The (lack of?) Science of Machine Learning for Healthcare

Matthew McDermott MATTHEW_MCDERMOTT@HMS.HARVARD.EDU
Harvard Medical School, Department of Biomedical Informatics, Boston, USA

The Science of Machine Learning for Healthcare

Consider the following scenario: A clinician scientist at a large health system emails you, asking for advice on designing a machine learning (ML) / artificial intelligence (AI) system for their population. They have a clinical problem and training dataset full of structured medical record data[1] in mind and have acquired funds and compute resources to train a model. Their only issue is that they don't know what kind of model would be best – and unfortunately they only have a limited window to use their resources. So, they turn to you, an expert, and ask precisely this: What kinds of models should they train to solve their clinical task?

When I have posed this question to colleagues, the typical response is "it depends." It depends on the size, makeup, and clinical properties of the covered population; it depends on what the clinical problem is; it depends on how much data we have and how informative we expect the data to be for this problem; and last but not least it depends on how the data are organized technically and conceptually. And, while this answer may be unsatisfying–and unhelpful to our hypothetical colleague–it is nevertheless *eminently reasonable*. The space of healthcare is vast and varied, and without answers to numerous questions about the dataset, problem, and deployment scenario in mind, expecting a one-size-fits-all model to exist and be clear at a glance is hubris.

Three key facts about health data synergize to hinder our ability to answer this question. First, *health datasets are highly complex*. The human system and the broader ecosystem of healthcare are extremely high-dimensional dynamical systems, with different diseases, health states, or healthcare settings perturbing that system in different ways. Atop all of this internal complexity, we measure health data in noisy, irregularly sampled, and biased manners, never seeing a true, raw view of the underlying system. Second, this complexity and the underlying states of data all manifest differently at different healthcare institutions, meaning that all *health datasets are unique*. It is not appropriate to expect the same patterns, data organizations, or population characteristics to be shared across different sites in the health world, necessitating local data expertise when working with health data. Finally, compounding both of these issues, even if we were to offer to examine our hypothetical colleague's dataset directly to find the answers to our questions, *health datasets are sensitive*, limiting our ability to answer our questions ourselves and imposing significant barriers to adapting techniques from one setting to another.

Unfortunately, while this description of health data is accurate—and while our hesitation to give an answer to this fictional question is appropriate—we nevertheless have exposed a significant problem. Namely, is there any point to studying how to build models (models that we don't necessarily intend to deploy) over health datasets if not than to offer inductive insight about how others should build models over their health datasets and problems? Moreover, is not the question posed by our fictional colleague the precise question that drives the field of "the science of machine learning for health"?[2] If we can't answer that question at all, then what have we learned in this field over all this time?

Some researchers in this community see this argument, and respond with the claim that, in fact, trying

1. While this perspective can apply to many sub-types of health data, it most directly applies to *longitudinal, structured medical record data*–data of categorical and/or numerical events occurring in continuous time in an irregularly sampled manner. Other modalities, in particular longitudinal data augmented with notes, imaging, or wearables data are also strong contenders for this perspective, but this work will focus most on longitudinal structured data.

2. Here, I mean very specifically the *computer* or *informatics* science question of how to best build models or informatics systems to work with health data for health problems. There are many other questions in the field of machine learning for health that are not included in this claim.

to study the science of machine learning in healthcare at all at the present time is not worthwhile, precisely because it is not possible to discover generalizable insights that span different datasets and clinical cultures with the current code and data ecosystem available in the academic world (Futoma et al., 2020; Miller, 2022). While this position is reasonable, it is unsatisfying. I *want* us to be able to study this science; I want our community to be able to discover new methods of model training, of injecting clinical knowledge, of leveraging complex, disparate data structures to work with health data and solve health tasks. And beyond my own intellectual curiosity, understanding such science is *important*. Understanding health and biomedical data through better models, capable of meaningfully forecasting diverse disease states, identifying latent phenotypes to help drive clinical discovery, or of helping link the growing inundation of clinical scientific articles to complex patient phenotypes could all have significant positive impacts on the state of human health. And it is clear to me that if this is a vision we want to pursue, then we need to change how we pursue research in our field to enable and empower scientific inquiry in this important topic. To do so, we must understand *why* our field is currently failing, and through this, how we can do better.

Empowering the Science of Machine Learning: Frictionless Reproducibility

Given the dramatic successes of machine learning and artificial intelligence in areas such as computer vision, natural language processing, and reinforcement learning, it is no surprise that precisely what factors drive this success has been the topic of significant study (Donoho, 2024; Salaudeen and Hardt, 2024; Liberman, 2015). In this work, the perspective that we will most draw upon is that of David Donoho, in his work exploring "Data Science at the Singularity" (Donoho, 2024). In this work, Donoho argues that the most essential aspect of the dramatic successes in the machine learning and artificial intelligence landscape is that of *frictionless reproducibility*–which is the property that it is extremely easy for other scientists to reproduce, and therefore iterate on, one another's work. This property enables the rapid pace of discovery in ML/AI precisely because it dramatically lowers the cost of meaningful experimentation. Donoho identifies three tenets that define frictionless reproducibility: the re-usability of research data (through easily accessible public data), research code (through technically reproducible and public code), and of research targets (through the adoption of shared benchmarks and challenges).

Unfortunately, while these tenets are achievable in many domains of ML/AI, it is also clear that they are, at best, wildly aspirational in ML/AI for healthcare in their current form. As we've noted, health datasets are highly sensitive, so frictionless access to data is not viable, and the complexity and uniqueness of health data mean that we cannot expect generalization from a small number of isolated public datasets to the broader world of health data. While reusability of code seems viable, in practice it is far less common than it should be, with a 2021 reproducibility study showing that fewer than 25% of published papers in this field released their code publicly (McDermott et al., 2021b). Finally, even what seemingly should be the easiest of the three tenets–frictionless adoption of shared, canonical research targets (typically through benchmarks or challenges)–seems out of reach in our field, *with past studies showing that even on public datasets, the basic definitions of the simple, widely studied mortality prediction problem differ greatly study to study and are irreproducible from the published papers alone (Johnson et al., 2017)*. With these significant failures to achieve even reproducibility, let alone *frictionless* reproducibility, we are forced to confront the troubling question: Is there anything we can do to achieve meaningful science in ML/AI for healthcare? Or, does the fact that our datasets are frustratingly complex, unique, and sensitive prohibit us from producing sufficiently reproducible and reliable results to drive empirical science forward?

Three True Lies about Health Data

Despite this dour picture, I do not feel that this is an unsolvable problem. In fact, if we take a step back from the context and the culture of the field of ML/AI for healthcare today and consider this proposition with fresh eyes, it actually seems somewhat ridiculous that we aren't doing better right now. As much as the 3 properties of health data that have caused such challenges are true, I argue that the inference that they imply frictionless reproducibility is functionally impossible in our field is false. Instead, I argue that claiming that we should allow the uniqueness and complexity of health data to hinder us from developing standardized, reproducible tools and models that can be examined across settings is as farci-

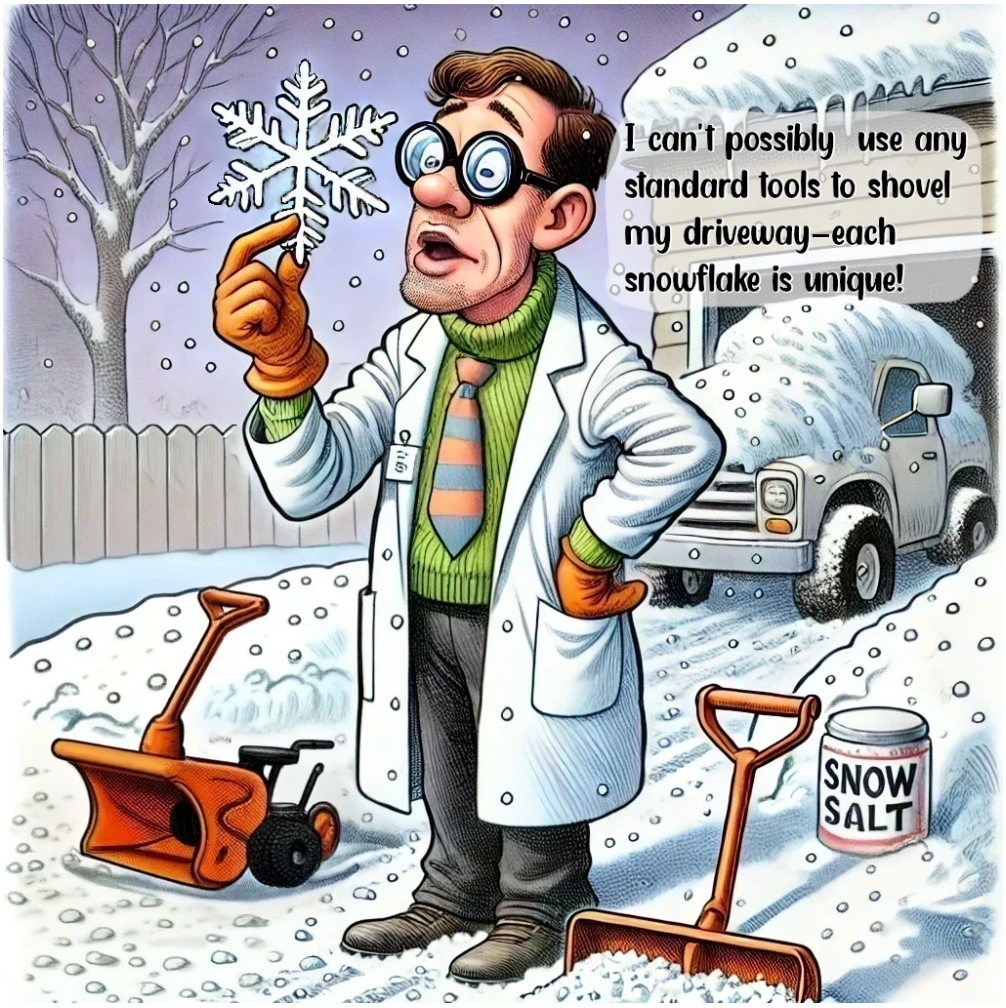

Figure 1: The Health AI Deep Learning scientist encounters a snowy driveway.

cal as claiming that the uniqueness and complexity of snowflakes should prevent one from using simple tools to clear one's driveway (Figure 1). In particular, we can re-frame our three health data properties into facts that empower us to build a reproducible, communal ecosystem of scientific discovery:

Firstly, while highly complex, *health datasets are also rich with simple insights and augmented with extensive scientific knowledge*, and are not clearly more daunting in functional form that domains such as natural language, protein structure, strategy and game playing, or global weather patterns, which are all areas where ML/AI has shown tremendous success. Secondly, while unique, *health datasets also share many commonalities that persist across sites*; in particular, all healthcare data can be represented in the form of an irregularly sampled temporal point process of events. This data form is arguably one of the defining aspects of longitudinal, structured health data, and is one reason why we might think there are meaningful, particular types of ML/AI models that are best suited to this space. And finally, while undeniably sensitive, *structured health data often can be readily de-identified*, and even unstructured health data permits automated de-identification system using modern ML techniques (Johnson et al., 2020; Murugadoss et al., 2021; Chambon et al., 2022).

If we believe that these three new facts are true, it suggests that we can grow a research community in this space that, while not necessarily as frictionless as other ML subdomains, still permits significant empirical exploration to be done in a much more meaningful way than exists today. In particular, these three facts evoke a world in which we could leverage the shared structure of health data to define data standards and develop frictionlessly re-trainable models, then share those model training recipes[3] across sites and datasets (public or private) through inter-institution collaborations powered by local, de-identified data access. In such a world we could begin to iteratively explore and refine model architectures over shared, transportable research tasks to build more effective models and develop the science of ML/AI in healthcare.

Unfortunately, even if we do believe this, the reality is that today, we don't exist in this world. We do have a reproducibility crisis and none of the above arguments have changed that. And if we wish to understand *why* this is, we must probe deeper into what has led to our reproducibility crisis, what barriers have prevented attempts to fix it to fail, and what, if anything, we can do about it.

Why do we have a reproducibility crisis/why have prior solutions failed?

If we wish to solve the reproducibility crisis in ML/AI for healthcare, we need to understand why it exists in the first place. To do so, we will start by exploring a longstanding area of research that seems poised to solve the reproducibility crisis: namely, data standardization. We will leverage the fact that existing standards have demonstrably *not* solved the reproducibility crisis[4] in ML/AI as a vehicle to understand what unique factors about ML/AI for healthcare contribute to our ongoing reproducibility crisis above and beyond factors that influence general medical informatics or general domain ML–and in so doing we will illuminate possible pathways to solve this problem more fully.

Existing data standards in health ML/AI Given the numerous advantages of data standardization, it is no surprise that there are a number of such systems for structured health data, many of which are in widespread use. These include the OHDSI OMOP common data model (CDM) (Reich et al., 2024), the i2b2 CDM (Wagholikar et al., 2022), the PCORnet CDM (Fleurence et al., 2014), among others.[5] Beyond CDMs, there are also a number of standardized data processing packages and benchmarks in this space that would further seem to solve this problem, including notable libraries such as FIDDLE, OMOP-learn, TemporAI, MIMIC-Extract, ESGPT, PyHealth, the Multi-task MIMIC benchmark, EHRShot, the EHR Pre-training Benchmark, and more (Tang et al., 2020; Kodialam et al., 2021; Saveliev and van der Schaar, 2023; Wang et al., 2020;

3. Note that I am not suggesting sharing model weights, but rather sharing model code in a frictionlessly re-usable manner, so that model architectures and training algorithms can be reliably assessed across sites, even when the underlying pre-trained weights would not transfer.

4. Note that this is not to imply that data standards have not been extremely impactful and essential for many important developments and areas of research; rather, this is simply to note that despite many health data models having existed for well over a decade, reproducibility problems in ML/AI in healthcare still persist (Johnson et al., 2017; McDermott et al., 2021b)

5. Note that I am excluding HL7 FHIR (International, 2024) here as, though it is a widely used CDM, it is designed for data transfer, not for data analytics.

McDermott et al., 2023; Yang et al., 2023; Harutyunyan et al., 2019; Wornow et al., 2023; McDermott et al., 2021a; Jarrett et al., 2021).

This breadth of research poses a key question: with all of these diverse standards and tools, why do we still have a reproducibility crisis at all? To answer this question, I believe we must examine two aspects of this problem-incentive-community space that come together to synergistically hobble efforts to solve the reproducibility crisis in health ML/AI: Namely, (1) key misalignment between the unique problems faced in ML/AI relative to more traditional informatics and (2) misaligned incentives within academia that promote irreproducible research.

Differentiating standards for informatics from standards for ML/AI The first fact that we must acknowledge is that all existing health common data models (e.g., OMOP, i2b2, PCORnet) were designed for more traditional informatics purposes, focusing on lower-capacity, typically federated and cross-system analyses, rather than on the more recent, extremely high-capacity models of the AI era. This distinction has several key consequences:

1. First, lower-capacity analyses require greater control and mechanistic understanding over your data, in order to power manual feature extraction efforts that are essential to power analyses. In contrast, higher-capacity neural network approaches often leverage all data available, without requiring that same degree of data control. Relatedly, higher-capacity systems are significantly more flexible in their ability to operate across data modalities in diverse ways, placing a greater premium on flexibility and simplicity for underlying data standards than lower-capacity analyses warrant.

2. Second, lower-capacity models are generally simultaneously (a) more easily trainable in a federated capacity due to their reduced parameter count and training needs, (b) more generalizable across populations or institutions, and (c) significantly less sensitive (in their pre-trained parameters) than higher-capacity neural networks. This greatly increases the extent to which developers of such tools are incentivized to realize impact through public or multi-institution deployment of their models, something that is often out of reach for higher-capacity systems due to privacy concerns around the pre-trained weights and inability to safely train multi-institution models without risking data security.

3. Third, both due to differences in community and the length of time that these tools have been in use, the core technologies used to power more classical informatics analyses (e.g., relational databases, CPU-powered computation, statistical programming languages) are very different than those used to power higher-capacity analyses (e.g., non-relational data storage systems, GPU-accelerated computation, python and deep-learning specific domain specific languages). These differences are reflected in how existing CDMs are built and optimized.

4. Finally, fourth, higher-capacity models are much more able to extract sources of bias and lower-power signal from data, placing a much greater onus on maintaining data fidelity with the underlying generative process rather than risking introducing systemic data bias through complex transformations.

These four properties paint clearly different optimal design principles for standards designed for lower-capacity analyses vs. those needed for ML/AI in healthcare. In particular, lower-capacity analytic tools must prioritize *data harmonization*[6] whereas tools for ML/AI do not need this additional complexity to the same degree, and can make significant advancements focusing on *data standardization*[7] only. This means that existing standards impose a significantly greater complexity barrier than what is needed to enable reproducibility in ML/AI. Relatedly, standards for more classical informatics analyses often can assume access to a particular technological stack which is not as familiar to the ML community as it is to the traditional informatics audience, thereby imposing a similar knowledge barrier relative to the needs of the ML/AI audience. Finally, standards for higher-capacity ML/AI must focus more on simplicity and extensibility to diverse health data modalities, something that is not as important with lower capacity systems, requiring additional work to be needed to

6. Meaning the task of ensuring the same conceptual vocabularies are used across sites to represent underlying clinical elements–e.g., all diagnostic codes must be converted to ICD-10.
7. Meaning the task of ensuring only that the technical organization of data into tables and on disk is identical–e.g., all sites must have data organized into common tables, with diagnoses, regardless of how they are encoded, captured in a particular column.

attempt to use existing CDMs as defining standards for ML/AI analyses.

Given that our goal is to get as close to frictionless *reproducibility as possible, all of these barriers are significant and diminish the extent to which we should expect these standards to solve this problem.*

This argument–that existing CDMs are not well optimized for ML/AI use cases given they are (appropriately) designed for more traditional informatics applications and the broader health informatics community–is borne out precisely by the existence of data standardization and pre-processing packages for the health ML/AI community highlighted above. While several of these can ingest data from existing formats, all of them internally rely on formats better optimized for ML/AI needs, often featuring properties such as event log structures, efficient internal conversion to PyTorch or TensorFlow structures, and more. However, while these packages are steps in the right direction, they too have not yet solved this problem. I argue that while this has many causes, the biggest reason is simply because the health ML/AI community is, in fact, not actually incentivized to solve this problem. Instead, as I argue next, the landscape of academic incentives directly promote a culture of irreproducible research in a manner we must confront if we hope to solve this problem fully.

Misaligned Incentives To complete our picture of why the reproducibility crisis in ML/AI for healthcare exists, we have to confront an unpleasant reality: that the structure of individual academic incentives in our research field actively supports irreproducible research in the short term. In particular, we can realize that irreproducibility acts as a protective mechanism over research silos, model directions, datasets, and invalid results, thereby enabling individual scientists to publish more papers faster, with less competition and risk of retraction. As publications are a major driver of academic notoriety and success, this means that individual researchers are incentivized in the short term to preserve the status-quo of irreproducibility.

Clearly, this fact is neither desirable nor sustainable. Academic incentives only promote irreproducibility in a short term sense; in the long term, as irreproducible research does not drive true scientific progress, impact will not emerge from such research, and lacking impact, grant funding will eventually too fall out of circulation. Further, organizations wishing to capitalize on the potential impact ML/AI can offer in healthcare will not be able to rely on the academic research in the space. Even outside of these more long-term constraints, the culture of irreproducibility also prevents us more practically from developing the rich, open-source communities that have developed in computer vision, natural language processing, and robotics that have been instrumental in their development. Note that these communities are especially essential considering the development of increasingly engineering and resource heavy ML/AI paradigms such as foundation models.

A better future

The arguments so far suggest that the pursuit of science in machine learning and AI in healthcare is fundamentally limited by the systemic culture of irreproducibility in our research field—and further that this culture of irreproducibility is one that is actively fostered both by the significant challenges we face in working with health data and by the short-term academic incentives that all too often drive research decisions.

To address this issue, then, we must find a way, both as individual researchers and, more importantly, as a research community, to promote reproducible, scientific research over our current practices. Such changes must be motivated simultaneously by developments in the technical landscape of health ML/AI that make it easier to work with health data in a reproducible fashion than it is to work with it in an irreproducible fashion; development of an active open-source community of shared tools and models; promotion, development, and community engagement in new, distributed forms of benchmarking suited to health data; and finally, changes in our communal incentive structure, as realized through our behaviors in reviewing, hiring, and submitting work reflective of the growing necessity of reproducibility in this field. In the rest of this section, I will comment briefly on each of these layers.

A new kind of standardization In order to ensure reproducibility, we need to shift away from the current status quo of entirely bespoke data standards and arrangements and towards widely used data standards that permit and encourage the use of shared tools, data pre-processing pipelines, and model algorithms. The critical goal of this transition is to ensure that if you write code to train a model on one dataset, that same code could run from scratch on

another dataset with minimal dataset-specific configuration. Note that this does not require harmonization of datasets – many algorithms do not care whether diagnoses are encoded using ICD-9 or ICD-10, or whether a particular observation originally came from one database table vs. another (conditioned on what is being observed, of course). This allows us to leverage greatly simplified standards than existing schemas such as OMOP as we require less reorganization of data to achieve standardization only, thereby greatly lowering the barriers hindering usage of such data standards and enabling more frictionless reproducibility. Critically, such efforts must also be focused on inter-compatibility across tools, contributors, and use-cases, rather than developing internal-only standards that must be used within a certain package, library, or dataset, a property common in existing health ML/AI specific data packages (Wang et al., 2020; McDermott et al., 2023; Saveliev and van der Schaar, 2023). Recent efforts such as the Medical Event Data Standard (MEDS) (Arnrich et al., 2024) epitomize these properties, focusing only on standardizing the minimum amount of information necessary to achieve code transportability alongside active community building across data owners and use-cases to ensure sufficient momentum and adoption for community change.

A new kind of open-source If we adopt a consistent data standard such as MEDS, this permits us to make code release much more impactful as said code can be reliably re-used on new datasets (in contrast, today, such re-use can require a significant amount of code revision). This capability further allows us to create open-source ecosystems of not just full models, but also of model components, task extraction tools, data pre-processing pipeline steps, and other tools. Such communities already reliably exist in other areas of machine learning, such as the `torchvision` (Marcel and Rodriguez, 2010) or `transformers` (Wolf et al., 2020) libraries which provide customizable, composable tools for working with computer vision or natural language processing modalities, respectively. Examples of tools like these already exist in the MEDS ecosystem, including tools for efficient and pythonic MEDS data processing (Steinberg et al., 2024), for deterministic, configurable task selection (Xu et al., 2024), for composing complex data extraction and pre-processing pipelines from isolated units (McDermott et al., 2024), to evaluate predictions made on MEDS datasets (Stankevičiūtė et al., 2024), and to easily train competitive tree-based ensemble baselines for tasks over MEDS datasets (Oufattole et al., 2024). The existence of these tools also highlights the ease of contribution to and development of the MEDS ecosystem, which is critical to ensure individual researchers have the capability to contribute and leverage these tools. But for sustained use, we need a larger community to embrace developing methods and pipelines in a fully reproducible manner while leveraging and contributing to this ecosystem of tools.

A new kind of benchmark With a widely adopted data standard and a communal, open-source ecosystem of models and tools, we can further begin to leverage more comprehensive and informative benchmarks in the field of health AI, much like fields such as computer vision and natural language processing have done. This benchmarking will need to be in a different form — the fragmentation of health data necessitates a distributed kind of benchmarking and the diversity of clinical needs warrants an extensible benchmark suitable to adapt to new tasks — but nevertheless any reliably used benchmarking efforts will be significantly superior to the current state of the field, where benchmarks are not meaningfully used over structured, longitudinal health data. Creating such a benchmark will require several significant changes, including adoption of deterministic, communicable definitions of tasks of interest (e.g., through ACES (Xu et al., 2024) configuration files), community alignment on what diverse sets of tasks are of interest in different clinical settings, and alignment by the community that an increased degree of cross-institution collaboration, by which a collaborator at institution A can train and evaluate from scratch the model of a collaborator at institution B in a feasible, efficient, ethics-board-approved manner, but it is not an infeasible prospect. A preliminary effort towards this end is being explored in the MEDS community through the MEDS-DEV benchmarking effort (Kolo et al., 2024), but such a change will require significant community alignment and contribution.

A new research culture Last, but far from least, we need to change our research culture, at both an individual and an institutional level, to ensure that these changes can be reliably made and can persist long enough such that the increased impact and effectiveness they offer become self-sustaining. At an individual level, this means not publishing papers proposing new models unless your model training results

are reliably reproducible across datasets; not leveraging training tasks that are only defined through inscrutable code and through your prose, but rather are dictated deterministically through some configuration specification; and participating in more cross-institution collaborations to permit multi-site evaluations of proposed algorithms. At an institutional level, we need to enforce significantly greater checks on the reproducibility of published works and be less forgiving of works that propose new model modifications leveraging only private datasets to do so or of works that do not compare against reliable competitors in the literature. Naturally, such changes need to respect the ethical and legal challenges in working with health data, ensuring that data is used in a manner that respects its sensitivity and the privacy constraints of health data. However, given that many of the changes suggested in this perspective can be done without requiring additional data sharing (though our community would naturally greatly benefit from such), I believe these challenges can be navigated while enabling more reproducible, insightful research. These changes will challenge our typical academic incentives in the short term, but in the long term, our field cannot exist without them.

Moving Beyond Longitudinal, Structured Data As indicated in footnote 1, throughout this work, I have largely discussed *longitudinal, structured, medical record data*; however, many of these ideas apply to multimodal health data or other health data as well. For example, longitudinal sequences of images or notes are also very amenable to "MEDS" style standardization and use in communal, distributed benchmarks. Other data modalities unique to health may also benefit from new forms of standardization and an increased focus on reproducibility but have different technical challenges, such as high-frequency, regularly sampled data such as medical waveforms. What is most important, when considering the diverse set of health data modalities, however, is not to let the fact that the general domain has or has not succeeded over a data modality hide our successes or failures in the medical analog of that domain–e.g., we should aim to succeed or fail in the development of health ML/AI solutions beyond simply in re-purposing known successes from the general domain. Health data is rich and has many nuances not found as commonly in other areas of ML/AI; irregularly sampled, continuous time sequences of complex events, fully observed 3D data, high-frequency spatio-temporal data over complex geometries, and more properties are all highly important in health-care and much less common in general domain ML/AI, and we should leverage these properties as opportunities to develop the computational and informatics sciences to succeed in working with these modalities, even if doing so requires us to develop new data standards, open source communities, and research incentives.

Conclusion

In this perspective, I have argued that the "Science of machine learning / artificial intelligence for health" is both important and drastically under-studied in our field today. Further, I state that this failure is a natural consequence of our field's significant reproducibility crisis and the academic incentives that promote that crisis. While at times these concerns can seem small, I claim that the reality is that if we don't fix them, we will effectively cede ownership and development of the future advances of AI on healthcare. We will not be able to help hospitals and colleagues solve problems at their health systems, we will not be able to effectively study systems that require significant engineering overhead to create and explore, and we will not be able to answer the interesting technical questions that seed the field of ML/AI in healthcare. And lastly, thankfully, I hold that while these concerns are daunting, we actually are empowered to fix these concerns. Through the use of new, simplified, minimal standards, an open-source ecosystem of shared tools and models, new kinds of benchmarks suited to the fragmented landscape of health, and finally with changes to our research culture, we can help engender a more seamlessly reproducible, scientific field that realizes more of the impact we set out to achieve.

Acknowledgments

This paper benefited significantly from many contributions, including conversations with Olawale Salaudeen; the entire MEDS working group, including but not limited to Jason Fries, Ethan Steinberg, Tom Pollard, Edward Choi, Robin van de Water, Patrick Rockenschaub, Pawel Renc; Shalmali Joshi; Deb Raji; Maya Peterson; Stephanie Hyland; Haoran Zhang; Marzyeh Ghassemi; Isaac Kohane; Noa Dagan; and many others. In addition, MBAM gratefully

acknowledges support from a Berkowitz Postdoctoral Fellowship at Harvard Medical School.

References

Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew B.A. McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (MEDS): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024. URL https://openreview.net/forum?id=IsHy2ebjIG.

Pierre J Chambon, Christopher Wu, Jackson M Steinkamp, Jason Adleberg, Tessa S Cook, and Curtis P Langlotz. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *Journal of the American Medical Informatics Association*, 30(2):318–328, 11 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac219. URL https://doi.org/10.1093/jamia/ocac219.

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

Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4): 578–582, 05 2014. ISSN 1067-5027. doi: 10.1136/amiajnl-2014-002747. URL https://doi.org/10.1136/amiajnl-2014-002747.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

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.

Health Level Seven International. Fhir specification, 2024. URL https://www.hl7.org/fhir/. Accessed: 2024-10-29.

Daniel Jarrett, Jinsung Yoon, Ioana Bica, Zhaozhi Qian, Ari Ercole, and Mihaela van der Schaar. Clairvoyance: A pipeline toolkit for medical time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xnC8YwKUE3k.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/johnson17a.html.

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 214–221, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384455. URL https://doi.org/10.1145/3368555.3384455.

Rohan Kodialam, Rebecca Boiarsky, Justin Lim, Aditya Sai, Neil Dixit, and David Sontag. Deep contextual clinical prediction with reverse distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 249–258, 2021.

Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A. Fries, Jeffery N. Chiang, Jungwoo Oh, Justin Xu, Kamilė Stankevičiūtė, Kiril V. Klein, Matthew B. A. McDermott, M. Odgaard, Nassim Oufattole, Nigam H. Shah, Patrick Rockenschaub, Pawe Renc, Robin P. van de Water, Shalmali Joshi, Simon A. Lee, Teya S. Bergamaschi, Tom J. Pollard, Vincent Jeanselme, and Young Sang Choi. The meds decentralized, extensible validation benchmark: Establishing reproducibility and comparability in ml for health, October 2024. URL https://github.com/mmcdermott/MEDS-DEV. original-date: 2024-05-20T19:50:35Z.

Marc Liberman. Reproducible research and the common task method. *Simmons Foundation Lecture*, 2015.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceed-*

ings of the 18th ACM International Conference on Multimedia, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL https://doi.org/10.1145/1873951.1874254.

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*, CHIL '21, page 257–278, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383592. doi: 10.1145/3450439.3451877. URL https://doi.org/10.1145/3450439.3451877.

Matthew McDermott, Pawel Renc, and Nassim Oufattole. mmcdermott/MEDS_transforms, October 2024. URL https://github.com/mmcdermott/MEDS_transforms. original-date: 2024-05-17T15:19:57Z.

Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586): eabb1655, 2021b. doi: 10.1126/scitranslmed.abb1655. URL https://www.science.org/doi/abs/10.1126/scitranslmed.abb1655.

Matthew B.A. McDermott, Bret Nestor, Peniel N 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 *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=hi00735tmc.

Katherine Miller. Healthcare Algorithms Don't Always Need to Be Generalizable, June 2022. URL https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable.

Karthik Murugadoss, Ajit Rajasekharan, Bradley Malin, Vineet Agarwal, Sairam Bade, Jeff R Anderson, Jason L Ross, William A Faubion, John D Halamka, Venky Soundararajan, et al. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, 2(6), 2021.

Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M. Stultz, and Matthew B. A. McDermott. Meds-tab: Automated tabularization and baseline methods for meds datasets, 2024. URL https://arxiv.org/abs/2411.00200.

Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*, 31(3):583–590, 01 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad247. URL https://doi.org/10.1093/jamia/ocad247.

Olawale Salaudeen and Moritz Hardt. Imagenot: A contrast with imagenet preserves model rankings, 2024. URL https://arxiv.org/abs/2404.02112.

Evgeny S Saveliev and Mihaela van der Schaar. Temporai: Facilitating machine learning innovation in time domain tasks for medicine. *arXiv preprint arXiv:2301.12260*, 2023.

Kamilė Stankevičiūtė, Vincent Jeanselme, and Nassim Oufattole. MEDS Evaluation, October 2024. URL https://github.com/kamilest/meds-evaluation. original-date: 2024-08-07T04:13:17Z.

Ethan Steinberg, Michael Wornow, Suhana Bedi, Jason Alan Fries, Matthew B. A. McDermott, and Nigam H. Shah. meds_reader: A fast and efficient ehr processing library, 2024. URL https://arxiv.org/abs/2409.09095.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 10 2020. doi: 10.1093/jamia/ocaa139.

Kavishwar B Wagholikar, Layne Ainsworth, David Zelle, Kira Chaney, Michael Mendis, Jeffery Klann, Alexander J Blood, Angela Miller, Rupendra Chulyadyo, Michael Oates, William J Gordon,

Samuel J Aronson, Benjamin M Scirica, and Shawn N Murphy. I2b2-etl: Python application for importing electronic health data into the informatics for integrating biology and the bedside platform. *Bioinformatics*, 38(20):4833–4836, 09 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac595. URL https://doi.org/10.1093/bioinformatics/btac595.

Shirly Wang, Matthew B. A. 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*, CHIL '20, page 222–235, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384469. URL https://doi.org/10.1145/3368555.3384469.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Alan Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=CsXC6IcdwI.

Justin Xu, Jack Gallifant, Alistair E. W. Johnson, and Matthew B. A. McDermott. Aces: Automatic cohort extraction system for event-stream datasets, 2024. URL https://arxiv.org/abs/2406.19653.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5788–5789, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599178. URL https://doi.org/10.1145/3580305.3599178.