MEDS Decentralized, Extensible Validation (MEDS-DEV) Benchmark: Establishing Reproducibility and Comparability in ML for Health

This effort is a collaboration of many groups and individuals. In this listing, we list working groups within this effort, and individuals involved in those groups in alphabetical order by first name.

The Core MEDS-DEV Working Group:
Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A Fries, Jeffrey N. Chiang, Jungwoo Oh, Justin Xu, Kamilė Stankevičiūtė, Kiril V. Klein, Matthew B. A. McDermott, Mikkel Odgaard, Nassim Oufattole, Nigam H. Shah, Patrick Rockenschaub, Pawel Renc, Robin P. van de Water, Shalmali Joshi, Simon A. Lee, Teya S. Bergamaschi, Tom J. Pollard, Vincent Jeanselme, Young Sang Choi
MATTMCDERMOTT8@GMAIL.COM

Core MEDS-DEV Tools: ACES, FEMR, MEDS-Evaluation, MEDS-Reader, MEDS-Transforms
Chao Pang, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A Fries, Justin Xu, Kamilė Stankevičiūtė, Matthew B. A. McDermott, Michael Wornow, Nassim Oufattole, Nigam H. Shah, Pawel Renc, Teya S. Bergamaschi, Vincent Jeanselme

MIMIC-IV MEDS Dataset
Ethan Steinberg, Matthew B. A. McDermott, Nassim Oufattole, Pawel Renc, Tom J. Pollard

Columbia MEDS Dataset
Apara Kashyap, Chao Pang, Shalmali Joshi, Vincent Jeanselme, Xinzhuo Jiang, Yanwei Li, Young Sang Choi, Yuta Kobayashi

Other MEDS-DEV Ready Datasets (AUMCdb & eICU)
Matthew B. A. McDermott, Patrick Rockenschaub, Robin P. van de Water, Ryan C King

Profiled MEDS Models
Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jason A Fries, Jungwoo Oh, Michael Wornow, Nassim Oufattole, Teya S. Bergamaschi, Xinzhuo Jiang

1. Introduction

Standardized benchmarks have driven significant progress in machine learning (ML) (Donoho, 2024; Deng et al., 2009). Benchmarks establish clearly defined metrics for success, allow researchers to fairly compare methods, and facilitate reproducibility and open science. Despite this, benchmarking remains underdeveloped in ML for healthcare.

Benchmarking in healthcare is hard for several reasons: (a) the lack of standardized schemas for sharing and processing data, which prohibit frictionless reproductions of published models over private health datasets; (b) ambiguity in how tasks and labels are defined, leading to irreproducible task definitions across papers; and (c) the inability to meaningfully compare model performance across the fragmented health data landscape Johnson et al. (2017); McDermott et al. (2021); Wang et al. (2020); Liao and Voldman (2023); Harutyunyan et al. (2019).

To resolve these limitations, we propose the MEDS Decentralized, Extensible Validation (MEDS-DEV) benchmark, a distributed benchmarking framework that enables seamless reproduction of model results with conceptually identical task definitions across a diverse set of source datasets, including both public and private datasets. MEDS-DEV differs from traditional benchmarks in a number of ways in order to be best suited to the ML4H ecosystem:

Decentralized Evaluation By default in MEDS-DEV, data is *not* presumed to be shareable or publicly available, and as such model architectures[1] will be evaluated on different datasets in a sparse, decentralized fashion driven by local collaborations and, eventually, larger competitions and curated efforts.

Extensible Task Landscape Secondly, MEDS-DEV is designed to operate over a large number of community curated, clinically meaningful tasks that are consistently defined in a dataset-agnostic manner. This permits the benchmark to both expand to diverse clinical areas of interest and to cover a much more rigorously curated and refined set of tasks through community engagement.

Rigorous, Comparable Validation Finally, by virtue of the MEDS standard and the seamless reproducibility it offers, MEDS-DEV can use identical evaluation systems across submitted models and tasks, streamlining analysis of not only performance, but fairness metrics, calibration, computational costs of training and evaluation, dataset-size sensitivity, and more. This offers a significantly expanded set of analysis opportunities for models in the ML4H field. While only a subset of these evaluation metrics are currently implemented, all are clearly operationalizable in the MEDS-DEV model and on the planned future roadmap.

2. Method

Adding new results For a new task, model, and dataset combination, the results can be submitted via a pull request to the MEDS-DEV GitHub repository in the form of a standardized MEDS-Evaluation results file. The pull request will be reviewed by the maintainers of MEDS-DEV and the results incorporated into the leaderboard upon approval.

Adding new tasks MEDS-DEV prediction tasks are defined through the ACES configuration system (Xu et al., 2024). This library enables the expression of task predicates (e.g. phenotype and event definitions) and inclusion/exclusion criteria in a dataset-agnostic way. To propose a new task in MEDS-DEV, users submit a pull request to the GitHub repository with an ACES configuration file and (optionally) any dataset-specific mappings of ACES predicates in the corresponding dataset configuration files. The pull request should also contain a README describing the task and its associated clinical utility. The community can comment on this pull request to suggest changes to the task or to contest its clinical utility before its final inclusion in MEDS-DEV. Note that adding a new task to MEDS-DEV does not inherently generate new model results for that task; instead, the results can be added over time, reflecting the sparse, decentralized nature of MEDS-DEV. We hope that these clear, standardized and reproducible task definitions is a valuable contribution to ML4H given its current challenges with reproducing basic concepts such as mortality (Johnson et al., 2017).

Adding new models For participation in MEDS-DEV, a model must be runnable on any MEDS dataset with its outputs conforming to the standardized MEDS-Evaluation schema. Model description is submitted via a pull request to the GitHub repository along with any relevant metadata and usage instructions as required by the pull request template.

1. Note that MEDS-DEV primarily helps compare model architectures and training recipes rather than pre-trained models, given the lack of widespread data availability.

Model	MIMIC-IV/Columbia	
	Long LOS	ICU Mortality
Log. Reg.	0.752/0.677	0.754/0.509
LightGBM	0.783/0.757	0.798/0.661
MOTOR	0.804/0.735	**0.854**/0.727
CEHR-BERT	0.808/0.741	0.845/0.726
MEDS-Tab	**0.811**/**0.761**	0.830/**0.785**
GenHPF	0.779/0.662	0.790/0.633

Table 1: Proof of viability results (AUC-ROC). Results format: "MIMIC-IV"/"Columbia".

Adding new datasets To support a MEDS-compatible dataset in MEDS-DEV, the contributor must define a predicate configuration file mapping dataset-specific features to task-specific concepts as defined by their ACES configuration files, document any limitations and incompatibilities (in terms of censoring, inclusion/exclusion criteria, and other potential biases), and describe the access policy. As with other contributions, this information is submitted as a pull request to the MEDS-DEV GitHub repository.

3. Results

MEDS-DEV is designed for extensibility and community contribution; however, we have a number of existing, proof-of-viability results demonstrating this style of benchmarking, with reproducibility as a first class citizen. In particular, MEDS-compliant datasets for use in MEDS-DEV have already been curated for a number of public and private datasets, including *(public)* MIMIC-IV (Johnson et al., 2023), eICU (Pollard et al., 2018), AUMCdb (Thoral et al., 2021), EHRShot (Wornow et al., 2023), *(private)* Stanford data, Columbia data, and cohort-specific datasets from Toronto, Copenhagen, and Mass General Brigham, with further datasets still under construction. In addition, MEDS-complaint versions of various published model architectures, such as MOTOR (Steinberg et al., 2023), CLMBR (Steinberg et al., 2021), EBCL (Jeong et al., 2024), GenHPF (Hur et al., 2023), CEHR-BERT (Pang et al., 2021), and MEDS-Tab (Oufattole et al., 2024) already in use, with further models actively being converted for inclusion, such as ETHOS (Renc et al., 2024), ESGPT (McDermott et al., 2023), CORE-BEHRT (Odgaard et al., 2024) and more.

For all of these models and datasets, MEDS-DEV contains a preliminary collection of 12 tasks across different clinical areas and challenges. MEDS-DEV is designed for the set of examined tasks to grow and change over time through community contribution, and we have already seen some of the most extensive and involved community discussions on the merits of different task inclusion/exclusion criteria in the GitHub Issues for MEDS-DEV that most authors in this project have encountered professionally to date.

To demonstrate the viability of transporting these models across public and private datasets, we show a subset of preliminary model results from MEDS-DEV in Table 1. This table shows comparison of newly trained models across 6 different model architectures from 4 different author groups across a public and private dataset (MIMIC-IV and a subset of Columbia data, respectively), demonstrating that these models can be reliably reproduced across sites via the MEDS and MEDS-DEV frameworks. Note that these results are preliminary—and in particular the Columbia data used is only a 10K patient subset of their entire cohort—but they nevertheless establish the viability of the MEDS-DEV system, thus motivating its presentation as a demonstration to the ML4H community to help encourage this new, significantly more reproducible style of learning in our field.

4. Discussion

MEDS-DEV represents a first step towards building a standardized benchmarking infrastructure for healthcare research. It addresses the three main limitations of prior benchmarks: data standardization, task definition consistency, and multi-institution participation. As shown in Section 3, MEDS-DEV enabled us, for the first time, to quickly evaluate four state-of-the-art EHR foundation models across multiple institutions on a common set of tasks. As health systems begin to deploy models into the clinic, benchmarking efforts such as MEDS-DEV will serve an increasingly important role in validating models and help accelerate the development of ML methods for EHR data.

With several dozen members already in the MEDS-DEV community, we are excited to build upon this momentum and present MEDS-DEV to the broader ML4H audience. We invite anyone who resonates with our vision for more rigorous, reproducible science to join the MEDS-DEV community and contribute models, datasets, and tasks here: `https://github.com/mmcdermott/MEDS-DEV`.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Hyewon Jeong, Nassim Oufattole, Matthew Mcdermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series, 2024. URL https://arxiv.org/abs/2312.10308.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine learning for healthcare conference*, pages 361–376. PMLR, 2017.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023.

Wei Liao and Joel Voldman. A multidatabase extraction pipeline (metre) for facile cross validation in critical care research. *Journal of Biomedical Informatics*, 141:104356, 2023.

Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive ehr timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 257–278, 2021.

Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24322–24334. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c8f197b24e9b05d22028c2de16a45d2-Paper-Datasets_and_Benchmarks.pdf.

Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv preprint arXiv:2404.15201*, 2024.

Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin Stultz, and Matthew McDermott. Meds tab: Scalable tabularization and tabular feature usage utilities over generic meds datasets. https://github.com/mmcdermott/MEDS_Tabular_AutoML, 2024. Accessed: 2024-10-31.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ Digital Medicine*, 7(1):256, 2024.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113: 103637, 2021.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

Patrick J Thoral, Jan M Peppink, Ronald H Driessen, Eric JG Sijbrands, Erwin JO Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49 (6):e563–e577, 2021.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. 2023.

Justin Xu, Jack Gallifant, Alistair EW Johnson, and Matthew McDermott. Aces: Automatic cohort extraction system for event-stream datasets. *arXiv preprint arXiv:2406.19653*, 2024.