

# Human-in-the-loop machine learning for healthcare: Current progress and future opportunities in electronic health records

Han Yuan<sup>1</sup>  | Lican Kang<sup>2</sup>  | Yong Li<sup>3</sup>  | Zhenqian Fan<sup>1</sup> 

<sup>1</sup>Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

<sup>2</sup>Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore, Singapore

<sup>3</sup>Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

## Correspondence

Han Yuan, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.

Email: [yuan.han@u.duke.nus.edu](mailto:yuan.han@u.duke.nus.edu)

## KEYWORDS

electronic health records, human-in-the-loop, machine learning

Machine learning (ML), particularly deep learning, has emerged as a fundamental analytical tool for various medical tasks in electronic health records (EHRs) [1]. However, the purely data-driven methods do not serve as a panacea for all encountered problems such as data annotation. To address these issues, human-in-the-loop (HITL) has increasingly gained prominence. It leverages human expertise to improve ML-based analyses [2]. In this commentary, we perform a literature search to identify the current progress (Figure 1), determine research gaps, and highlight future opportunities for HITL across the ML lifecycle, including data preparation, feature engineering, model development, and model deployment.

The first phase in which HITL enhances ML is data preparation. This phase includes data extraction, data preprocessing, and data annotation of large-scale raw EHRs into formats operable for downstream modeling [3]. Across the three data preparation steps, data annotation is the focal point in the latest HITL research because the traditional paradigm indiscriminately annotates all available samples by default, which places an unnecessary burden on human experts in time-urgent

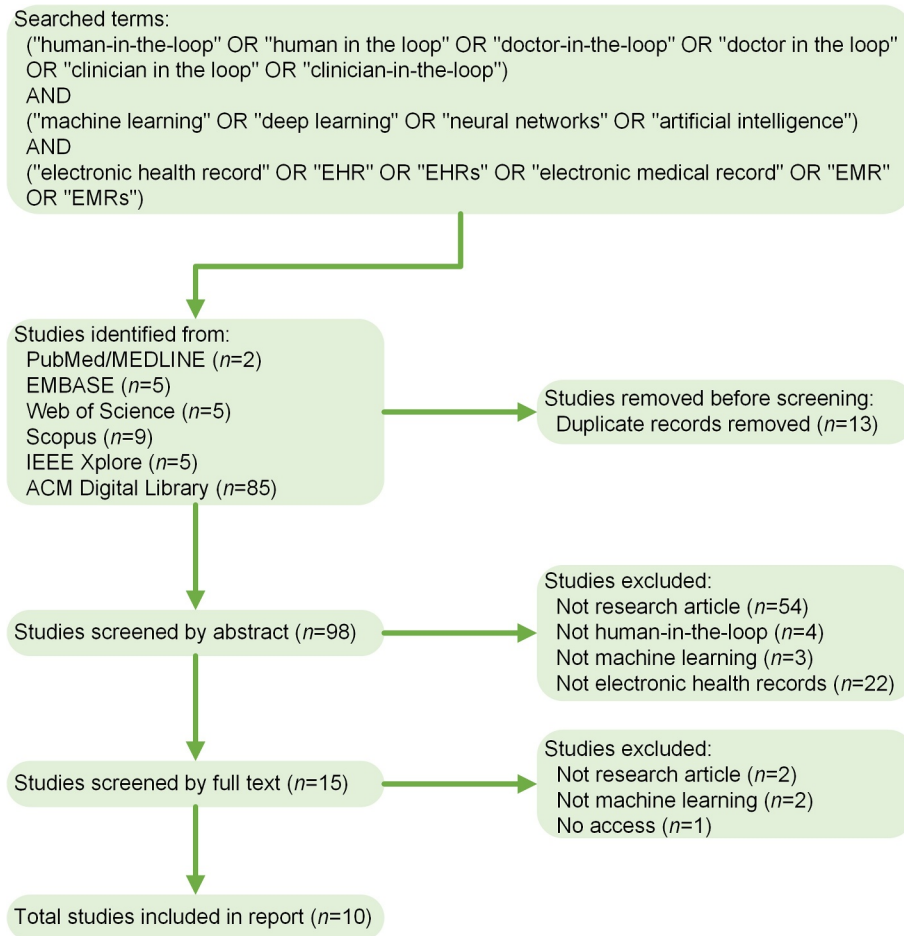
medical settings. Bull et al. [4] designed an interactive HITL platform that enables clinicians to verify or correct labels predicted by ML. Evaluated on two EHRs databases, the developed platform quickly generated accurate ML models with reduced annotation needs. Similar strategies have been implemented for detecting unauthorized access in data extraction [5] and deidentifying free text [6] in data preprocessing. Given the powerful ability of foundation models in zero-shot inference [7], future studies may use them, such as GPT-4, to replace homemade ML models in computer-aided annotations [8]. Moreover, current studies predominantly focus on data annotation; hence, there remains a vast and unexplored blue ocean for HITL in data extraction, such as data integration, and data preprocessing, such as noise filtering and missing value imputation [9].

Building on well-prepared datasets, the subsequent applications of HITL-ML to EHRs span feature engineering and model development. Feature engineering without HITL relies on either fully automated or fully manual methods, which demand large amounts of computation resources or expert involvement. The incorporation of

**Abbreviations:** EHRs, Electronic Health Records; HITL, Human-in-the-loop; ML, Machine Learning.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Medicine Advances* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.



**FIGURE 1** Pipeline of the literature search of human-in-the-loop machine learning in electronic health records.

HITL has enabled the generation of novel features of comparable quality at speeds up to 20 times faster than the original methods [10]. In classic ML, feature engineering has long been deemed essential, preceding model development in numerous contexts because of its demonstrated efficacy in enhancing model performance. However, in recent years, a notable shift has occurred toward the end-to-end paradigm for model development, gradually rendering traditional feature engineering less pivotal [1]. An exemplification of this trend can be observed in artificial neural networks, where shallow layers undertake the task of feature engineering for deep layers, thereby enabling automatic and seamless optimization during model development. Within this paradigm, HITL improves both model architecture design and parameter optimization. Sheng et al. [11] invited doctors to modify the structure and causal relationships of a knowledge graph distilled from EHRs, thereby demonstrating that HITL elevated not only the accuracy but also the interpretability of ML. Rather than adjusting models post-training, Tari et al. [12] applied HITL during the

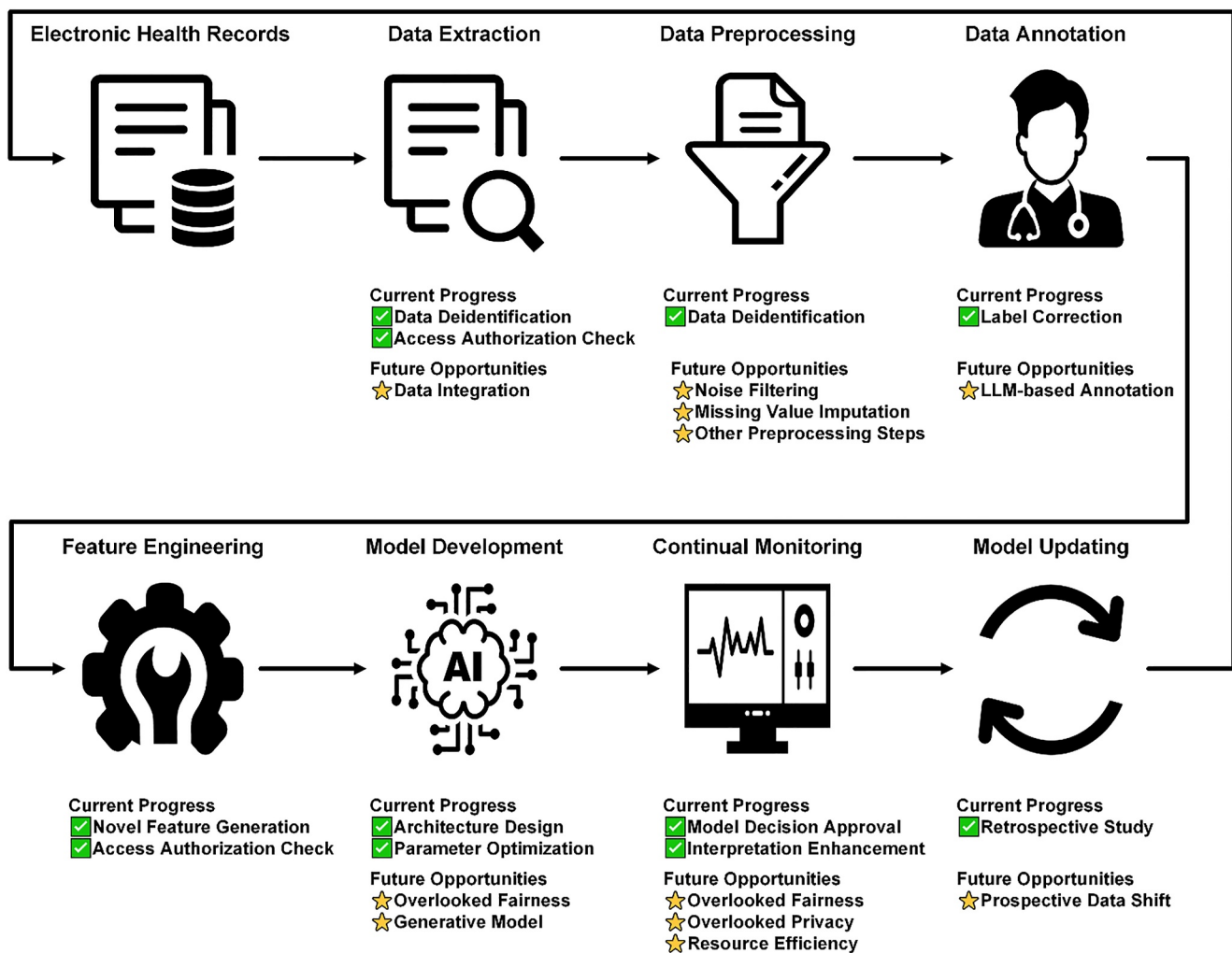
training phase by adding human preferences to the classic optimization target of gold-standard labels. The accuracy [11, 12] and interpretability [11] of ML for EHRs have been augmented by HITL; however, the aspect of fairness has not been sufficiently emphasized. Future researchers may introduce post-hoc recalibrations following [11], or alternatively, embed fairness as an optimization objective in the process of model development like [12] to mitigate potential inequities [13]. Furthermore, the present tasks of model development focus on EHR classification and regression. Future research endeavors could venture into EHR generation to support privacy-preserving analyses on synthetic pseudo samples [14].

Once model development is complete, the final phase of the ML lifecycle is model deployment, which encompasses the continuous monitoring and updating of trained models. HITL has been incorporated into this phase to ensure ML accuracy, interpretability, and compatibility toward temporal and spatial shifts. Doctors have been engaged to double-check intervention suggestions from developed ML models, such as positive

infection cases [15] and medication doses [16]. Instead of seeking approval for all decisions from clinicians, Zheng et al. [17] proposed that ML should be able to distinguish difficult and simple cases so that such cases could be solved by medical experts and models, respectively. In addition to ensuring ML accuracy through HITL [15–17], Elshawi et al. [18] and Yuan et al. [19] used clinician-labeled concepts to interpret ML behaviors and clarified their advantages over explanations solely generated by ML. Research on model deployment should also broaden its focus to consider fairness and privacy. Furthermore, the resource efficiency of ML should not be neglected in model deployment because models could be executed on mobile devices with limited computation capability [20]. Even with access to powerful cloud infrastructure, for time and privacy-sensitive medical applications, efficient models should run on edge devices because of their low latency and privacy-preserving benefits. Finally, most previous model deployments

were simulated using retrospective EHRs, which prompts future HITL research to resolve performance deterioration in prospective clinical landscapes [21].

In this commentary, we have shown that HITL not only refines data preparation and feature engineering but also catalyzes advancements in model development and deployment, thereby yielding ML tools that are accurate and interpretable. Despite the elucidation of the advantages of leveraging HITL in ML for EHRs, the full potential of HITL is yet to be harnessed. Figure 2 shows an overview of existing gaps and highlights future opportunities. The synergistic interaction among healthcare professionals, ML engineers, and high-performance computers is poised to fulfill the potential of HITL in the enhancement of ML. This human-computer interaction promises not only to improve accuracy, efficiency, and robustness but also to foster interpretability, impartiality, and privacy preservation in healthcare ML systems.



**FIGURE 2** Schematic plot of the current progress and future opportunities of human-in-the-loop across the machine learning lifecycle.

## AUTHOR CONTRIBUTIONS

**Han Yuan:** Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); project administration (lead); visualization (lead); writing – original draft (lead); writing– review & editing (lead). **Lican Kang:** Data curation (supporting); formal analysis (supporting); writing – review & editing (supporting). **Yong Li:** Investigation (supporting); methodology (supporting); writing – review & editing (supporting). **Zhenqian Fan:** Investigation (supporting); methodology (supporting); writing – review & editing (supporting).

## ACKNOWLEDGMENTS

None.

## CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data was created or analyzed in this study.

## ETHICS STATEMENT

Not applicable.

## INFORMED CONSENT


Not applicable.

## ORCID

Han Yuan  <https://orcid.org/0000-0002-2674-6068>

Lican Kang  <https://orcid.org/0000-0003-3136-9225>

Yong Li  <https://orcid.org/0000-0002-8949-8612>

Zhenqian Fan  <https://orcid.org/0000-0001-9127-8355>

## REFERENCES

- [1] Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. *J Biomed Inf.* 2022;126:103980. <https://doi.org/10.1016/j.jbi.2021.103980>
- [2] Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. *Future Generat Comput Syst.* 2022;135:364–81. <https://doi.org/10.1016/j.future.2022.05.014>
- [3] Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle. *ACM Comput Surv.* 2022;54(5):1–39. <https://doi.org/10.1145/3453444>
- [4] Bull NJ, Honan B, Spratt NJ, Quilty S. A method for rapid machine learning development for data mining with doctor-in-the-loop. *PLoS One.* 2023;18(5):e0284965. <https://doi.org/10.1371/journal.pone.0284965>
- [5] Boddy A, Hurst W, MacKay M, El Rhalibi A. A hybrid density-based outlier detection model for privacy in electronic patient record system. In: 2019 5th international conference on information management (ICIM). Cambridge; 2019. p. 92–6.
- [6] Liu L, Perez-Concha O, Nguyen A, Bennett V, Blake V, Gallego B, et al. Web-based application based on human-in-the-loop deep learning for deidentifying free-text data in electronic medical records: development and usability study. *Interact J Med Res.* 2023;12:e46322. <https://doi.org/10.2196/46322>
- [7] López Espejel J, Ettifouri EH, Yahaya Alassan MS, Chouham EM, Dahhane W. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Nat Lang Process J.* 2023;5:100032. <https://doi.org/10.1016/j.nlp.2023.100032>
- [8] Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology.* 2023;308(3):e231362. <https://doi.org/10.1148/radiol.231362>
- [9] García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. *Big Data Anal.* 2016;1(1):9. <https://doi.org/10.1186/s41044-016-0014-0>
- [10] Salam MA, Koone ME, Thirumuruganathan S, Das G, Basu Roy S. A human-in-the-loop attribute design framework for classification. In: The world wide web conference. San Francisco; 2019. p. 1612–22. <https://doi.org/10.1145/3308558.3313547>
- [11] Sheng M, Li A, Bu Y, Dong J, Zhang Y, Li X, et al. DSQA: a domain specific QA system for smart health based on knowledge graph. In: International conference on web information systems and applications: Springer; 2020. p. 215–22. [https://doi.org/10.1007/978-3-030-60029-7\\_20](https://doi.org/10.1007/978-3-030-60029-7_20)
- [12] Tari L, Mulwad V, von Reden A. Interactive online learning for clinical entity recognition. In: Proceedings of the workshop on human-in-the-loop data analytics. San Francisco; 2016. p. 1–6. <https://doi.org/10.1145/2939502.2939510>
- [13] Thompson HM, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *J Am Med Inf Assoc.* 2021;28(11):2393–403. <https://doi.org/10.1093/jamia/ocab148>
- [14] Lee SH. Natural language generation for electronic health records. *NPJ Digit Med.* 2018;1:63. <https://doi.org/10.1038/s41746-018-0070-0>
- [15] Bonde A, Lorenzen S, Brixen G, Troelsen A, Sillesen M. Assessing the utility of deep neural networks in detecting superficial surgical site infections from free text electronic health record data. *Front Digit Health.* 2024;5:1249835. <https://doi.org/10.3389/fdgh.2023.1249835>
- [16] Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Orlando; 2016. p. 2978–81.
- [17] Zheng K, Chen G, Herschel M, Ngiam KY, Ooi BC, Gao J. PACE: learning effective task decomposition for human-in-the-loop healthcare delivery. In: Proceedings of the 2021 international conference on management of data. Virtual Event China; 2021. p. 2156–68. <https://doi.org/10.1145/3448016.3457281>

- [18] El Shawi R, Al-Mallah MH. Interpretable local concept-based explanation with human feedback to predict all-cause mortality. *Jair*. 2022;75:833–55. <https://doi.org/10.1613/jair.1.14019>
- [19] Yuan H, Hong C, Jiang PT, Zhao G, Tran NTA, Xu X, et al. Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification. *J Biomed Inf*. 2024;156:104673. <https://doi.org/10.1016/j.jbi.2024.104673>
- [20] Fawagreh K, Gaber MM. Resource-efficient fast prediction in healthcare data analytics: a pruned Random Forest regression approach. *Computing*. 2020;102(5):1187–98. <https://doi.org/10.1007/s00607-019-00785-6>
- [21] Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat Biomed Eng*. 2022;6(12):1330–45. <https://doi.org/10.1038/s41551-022-00898-y>

**How to cite this article:** Yuan H, Kang L, Li Y, Fan Z. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. *Med Adv*. 2024;2(3):318–22. <https://doi.org/10.1002/med4.70>