CuraBench: A Benchmark Dataset Generation System for Healthcare AI Evaluation

Hejie Cui
hejie.cui@stanford.edu
Stanford University
Palo Alto, California, USA

Alyssa Unell
aunell@stanford.edu
Stanford University
Palo Alto, California, USA

Haoran Zhang
haoranz@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Caleb Winston
calebwin@cs.stanford.edu
Stanford University
Palo Alto, California, USA

Jason Alan Fries
jfries@stanford.edu
Stanford University
Palo Alto, California, USA

Sanmi Koyejo
sanmi@stanford.edu
Stanford University
Palo Alto, California, USA

Nigam Shah
nigam@stanford.edu
Stanford University
Palo Alto, California, USA

Abstract

Ensuring that artificial intelligence (AI) tools in healthcare operate safely and effectively requires robust evaluation within realistic clinical contexts. Traditional evaluation methods often rely on standardized benchmarks that fail to capture the full complexity of patient care, while manually curating a dataset for a specific deployment scenario can be time-consuming and limiting. We propose **CuraBench**, a configurable benchmark generation system designed to create customized synthetic datasets tailored to specific clinical use cases. CuraBench's taxonomy-driven configurable approach enables diverse evaluation scenarios—from assessing how AI systems interpret longitudinal patient histories, to evaluating clinical note summarization. By leveraging real-world healthcare data, CuraBench produces synthetic yet realistic scenarios configured to match the requirements of various medical settings, specialties, and patient demographics. Preliminary validation (TIMER) demonstrates the effectiveness of configurable benchmark generation in revealing evaluation biases undetectable with existing benchmarks. By streamlining the creation of comprehensive benchmark datasets, CuraBench represents a significant step toward responsible AI deployment, ensuring that models are rigorously tested in environments that mirror their intended clinical use.

ACM Reference Format:
Hejie Cui, Alyssa Unell, Haoran Zhang, Caleb Winston, Jason Alan Fries, Sanmi Koyejo, and Nigam Shah. 2025. CuraBench: A Benchmark Dataset Generation System for Healthcare AI Evaluation. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Healthcare AI evaluation faces critical limitations that hinder the responsible deployment of AI systems in medical settings. Current approaches fall short in two key aspects: inappropriate benchmarks that rely on simplified test scenarios rather than real clinical complexity [1] and lack of benefit specification [4]. For example, many studies benchmark AI models using multi-choice question answering datasets [8], such as the United States Medical Licensing Examination (USMLE), which poorly represent real-world clinical settings [12]. These standardized tests fail to capture the complexity of actual clinical decision-making. They present simplified, idealized case presentations that do not reflect real patient care, which involves temporal complexity, incomplete information, and contextual ambiguity [7, 9, 10].

Furthermore, evaluations often lack clear specifications of expected clinical benefits, making it impossible to assess whether AI tools deliver real-world value [3]. Additionally, these static approaches cannot accommodate diverse healthcare settings with unique requirements, where context, patient populations, and workflows vary significantly across institutions and specialties [11, 14, 15]. For instance, evaluating AI for pediatric emergency departments requires different scenarios than geriatric oncology clinics.

While traditional benchmark creation attempts to address these limitations through extensive manual curation by clinical experts [6, 13], this approach makes comprehensive evaluation prohibitively expensive and time-consuming. Clinical experts must invest significant time reviewing patient records, formulating questions, and validating responses and causal reasoning — a process that scales poorly and creates bottlenecks in AI development cycles.

The critical gap in healthcare AI evaluation lies in creating representative, context-specific benchmarks [3]. Such evaluation requires representative datasets that enable customized, task-specific assessments tailored to local clinical contexts. This necessitates a systematic approach to benchmark generation that produces curated yet realistic scenarios matching diverse medical settings while ensuring meaningful translation to real-world clinical values.

2 CuraBench: A Configurable Benchmark Generation System

We propose **CuraBench**, a benchmark dataset generation system that facilitates adaptive AI evaluation across diverse healthcare scenarios. The system addresses the fundamental limitations of static benchmarks by operating on a comprehensive taxonomy, accepting user-specified configurations, and generating tailored benchmarks for specific evaluation needs. By leveraging real-world healthcare data as the starting point, CuraBench produces curated yet realistic scenarios that enable cost-effective verification of AI benefits while maintaining clinical authenticity.

2.1 Multi-faceted Taxonomy and Architecture

CuraBench's taxonomy includes various key configuration dimensions and their corresponding possible values derived from electronic health records and clinical workflows, enabling systematic customization of benchmark generation:

- **Patient Demographics**: Age, gender, race, socioeconomic status, and geographic information to ensure evaluation across diverse patient populations and identify potential disparities.
- **Clinician Information**: Specialty, level of training, role, and institutional context to capture varied healthcare provider perspectives and needs across different care settings.
- **Healthcare Tasks**: Retrieval, diagnostic support (e.g., differential diagnosis generation), care planning (e.g., treatment recommendation), and clinical documentation (e.g., visit summarization) spanning the broad spectrum of AI applications in healthcare [2].

The benchmark generation pipeline operates by accepting user-specified configurations and systematically creating representative benchmark sets from real patient EHRs. The system can generate diverse evaluation scenarios, from instruction-response pairs derived from patient timeline chunks for evaluating longitudinal reasoning capabilities [5], to diagnostic scenarios that test AI systems' ability to synthesize information across multiple clinical encounters and generate causal reasoning. This flexibility allows healthcare organizations to create proxy golden evaluation sets that reflect their specific patient populations, clinical workflows, and evaluation priorities.

2.2 Foundational Work and Validation

Our approach expands upon preliminary investigations that demonstrate the feasibility of generating clinical instructions using large language models. We developed and validated **TIMER** [5], which demonstrates the practical effectiveness of CuraBench's benchmark generation approach for addressing critical gaps in healthcare AI evaluation. TIMER introduced a novel evaluation methodology that creates timestamp-linked instruction-response pairs from longitudinal EHR data, enabling systematic assessment of AI models' temporal reasoning capabilities across patient histories.

This benchmark generation approach revealed previously unrecognized biases in existing evaluation datasets, where over 55% of questions focused only on the final 25% of patient timelines. TIMER's automatically generated benchmarks (**TIMER-Eval**) provided more comprehensive temporal coverage compared to human-curated benchmarks like MedAlign, enabling detection of model capabilities that were previously unmeasurable due to evaluation limitations. TIMER's controlled temporal sampling strategies showed that distribution-matched training demonstrated advantages up to 6.5% in temporal reasoning evaluation, highlighting how benchmark design directly impacts model performance.

Additionally, we generated instruction-response QA pairs across 5 core task categories and 261 distinct clinician personas, achieving 45% classification alignment with expert clinicians - a promising baseline that demonstrates feasibility for programmatic benchmark generation while highlighting opportunities for improvement. These results collectively validate CuraBench's core hypothesis that configurable benchmark generation enables realistic contexts for meaningful clinical AI evaluation.

2.3 Benchmark Quality Assurance

CuraBench can be used to evaluate both open-source and closed-source LLMs with standardized metrics such as F1 score, ROUGE, and task-specific measures. To ensure benchmark validity and clinical relevance, CuraBench employs a comprehensive two-stage quality control process: Stage 1 implements automated data selection using classification models trained on clinician-annotated preference data. We systematically compare the performance of distilled LLM-as-a-judge versus multi-way classifiers to identify high-quality benchmark candidates while maintaining computational efficiency. Stage 2 involves rigorous multi-faceted human evaluation conducted by clinical fellows from multiple specialties, ensuring cross-domain validity. Statistical analysis using Cohen's kappa coefficient measures inter-rater agreement, with established thresholds ensuring evaluation consistency and clinical relevance across generated benchmarks.

2.4 Long-term Vision and Impact

CuraBench could fundamentally reshape the healthcare AI development lifecycle by enabling continuous evaluation throughout model development. This paradigm shift toward configurable, context-aware evaluation could accelerate responsible AI deployment while addressing current bottlenecks where manual benchmark creation requires months of expert time at costs exceeding $100,000 per specialized evaluation. Healthcare organizations could adopt this through phased implementation, starting with pilot evaluations in specific departments before scaling to institution-wide AI assessment frameworks that adapt to their evolving clinical needs.

3 Conclusion

CuraBench addresses critical limitations in healthcare AI evaluation by providing a configurable, scalable solution, demonstrating the potential to transform healthcare AI evaluation from static, one-size-fits-all approaches to dynamic, context-aware assessment. CuraBench enables healthcare organizations to create customized evaluation datasets reflecting their specific patient populations and clinical workflows. Our validation demonstrates feasibility for programmatic benchmark generation. CuraBench represents a significant step toward responsible AI deployment, providing tools to rigorously verify AI benefits before clinical implementation.

Acknowledgments

We thank the clinical informatics fellows at Stanford Medicine and healthcare providers at Stanford Healthcare for their valuable insights and contributions to this work.

References

[1] Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical Large Language Model Benchmarks Should Prioritize Construct Validity. *arXiv preprint arXiv:2503.10694* (2025).

[2] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. *arXiv preprint arXiv:2505.23802* (2025).

[3] Suhana Bedi, Sneha S Jain, and Nigam H Shah. 2024. Evaluating the Clinical Benefits of LLMs. *Nature Medicine* (2024). doi:10.1038/s41591-024-03181-6

[4] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2024. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* (2024).

[5] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. 2025. TIMER: Temporal Instruction Modeling and Evaluation for Longitudinal Clinical Records. *CoRR* abs/2503.04176 (2025). https://doi.org/10.48550/arXiv.2503.04176

[6] Scott L Fleming et al. 2024. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38.

[7] Gordon Guyatt, Drummond Rennie, and S Satya-Murti. 2002. Users' guides to the medical literature: a manual for evidence-based clinical practice. *JAMA-Journal of the American Medical Association-International Edition* 287, 11 (2002), 1463.

[8] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[9] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? *arXiv preprint arXiv:2403.17752* (2024).

[10] Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. 2025. It's Time to Bench the Medical Exam Benchmark. AIe2401235 pages.

[11] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. 2020. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR medical informatics* 8, 7 (2020), e15182.

[12] Nigam H Shah, David Entwistle, and Marc A Pfeffer. 2023. Creation and Adoption of Large Language Models in Medicine. *JAMA* 330, 9 (2023), 866–869. doi:10.1001/jama.2023.14217

[13] Nikos Sourlos, Rozemarijn Vliegenthart, Joao Santinha, Michail E Klontzas, Renato Cuocolo, Merel Huisman, and Peter van Ooijen. 2024. Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. *Insights into Imaging* 15, 1 (2024), 248.

[14] Stanford Institute for Human-Centered Artificial Intelligence. 2024. Healthcare Algorithms Don't Always Need to Be Generalizable. https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable Accessed: 2025-05-20.

[15] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, et al. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine* 181, 8 (2021), 1065–1070.