

Breaking the Data Bottleneck in Exercise Health AI

Xiangxun Lai, Yue Lai, Qi Gao, caihua huang

Submitted to: Journal of Medical Internet Research
on: April 12, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 12

 Figures 13

 Figure 1..... 14



Breaking the Data Bottleneck in Exercise Health AI

Xiangxun Lai^{1*} PhD; Yue Lai^{2*} BSc; Qi Gao¹ Prof Dr; caihua huang³ Prof Dr

¹ Beijing Sport University Beijing CN

² Jimei University Xiamen CN

³ Xiamen University of Technology Xiamen CN

*these authors contributed equally

Corresponding Author:

caihua huang Prof Dr

Xiamen University of Technology

Xiamen University of Technology, 600 Ligong Road, Jimei District, Xiamen, China

Xiamen

CN

Abstract

Artificial intelligence (AI) holds significant potential to transform exercise health through applications like personalized training, injury prevention, and targeted rehabilitation. However, the scarcity of high-quality, real-world data—due to issues like data fragmentation, inconsistent formats, privacy constraints, and confidentiality concerns—severely restricts AI development and deployment in this domain. Learning from advancements in healthcare AI, strategies such as federated learning for privacy-preserving collaboration, synthetic data generation to augment limited datasets, transfer learning to leverage existing models, and structured data-sharing protocols offer viable solutions. Furthermore, an innovative closed-loop data generation process is proposed: leveraging large language model APIs to create extensive synthetic datasets from initial high-quality inputs, followed by rigorous human expert validation and refinement through real-world feedback. This paper argues that overcoming the data bottleneck necessitates an integrated strategy combining these methods. Implementing such comprehensive measures will enhance AI model robustness and generalizability, ultimately accelerating the delivery of tangible benefits to athletes and health practitioners.

(JMIR Preprints 12/04/2025:75895)

DOI: <https://doi.org/10.2196/preprints.75895>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, my title and abstract will remain visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Title Page

Breaking the Data Bottleneck in Exercise Health AI

Authors

Xiangxun Lai^{1,2*}, Yue Lai^{3*}, Qi Gao², Caihua Huang^{1†}

Authors details

1. Research and Communication Center for Exercise and Health, Xiamen University of Technology, Xiamen, Fujian Province, China

2. School of Sport Medicine and Rehabilitation, Beijing Sport University, Beijing, China

3. A Department of Mathematics and Digital Science, Chengyi College, Jimei University, Xiamen, China

*Equal contribution, ordered alphabetically. W

†Corresponding author: Caihua Huang

Address: Xiamen University of Technology, 600 Ligong Road, Jimei District, Xiamen, China

Zip Code: 361024

Tel: +86 15606951380

Email: caihua.huang@foxmail.com

Artificial intelligence (AI) holds significant potential to transform exercise health through applications like personalized training, injury prevention, and targeted rehabilitation. However, the scarcity of high-quality, real-world data—due to issues like data fragmentation, inconsistent formats, privacy constraints, and confidentiality concerns—severely restricts AI development and deployment in this domain. Learning from advancements in healthcare AI, strategies such as federated learning for privacy-preserving collaboration, synthetic data generation to augment limited datasets, transfer learning to leverage existing models, and structured data-sharing protocols offer viable solutions. Furthermore, an innovative closed-loop data generation process is proposed: leveraging large language model APIs to create extensive synthetic datasets from initial high-quality inputs, followed by rigorous human expert validation and refinement through real-world feedback. This paper argues that overcoming the data bottleneck necessitates an integrated strategy combining these methods. Implementing such comprehensive measures will enhance AI model robustness and generalizability, ultimately accelerating the delivery of tangible benefits to athletes and health practitioners.

Keywords: Artificial Intelligence (AI); Data Science; Exercise Health; Exercise Therapy;

Background: the challenge of data scarcity and fragmentation

Artificial intelligence (AI) offers transformative potential in exercise health applications, including personalized training regimens, injury prevention strategies, and precision rehabilitation programs [1]. However, progress in this domain is significantly hindered by limited access to high-quality, real-world data [2,3]. The fragmentation of datasets, inconsistent data formats, stringent privacy protections, and competitive confidentiality are major challenges constraining AI model development and generalization [4,5].

Translating healthcare AI innovations to exercise health

The healthcare sector provides valuable insights for addressing similar data-related challenges. Recent advances in medical AI demonstrate effective approaches such as federated learning, synthetic data generation, transfer learning, and structured data-sharing protocols [6]. Federated learning enables multiple institutions to collaboratively train robust AI models while preserving data privacy by sharing only aggregated model parameters rather than raw data. Synthetic data generation, leveraging advanced generative AI models, has emerged as a powerful method to supplement scarce real-world datasets, thereby significantly improving model accuracy and generalization. Transfer learning also allows leveraging existing AI models developed for related tasks to expedite model development in target applications. Moreover, structured data-sharing frameworks enable secure collaborations among institutions, facilitating broader data access without compromising privacy.

Integration of weak supervision in exercise health AI data generation

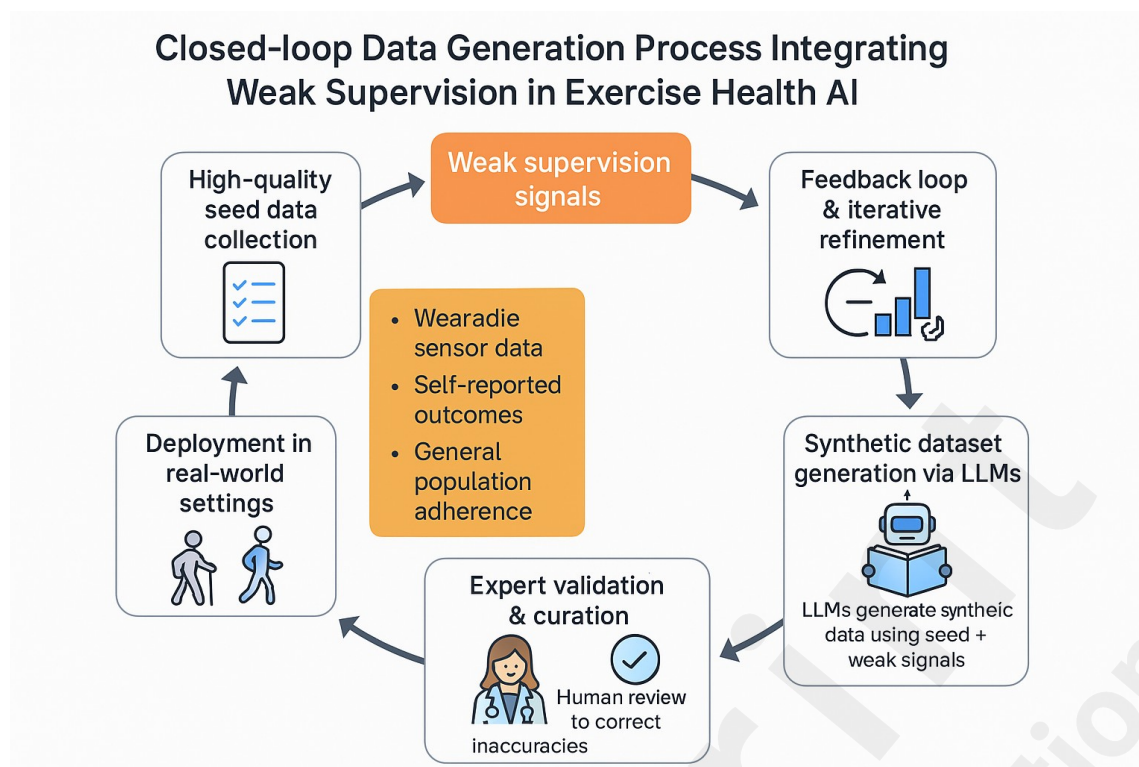


Figure1. Integrates weak supervision signals, guides LLM-based synthetic data creation, includes expert validation, real-world deployment, and continuous refinement for improved model performance..

An innovative and particularly promising approach for exercise health AI utilizes a closed-loop data generation process (Figure 1). This commences with the collection of limited but high-quality real-world data, such as clinically validated training regimens or rehabilitation plans. However, acquiring sufficient high-fidelity, expert-annotated data is often resource-intensive. In such scenarios, weak supervision-learning approaches become highly relevant [7,8]. These techniques leverage more abundant yet potentially noisy, limited, or imprecise data sources—for instance, user self-reported outcomes, unlabeled data from wearable sensors, or general population exercise adherence patterns—as cost-effective training signals when precise ground truth is scarce. This weakly supervised information, potentially complementing the initial high-quality seed data, can guide powerful large language models in generating extensive synthetic datasets. Subsequently, human experts rigorously validate and curate these generated datasets to mitigate inaccuracies and inconsistencies [9]. Following deployment in practical settings, real-world feedback and iterative adjustments continuously refine and enhance data quality. This integrated methodology, strategically incorporating weak supervision where appropriate, yields highly reliable and generalizable datasets essential for robust AI model training and application in personalized exercise science and health.

In conclusion, overcoming the data bottleneck in exercise health AI requires an integrated strategy combining

standardized data collection practices, federated learning, synthetic data generation, transfer learning, and structured data-sharing methods. Adopting these comprehensive measures can substantially enhance the effectiveness and applicability of AI technologies, accelerating advancements and delivering tangible benefits to athletes and health practitioners.



REFERENCES

1. Sun T, Xu Y, Xie H, Ma Z, Wang Y. Intelligent personalized exercise prescription based on an eHealth promotion system to improve health outcomes of middle-aged and older adult community dwellers: pretest–posttest study. *J Med Internet Res*. 2021;23(5):e28221. doi:10.2196/28221.
2. Mohsen F, Al-Absi HRH, Yousri NA, El Hajj N, Shah Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digit Med*. 2023;6(1):141. doi:10.1038/s41746-023-00933-5.
3. Li H, Jiang Z, Guan Z, et al. Large language models for diabetes training: a prospective study. *Sci Bull*. 2025. doi:10.1016/j.scib.2025.01.034.
4. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligence applications. *npj Digit Med*. 2023;6(1):116. doi:10.1038/s41746-023-00834-7.
5. Gao S, Zhu R, Kong Z, et al. TxAgent: an AI agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970*. Published March 14, 2025.
6. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. doi:10.1001/jama.2024.21700.
7. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci*. 12, 7866–7881 (2021).
8. Janet, J. P. et al. Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization. *ACS Cent. Sci*. 6, 513–524 (2020).
9. Šuvalov H, Lepson M, Kukk V, et al. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. *J Med Internet Res*. 2025;27:e66279. doi:10.2196/66279.

Acknowledgements

None.

Author contributions

Xiangxun Lai and Caihua Huang conceptualized the **viewpoint**. Xiangxun Lai wrote the initial draft.

Xiangxun Lai, Yue Lai, Qi Gao, and Caihua Huang contributed to the first draft and provided critical

revisions. All authors have read, reviewed, and approved of the final manuscript.

Competing interests

None.



Supplementary Files

Figures

Closed loop data generation process.

