreement ICC (ICC[2,1]) [13]. Significance threshold was $p < 0.05$.

3 Results

3.1 Study Cohort

The final cohort included 12,829 examinations after screening, spanning three centers (shanxi: 8,785; xuhui: 2,411; zhongshan: 1,633). Device vendors were center-specific (shanxi: Ganshorn; xuhui/zhongshan: COSMED). Cohort-level summary: age median 60 years (IQR 51–66) and BMI median 24.7 kg/m^2 (IQR 22.6–27.0). At AT, VO_2/kg was 12.5 mL/kg/min (IQR 10.6–14.9), HR 107 bpm (IQR 96–118), time 430 s (IQR 350–547), and RER 0.98 (IQR 0.93–1.03). Baseline characteristics by center are shown in Table 1.

Flow chart placeholder

Insert flow-chart of study cohort: screened [N], excluded [reasons], final [12,000].

Figure 1: Study flow diagram.

Table 1: Baseline characteristics by center. Data are median (IQR) or n (%).

	Center A ([Zhongshan])	Center B ([Shanxi])	Center C ([Xuhui])
N	[N_A]	[N_B]	[N_C]
Age (years)	[X]	[Y]	[Z]
Female, n (%)	[X]	[Y]	[Z]
BMI (kg/m^2)	[X]	[Y]	[Z]
Peak VO_2 (mL/kg/min)	[X]	[Y]	[Z]
Device (Ganshorn/Cosmed, %)	[X/Y]	[X/Y]	[X/Y]

3.2 Model Performance and Generalization

Across mixed cross-validation, CPET-former outperformed traditional ML models (Table 2). Under LOCO, GroupDRO improved performance at unseen centers, reducing worst-center error and narrowing inter-center variability (Figure 2).

Table 2: 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]

LOCO generalization placeholder

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

Figure 2: 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).

Reader study boxplots placeholder

Insert boxplots of MAE by reader seniority vs AI.

Figure 3: Reader study accuracy comparison.

Bland–Altman plots placeholder

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

Figure 4: Agreement analyses against expert consensus.

3.4 Secondary Analyses

A self-supervised variant ([CPET-former-SSL]) achieved [˜95%] of full-supervision performance using [10%] labels (Figure 5); 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 CPET automation studies are typically single-center with limited validation. Transformers capture long-range temporal dependencies [9] and align with the physiological progression of exercise. GroupDRO [12] explicitly minimizes worst-group risk, offering a principled route to domain robustness in multi-center settings.

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.

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.

SSL learning curve placeholder

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

Figure 5: Label efficiency via self-supervision.

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.

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/or former].

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]