Recent Advances, Applications and Open Challenges in Machine Learning for Health: Reflections from Research Roundtables at ML4H 2022 Symposium

Stefan Hegselmann[1] stefan.hegselmann@uni-muenster.de
Helen Zhou[1] hlzhou@andrew.cmu.edu
Yuyin Zhou[1] yzhou284@ucsc.edu
Jennifer Chien[2] jjchien@eng.ucsd.edu
Sujay Nagaraj[2] s.nagaraj@mail.utoronto.ca
Neha Hulkund[2] nhulkund@mit.edu
Shreyas Bhave[2] sab2323@cumc.columbia.edu
Michael Oberst[2] moberst@mit.edu
Amruta Pai[2] ap52@rice.edu
Caleb Ellington[2] cellingt@cs.cmu.edu
Wisdom Ikezogwo[2] wisdomik@cs.washington.edu
Jason Xiaotian Dou[2] jason.dou@pitt.edu
Monica Agrawal[2] magrawal@mit.edu
Changye Li[2] lixx3013@umn.edu
Peniel Argaw[2] peniel@g.harvard.edu
Arpita Biswas[2] arpitabiswas@hsph.harvard.edu
Mehak Gupta[2] mehakg@udel.edu
Xinhui Li[2] xli993@gatech.edu
Marta Lemanczyk[2] marta.lemanczyk@hpi.de
Yuhui Zhang[2] yuhuiz@stanford.edu
Christian Garbin[2] garbin.christian@gmail.com
Elizabeth Healey[2] ehealey@mit.edu
Heejong Kim[2] hek4004@med.cornell.edu
Claire Boone[2] claireboone@uchicago.edu
Roxana Daneshjou[3] roxanad@stanford.edu
Siyu Shi[3] sshi@fprimecapital.com
Nicola Pezzotti[3] nicola.pezzotti@gmail.com
Stephen R. Pfohl[3] spfohl@google.com
Edwin Fong[3] chef@novonordisk.com
Aakanksha Naik[3] aakankshanaik19@gmail.com
Ben Lengerich[3] blengeri@mit.edu
Ying Xu[3] yingxuuc@gmail.com
Jonathan Bidwell[3] jonathan.bidwell@ochsner.org
Mark Sendak[3] mark.sendak@duke.edu
Byung-Hak Kim[3] byunghakk@gmail.com
Nathaniel Hendrix[3] nhendrix@theabfm.org
Dimitris Spathis[3] dimitrios.spathis@nokia.com
Jun Seita[3] jun.seita@riken.jp
Bastiaan Quast[3] bastiaan.quast@itu.int
Megan Coffee[3] megan.coffee@nyulangone.org
Collin Stultz[3] cmstultz@csail.mit.edu
Irene Y. Chen[1] iychen@csail.mit.edu
Shalmali Joshi[1] shalmali@seas.harvard.edu
Girmaw Abebe Tadesse[1] girmaw@ieee.org

[1] Organizing committee for ML4H 2022 Research Roundtables, [2] Junior Chairs, [3] Senior Chairs
We asked all roundtable chairs to opt-in as co-authors; 11 chairs did not opt-in.

1. Introduction

The second Machine Learning for Health (ML4H) symposium was held both virtually and in-person on November 28, 2022, in New Orleans, Louisiana, USA (Parziale et al., 2022). The symposium included research roundtable sessions to foster discussions between participants and senior researchers on timely and relevant topics for the ML4H community. Encouraged by the successful virtual roundtables in the previous year (Roy et al., 2021), we organized nine in-person and four virtual roundtables at ML4H 2022 (Parziale et al., 2022). A roundtable session included invited senior chairs (with substantial experience in the field), junior chairs (responsible for facilitating the discussion), and attendees from diverse backgrounds with interest in the session's topic. This document explains the organization process we used and compiles the takeaways from the roundtable discussions, including recent advances, applications, and open challenges for each topic. We conclude with a summary and lessons learned across all roundtables.

2. Organization Process

We identified potential roundtable topics from papers in the ML for health domain published in the last three years and we pooled suggestions from ML4H chairs and keynote speakers. After removing duplicates, there were 18 unique topic candidates. These topics were entered into a Google poll, which was broadcasted on Twitter to solicit feedback from the ML4H community. Between July 29th and September 30th 2022, 39 people answered the poll. The final votes are presented in Figure 1. For each of the top-ranked topics, we invited senior chairs with expertise in the respective field and aimed for two to three senior chairs for the in-person and one senior chair for the virtual roundtables. Next, we identified junior chairs that preferably had some experience in the discussed topic. Before the event, junior and senior chairs wrote an introduction paragraph shared on the ML4H website[1] and submitted three to five potential discussion questions. On the day of the symposium, we had two 25-minute slots for roundtables with a five-minute break to allow participants to join another roundtable session. After the event, we asked the chairs to write a summary of the main takeaways from the discussion.

3. Research Roundtables

We successfully recruited chairs for nine in-person and four virtual research roundtables. For the in-person roundtables we included the following topics:

1. Are our ML models really making an impact in the hospital? What do caregivers and clinicians want and what is still missing?

2. Evaluation of healthcare data prior to applying ML, e.g., representation analysis, annotation quality, Out of Distribution detection (OOD), clusters of In-distributions (IDs)

3. How to ensure generalizability of ML in healthcare?

4. How do we inject domain knowledge into deep learning (DL) models, in particular when not much data is available?

5. How to effectively integrate multiple data sources (e.g., Electronic Health Records (EHRs), images, genomics) for ML applications in healthcare?

6. How can we utilize foundation models (very large pre-trained models) for healthcare?

[1] https://ml4health.github.io/2022/

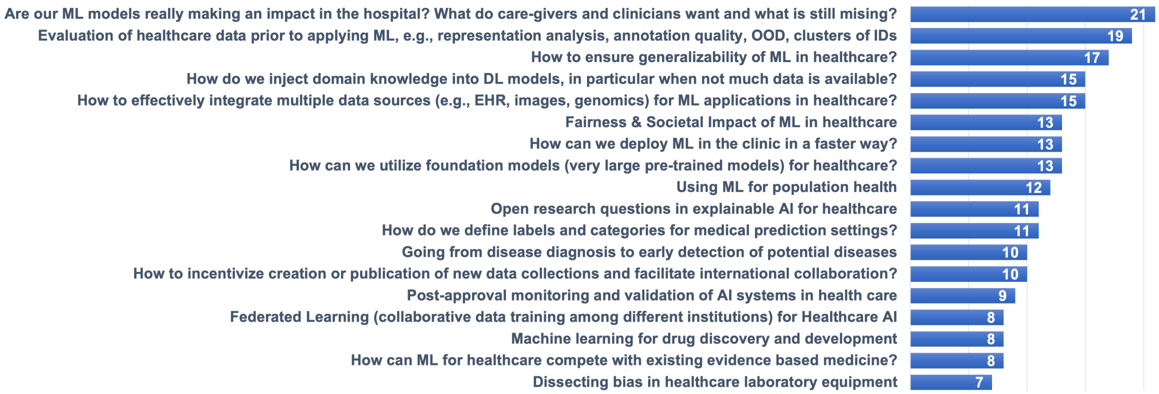

Figure 1: Results of public poll for roundtable topics filled out by 39 participants.

7. Using ML for population health

8. How to incentivize creation or publication of new data collections and facilitate international collaboration?

9. Post-approval monitoring and validation of ML systems in health care.

For the virtual roundtables we selected:

1. How ML can help prevent and respond to infectious disease outbreaks. Where are we now?

2. Are our ML models really making an impact in the hospital? What do care-givers and clinicians want and what is still missing?

3. How to effectively integrate multiple data sources for machine learning applications in healthcare?

4. How to evaluate of healthcare data and labels prior to applying machine learning?

We tried to identify senior and junior chairs for the top-rated topics (see Figure 1). However, for some of them, we could not identify suitable chairs. Hence, there are topics with many votes that were not included in the symposium. Also, in-person and virtual topics were allowed to overlap due to different participants. All chairs handed in the introductions and discussion questions before the event and prepared a summary afterwards.

In the following sections, we provide slightly revised versions of the introductions and summaries from all roundtables. Information from in-person and virtual roundtables on the same topics was combined.

3.1. Are our ML models really making an impact in the hospital? What do care-givers and clinicians want and what is still missing?

Chairs: *Jennifer Chien, Sujay Nagaraj, Roxana Daneshjou, Siyu Shi, Elizabeth Healey, Collin Stultz*

Background: Machine learning models have shown tremendous promise in advancing clinical care. However, adaptation in clinical practice has been slow due to the numerous barriers that exist in implementation. ML models can impact healthcare in different ways such as relieving workload through medical note summarization, augmenting clinical workflows through intensive care unit alert systems, and providing highperforming diagnostics and risk-stratification (Aung et al., 2021). Despite the promised potential of these methods, concerns about the clinical utility of these models hinder widespread adoption in hospital systems (Wilkinson et al., 2020). Further, despite FDA approval of many clinical ML models, few randomized controlled trials have been conducted to evaluate the utility of these interventions (Plana et al., 2022). In addition, there are potential harms that can arise if ethics, fairness, and privacy considerations are ignored (Kelly et al., 2019; Rajpurkar et al., 2022). We draw attention to the translation gap between theoretical/proof-of-concept ML research and clinical integration in the hospital (Seneviratne et al., 2020), and pose the following questions: What is the current landscape of ML models in hospitals? What are common problematic simplifying assumptions? How can we ensure various stakeholders are engaged in this pipeline? Who are these technologies helping, and who are they hurting?

Recent work in explainability, data visualization, and human-computer interaction has taken steps towards providing what caregivers and clinicians want (Tonekaboni et al., 2019; Ghassemi et al., 2021). However, many voices in the healthcare system may not be prioritized (e.g. nurses, allied health care workers, and especially patients). We also explore how ML fits into the healthcare pipeline, who gets to contribute, and what are key areas of focus moving forward.

Discussion: The role of ML in healthcare, whether as an automation or augmentation technology, was discussed in this session. Identifying the right problems requires expertise in healthcare, humility, and awareness. Though the healthcare system can work quite well, there are still many opportunities for augmentation. Medicine is becoming more complex and siloed, with specialists pursuing niche roles. There may be opportunities to bridge these gaps by allowing specialists to better perform in roles outside of their skill set. For example, giving a family doctor the tools to perform diagnosis at a subspecialist's level. The goal here is not to subsume other roles, but rather to improve baseline performance of all doctors. However, we must remain careful about whether the tools are truly improving performance or are just giving the illusion of confidence to a doctor. AI predictions can bias a doctor to perform worse than they would have without that prediction. This is where the idea of low-hanging fruit can be useful—there exist tons of important, simple problems that may not be as sophisticated as a sepsis-prediction tool but can still provide non-zero utility for a clinician. Two examples mentioned were: 1) an algorithm for patients to take better dermatology photos, 2) helping histopathologists count the number of cells in an image. These require simple solutions, with deterministic goals and consistent models, yet provide much utility for a clinician.

The other role that came up was that of labour augmentation. Especially now, there is increasing strain on the healthcare system, and algorithms could empower people to perform tasks they are not trained to do in order to improve efficiency. For example, empowering workers without sonography training to improve their ultrasound skills with ML-based gamified applications.

Lastly, we discussed what providers cannot currently do. One such role for ML is combining knowledge from multiple modalities - such as proteomics, genomics, mobile health, and other such sources of data.

We also considered how impact and importance could be measured in healthcare. This question elicited a sobering discussion of who truly measures impact in this space. As much as the roundtable agreed that patient outcomes should drive decision-making, in reality, it boils down to economics, an insight repeatedly shared among participants from venture capital and big tech. Hospitals, particularly in the US, work under razor-thin margins and most executives will expect solutions that will save (if not make) money. The companies in ML4H that are doing the most, are those that are saving hospitals money or are better optimizing the fee-for-service system (more billable services for a provider). The question people are asking is not "is my model technically good enough?" but "how do we incorporate it into a fee-for-service model?".

Another topic was how ML models can be explainable, and gain clinicians' trust. We discussed the difference between finding ways to explain ML models once they are built and designing ML models explicitly for explainability, with the latter being desirable. We talked about instances of successful implementations of ML models in hospitals, with examples from image recognition, and some of the reasons related to slow adoption of other models. Many researchers who joined the roundtable were interested in learning the practical steps they can take to make their research translational. The unifying theme of the discussion was that the best way to design ML models for useful deployment in hospitals is to involve clinicians early in the project.

3.2. Evaluation of healthcare data prior to applying ML, e.g., representation analysis, annotation quality, OOD detection, clusters of IDs

Chairs: *Neha Hulkund, Shreyas Bhave, Nicola Pezzotti, Pin-Yu Chen, Claire Boone, Ziad Obermeyer*

Background: Healthcare datasets often exhibit many challenging properties including non-random missingness processes, label noise, and high dimensionality (Ghassemi et al., 2020). These datasets also often reflect societal biases and algorithms trained using this data may exacerbate disparities (Obermeyer et al., 2019). Each of these properties is further complicated when considering scenarios where there may be dataset shift. Given these challenges, it is important to systematically evaluate datasets for the existence and severity of such pathologies prior to applying any ML methods. A related task is interrogating the robustness of methods to the introduction of different dataset shifts in these pathologies (Speakman et al., 2023; Subbaswamy et al., 2021). Several types of approaches exist for this kind of data exploration, including dimensionality reduction techniques for visualizing clusters of data (Pezzotti et al., 2019), evaluating and improving label quality using active learning (Chen et al., 2013), and tests for missingness assumptions. One purpose of this session is to discuss existing methods used by the community for exploration of healthcare data, explore what pathologies are rarely addressed or difficult to interrogate using existing methods and how we can mitigate common pitfalls in the analysis of large healthcare datasets.

Another topic for discussion is label definition. We often use ML to predict important but challenging to define concepts like health, or healthcare needs. When empirical definitions are hard to determine, an obvious next step is to use a proxy label (Mullainathan and Obermeyer, 2021). One example from healthcare is using patient cost as a label when we want to understand future healthcare needs. Previous work has shown that not only is cost a bad label for health, but it can be biased. An algorithm that was deployed by a large healthcare organization systematically predicted lower future health needs among Black patients compared to White patients in part because the

Black patients had systematically lower past spending (Obermeyer et al., 2019). Low past spending, in this case was correlated with higher health needs, not lower.

Discussion: We often want to use algorithms to make the world more ideal—for example, providing the best quality diagnosis for a patient. However, available data instead reflects the world as it is, not how it should be, and algorithms end up reproducing that when they are built using said data. The data generating process in healthcare depends on people—for example a physician deciding to write down a note—and people make mistakes and can be biased. For these reasons, evaluating datasets for selection bias and choosing appropriate labels is highly important. Doing this can be challenging without a deep understanding of healthcare context and medical expertise. Thus, working with multidisciplinary teams is necessary, and taking the time to invest in communication with collaborators from other disciplines is crucial to producing high quality work.

We also discussed various topics such as data exploration, the unique nature of health data, and creating ML tools that are ready for deployment. A senior chair shared their insights on checking datasets for any attacks or adversarial noise. Furthermore, he brought up the major tradeoffs between robustness, accuracy, and privacy inherent to all models and suggested a possible solution to be a tuneable multi-objective function keeping these tradeoffs in mind, allowing practitioners to make these tradeoffs themselves. Another chair brought up the idea of never trusting the first model you train, instead going to the first real-world deployable setting, testing the model there, and going back to the model to make adjustments accordingly. This led us to the idea of distribution shifts, where a discussant shared challenges of merging data from different hospitals, due to legacy factors such as templates and notation. Another senior chair mentioned a possible solution could be trying to build foundation models for EHRs, since masked language models have been useful in getting tokens for missing data and for internal temporal shifts that are more difficult for small models to work with. We ended our discussion by talking about the challenges of getting ML for healthcare models deployed at the bedside, with challenges in generalization, regulation, and long time-scales.

3.3. How to ensure generalizability of ML in healthcare?

Chairs: *Michael Oberst, Amruta Pai, Emmanuel Candès, Stephen R. Pfohl, Edwin Fong*

Background: Despite being deployed widely in practice, predictive models in healthcare can fail to perform well across clinics, patient populations, and time. Two recent examples illustrate the challenge: The University of Michigan Hospital deactivated their sepsis prediction model in April of 2020 due to COVID-related changes in the distribution of patients (Finlayson et al., 2021), and the widely-deployed Epic Sepsis Model was found (in an external validation study) to dramatically under-perform relative to the claims of the developer (Wong et al., 2021). Meanwhile, a recent study found that only 23% of ML-based healthcare papers used multiple datasets (McDermott et al., 2021).

A definition of generalizability is the ability of a model to perform well on data from a cohort of patients independent from its training data, whether from a different hospital or demographic subpopulation, or a different point in time. In this roundtable, we will discuss different aspects of what it means for a model to generalize well, the challenges inherent in building generalizable models, and potential paths forward.

Discussion: Although ML has made enormous strides in various healthcare applications, the focus has been on achieving good accuracy on individual datasets without rigorous external cross-validation. Recent works have found that models trained for the same tasks but on different datasets have low generalizability with drastic performance degradation when there is a distribution shift. However, a culture of sharing datasets and creating benchmark datasets to study generalizability is necessary to encourage external cross-validation. A culture of auditing models to identify regions where the model fails is essential.

Models trained on data with operational features such as hospital-specific protocols may not always generalize. Thus, methods need to be built to adapt these models to hospital-specific needs. An approach is to build supermodels trained on datasets from different sources, modalities, and hospitals and fine-tune them on individual subgroups before deployment.

Another approach is to develop uncertainty quantification measures to augment point estimates with confidence intervals. Health professionals can use uncertainty to trust the model's performance across different sub-groups. When the uncertainty of the estimates is high, the model would need to be retrained.

Generalization can be limited due to the dataset as well. A classic example is clinical trials. Recruitment in clinical trials is conducted with inclusion criteria often based on convenience. Hence the average treatment effects are biased. However, emerging registry trials can reach a broader population base leading to more diversity in the study population and subsequent dataset.

3.4. How do we inject domain knowledge into DL models, in particular when not much data is available?

Chairs: *Caleb Ellington, Wisdom Ikezogwo, Aakanksha Naik, Ben Lengerich, Ying Xu, Eran Halperin*

Background: When sampling is difficult or expensive, as is often the case in medicine, biology, language, and the social sciences, we typically obtain only a small dataset relating to any particular modeling or prediction task. In these situations, we can provide domain knowledge as an inductive bias to inform learning tasks and restrict the possible solution spaces of models, increasing our sample efficiency in these already sample-sparse domains. Common examples are knowledge graphs, but this representation is often unavailable, and knowledge in some domains may not align well with a graphical representation.

In this session, we focus on why domain knowledge is important (and when it might not be), and discuss the many ways of injecting domain knowledge when it is relevant. Our panelists represent diverse backgrounds from natural language processing, bio-informatics, and computational medicine, and each has unique experiences utilizing domain knowledge. We aim to provide a comprehensive overview of existing approaches and thrilling discussion about future research on injecting domain knowledge into DL.

Discussion: Domain knowledge is an umbrella term for the entities, associations, systems, rules, and goals that define a field of study. As such, there are a myriad of ways to represent domain knowledge beyond simple associative knowledge graphs, spanning a variety of constraints and structures, including scientific text corpora, fact banks, models pre-trained on related tasks, priors, regularizers, independence structures, as well as domain-aligned metrics that go beyond accuracy to evaluate a model's representation of a real-world system.

Expert-in-the-loop model design should be a gold standard for instilling domain knowledge and ensuring proper alignment of deep learning models with real-world objectives, but this approach is time-consuming and sometimes expensive. At a minimum, experts need to be included in the evaluation of a model for domain applicability. Using a set of expert-derived rules and priors can help instill domain knowledge during early development without requiring constant interaction with experts. Synthetic or imputed data can effectively incorporate domain knowledge through large volumes of unstructured data and align with existing formats. However, it is a debatable practice since the rules, principles, or models used to create this data are always a richer representation of knowledge than the derived data. For non-objective tasks, observer variability in medicine demands the need for experts-in-the-loop as most relevant applications predict ranges of observations rather than absolutes. Navigating uncertainties about multiple sets of competing domain knowledge, without knowing which sets are relevant or irrelevant for our tasks can be very expensive without an expert to help tease out differences between conflicting knowledge incorporating priors. This could be related to explainability on the question of conflicting domain knowledge. One major open question is iterative inclusion of domain knowledge, or inclusion of new knowledge after an expensive system has already been trained. We discussed one option, targeted parameter editing based on new facts (De Cao et al., 2021), but believe this is an interesting future direction. Another discussion point was the collection of up-to-date and new domain knowledge. For example, at the beginning of the COVID pandemic.

3.5. How to effectively integrate multiple data sources (e.g., EHR, images, genomics) for ML applications in healthcare?

Chairs: *Jason Dou, Jonathan Bidwell, Dominik Dahlem, Mark Sendak, Heejong Kim, Mert R. Sabuncu*

Background: Healthcare datasets are inherently multi-modal and collected from multiple sources (Lance et al., 2022; Wang et al., 2022; Mao et al., 2022). For example, medical datasets include electronic health records, medical imaging, lab tests, genetics, and patient demographics gathered from various sources and datasets from internet of things (IoT) devices like watches collecting fitness and health statistics data. Consolidating such datasets is an essential step for data analysis. Also, a potential avenue for research is to move the AI solution upstream with the aim to model entire decision processes end-to-end. This research roundtable discussed the challenges, opportunities, and possible solutions of healthcare data integration for ML applications.

Discussion: We discussed the advantages of data integration. In the clinical setting, today, much of the data interpretation is done by specialists. For example, radiologists read the medical imaging data and summarize findings in a report. However, physicians need to integrate these summaries (e.g. reports, test results, etc) to make important treatment decisions. This is critical for delivering optimal care. Since this way of delivering care is dependent on specialist reads and communication between physicians, ML offers the potential of positively impacting medicine by accessing raw multimodal data and optimizing the communication and integration between specialties. We also talked about advantages of deep learning approaches. The flexibility of the deep learning framework allows us to try numerous strategies for integrating data. Healthcare data is inherently multi-modal and acquired from multi-sources. Techniques like early fusion and late fusion can be simple but effective solutions for integrating healthcare data. Also, potential roadblocks of integrating healthcare data were mentioned. In real-world clinical settings, data is often collected as needed. For example, additional clinical tests will only be prescribed if basic diagnostic tests are not enough and further examination is necessary. This can cause sparsity and bias in the integrated datasets. This also makes it difficult to evaluate the necessity of data from additional experiments.

Senior chairs were asked about the biggest challenges in effectively integrating multiple data sources for ML applications in healthcare. One chair thought the biggest challenges arise between different data owners and stakeholders with different cultures and values. Another senior chair mentioned the challenge that adding data sources leads to additional complexity. This concurs with the cultural challenge mentioned before. Also, one chair reported their experience using a combination of claim and structural data. For instance, privacy issues due to the centralization of data might arise. Another chair proposed that innovation for simplicity is among the biggest challenges. Indeed more data sources can provide more information, but it remains unclear how we can obtain simple principles and guidelines. For example, we discussed how to design the reward function and optimal strategy if we want to apply reinforcement learning with a value function based on economic and social value.

3.6. How can we utilize foundation models (very large pre-trained models) for healthcare?

Chairs: *Monica Agrawal, Changye Li, Payel Das, Zachary C. Lipton, Byung-Hak Kim*

Background: A recent paradigm in ML is the concept of *foundation models*, very large pre-trained models that can be used as a base model for applications to build on top of (Bommasani et al., 2021). Such models have seen particularly impressive performance at zero- and few-shot generation tasks in text and vision (Brown et al., 2020; Radford et al., 2021). Given the 'emergent abilities' of such models (Wei et al., 2022), they hold remarkable promise for transforming the way we practice medicine and transform care. For example, large language models (LLMs) have already shown potential recently at tasks including clinical information extraction and medical exam question answering (Agrawal et al., 2022; Chung et al., 2022). However, there are practical questions and challenges that arise due to the high stakes of the clinical setting. In this roundtable session, we will discuss "How can we utilize foundation models (very large pre-trained models) for healthcare?"

Discussion: The roundtable started with participants discussing what made something a "foundation model" before moving on to where such models could be impactful. One area of interest in NLP was retrospective interpretation of clinical notes, to track trends in disease status over time. Another was AI-assisted clinical note writing, which could lift clerical burden off of clinicians. Others did bring up the concern that this could result in clinicians being less cognizant and in tune with what was being written. Another noted that a site-specific burn-in time should be relatively standard, from a regulatory perspective.

One senior chair posited that foundation models could make the largest impact where deep models have been shown to be transformative, but might struggle where older, statistical models are still close to optimal (e.g. survival modeling). Participants discussed if there could exist a foundation model for other modalities of health data (e.g. waveforms) as opposed to imaging alone, or if there were insufficient similarities for a shared representation to benefit from. The roundtable concluded with participants discussing what areas of medicine might be more generalizable.

3.7. Using ML for population health

Chairs: *Peniel Argaw, Arpita Biswas, Nathaniel Hendrix, Dimitris Spathis*

Background: Population health, in general, can be defined as the health outcome distributions within and across populations (Kindig and Stoddart, 2003). This research requires taking into account various cultural, social, and environmental factors and investigating their effects on the health of communities. To this end, ML is emerging as a possible way to automate complex tasks in population health that otherwise have required substantial human labor. ML utilizes these health outcomes to understand patterns of health determinants in order to identify high-risk groups and predict future disease burdens (Morgenstern et al., 2020; Ogallo et al., 2021). These models can be especially of interest to policy-makers to recommend public health policies and interventions. In this roundtable, we discuss the challenges in data quality and dimensionality, the emergence of new technologies and applications in the field, and the design of evaluation metrics in order to push for clinically and cost-effective models.

Discussion: We had an interesting discussion on ML for population health revolving around data quality and dimensionality. Data quality was a major topic of the session, where the senior chairs and participants discussed data quality assessment, resultant issues, and limitations. Trust in self-reported data, aligning ground truth to signal data and the variance in datasets resulting from the variance in populations were all points of reflection. Moreover, the topic of data dimensionality, particularly in low-resource settings, was discussed as members were engaged in the topics of heterogeneity in data and devices, transparency across populations and possibilities of transfer learning and multi-modality. Overall, the technologies and applications of ML for population health are exciting but challenging given the difficulty in building clinically and cost-effective models for heterogeneous populations.

3.8. How to incentivize creation or publication of new data collections and facilitate international collaboration?

Chairs: *Mehak Gupta, Xinhui Li, Jun Seita, Bastiaan Quast*

Background: While the potential for the application of ML in health is enormous, so are the challenges. Crucially, ML models require large amounts of training data, yet making datasets available while respecting patients' privacy has proven to be an immense conundrum (Liang et al., 2022). Furthermore, though we are constantly reminded that health issues know no borders, whether it be outbreaks of infectious diseases or rare genomic disorders, the jurisdictions governing health data most certainly do. As such, the global sharing of health data for the benefit of the citizens of the world remains elusive.

This roundtable will distill and articulate the exact bottlenecks that currently block progress in the greater sharing of health data in a responsible manner. Building on the precisely articulated bottlenecks, the roundtable will discuss the challenges and propose suggestions to incentivize creation or publication of new data collections and facilitate international collaboration.

Discussion: Our roundtable discussion covered challenges and suggestions to incentivize creation or publication of new data collections and facilitate international collaboration.

First of all, participants brought up that we need to show people the value of ML applications in the health domain in order to motivate people to invest more resources to develop the ecosystem. Specifically, the main benefit of ML is that it can assist with the diagnosis process and make diagnostics more efficient and affordable at earlier stages, thus leading to earlier detection and better treatment outcomes.

Next, we discussed the techniques to protect patient privacy. Particularly, we can utilize data encryption techniques and federated learning (McMahan et al., 2017; Yang et al., 2019). For example, the iDASH (Kuo et al., 2020) competition has demonstrated how homomorphic encryption can be used to perform computations on rare genomic conditions, with data from different jurisdictions, without revealing any private patient data. In addition, instead of sharing the data directly, we can distribute the ML models by building an Application Programming Interface (API) to share the models.

Then, we shared our experiences regarding data collection and sharing problems in developing countries. Participants mentioned that some hospitals do not store medical data such as X-rays after 6 months due to storage limits; some hospitals still use paper healthcare records instead of electronic healthcare records. Prior to sharing the datasets across countries, it is crucial to improve the local system for data maintenance and digitization.

Finally, we discussed practical suggestions to incentivize data collection at the individual level and at the international organization level. At the individual level, we can encourage participants to share their data and obtain informed consent prior to data collection. At the international organization level, we can learn and follow how international institutions collaborate to collect and share the data from international scientific research projects, such as the Medical Information Mart for Intensive Care Project (MIMIC) (Johnson et al., 2016), the Human Genome Project (HGP) (Consortium, 2001) and the PRIMatE Data Exchange Project (PRIME-DE) (Milham et al., 2018).

3.9. Post-approval monitoring and validation of AI systems in health care

Chairs: *Marta Lemanczyk, Yuhui Zhang, Berkman Sahiner, Anna Decker, Harvineet Singh*

Background: The rapid development and application of AI systems in healthcare have raised a wide range of concerns about their reliability. Even if an AI system is approved, it can become brittle when deployed in real-world scenarios, for instance, due to distribution shifts (Oakden-Rayner et al., 2020; Zhang et al., 2022; Kim et al., 2022). Therefore, post-approval monitoring and validation is a critical quality assurance process to ensure that these systems deliver accurate medical predictions in diverse real-world scenarios.
In this roundtable, we will discuss some of the challenges regarding the post-approval monitoring and validation process and how to address these issues from a policy, technical, and data perspective. In addition, we

will discuss the best way to establish collaboration between clinicians, ML experts, and the systems themselves to communicate these issues.

Discussion: In this roundtable, we discussed the opportunities and challenges of post-approval monitoring and validation of AI systems in health from both policy and technical perspectives. In terms of policy, the initial FDA approval process is similar to the processes for other devices and depends on the risk of the given AI system. For high-risk devices, it will typically require at least a year for monitoring and scientific studies to be reported. Now, the FDA has established a new policy that enables AI systems to be updated after approval, a much faster and less burdensome process than the initial approval process. There are also challenges from the technical side. Post-approval data collection is a large effort because standards and protocols and also supporting tools for those tasks are lacking. Additionally, the data must capture the requirements for the chosen evaluation metrics. Other challenges that were mentioned are distribution shifts in input data as well as outcomes. Monitoring shifts is crucial since they can lead to a decrease in performance and therefore to safety concerns for patients.

To achieve successful post-approval monitoring and validation, academia and industry should increase collaborations, for instance, by combining promising techniques from academic researchers and valuable clinical data from providers in the industry. Knowledge sharing and open communication are key to achieving goals.

3.10. How ML can help prevent and respond to infectious disease outbreaks. Where are we now?

Chairs: *Christian Garbin, Megan Coffee*

Background: We can trace the origins of infectious disease modeling to Bernoulli's smallpox models in 1760. Modeling became increasingly sophisticated in the 20th century with the mass action law, the Kermack-McKendrick epidemic model, stochastic models, and other advancements in modeling. In the 21st century, we have seen the rise of ML techniques powered by the surge in data collection and cheaper and cheaper computing resources.

Besides modeling the spread of diseases, ML has been applied to other areas, such as diagnosing, triaging, and predicting outcomes at the patient level, accelerating the study of protein structures for vaccine development, reducing the time to develop tests, detecting and suppressing misinformation, and predicting future outbreaks caused by complex environmental and climate changes. In this roundtable, we will discuss "how can ML help prevent and respond to infectious disease outbreaks?".

Discussion Climate and environmental change, along with mobility and structural factors, increase the risk of new outbreaks, like COVID-19, which will require new tools to prevent and respond to. In the initial stages, we suffer from the "fog of war". Minimal, noisy data is available; thus, models are limited. Although it would be ideal to have centralized data collection, clinicians' focus at this stage is to treat and contain current cases. Data is collected in the simplest format that can support the daily work, subject to current health system limitations. The goal of early outbreak models should be to understand the spread and support strained frontline personnel and health systems. Waiting for the perfect model may prevent us from deploying good models.

ML can also be applied to later outbreak stages, improving outbreak recognition and patient diagnostics, as well as community engagement (NLP for denial, stigma tracing), drug and vaccine discovery, and reducing repetitive tasks and workload. To effectively apply ML, we need platforms for rapid response, collaborative multidisciplinary teams, the ability to share data (while maintaining privacy), improved capacity to work from smaller datasets, and investment in good models instead of waiting for the perfect one.

4. Summary

The most requested topic was discussing the impact of ML models in the clinic and what researchers could do to better meet the needs of healthcare workers showing the interest of the ML4H community to bridge the current translation gap of ML for health. One in-person and one virtual roundtable addressed this topic. The discussants concluded that it is essential to choose the right problem to work on by carefully assessing whether ML can have a positive impact.

Apart from disease prediction and risk stratification which are very common tasks addressed by ML research, there are plenty of potentially impactful applications of ML. The roundtables also mentioned that economic considerations are relevant to ensure successful translation into hospitals that operate under financial pressure. An overarching theme of the roundtables was to involve clinicians early in a project to ensure useful deployment of ML.

The different steps towards successful translation were discussed in various roundtables. Starting with datasets, one discussion focused on incentives for creating new data collections. At the local level, they highlighted the need to invest in healthcare systems to provide data of a useful quality and to proactively use consent forms for existing patients. To establish a global data-sharing ecosystem they think it is crucial to show the value of ML for health to all people to motivate the necessary investment. Existing global initiatives can serve as blueprints to develop such ecosystems. Two further roundtables discussed evaluation of healthcare datasets before applying ML. They talked about the threats of using historic data for model development and concluded that different perspectives in a multidisciplinary team can help with identifying potential problems. Particularly, they highlighted the importance of scrutinizing the label definition to ensure that it is aligned with ethical considerations and the aim of the prediction. In addition, techniques and methods to detect and tackle potential dataset issues such as adversarial attacks and iterative model development were discussed.

Another potential problem of ML for health is missing generalizability across hospital systems. One roundtable identified the key problem that many studies only optimize on a single dataset which often leads to performance degradation when moving to a new environment. Instead, the participants asked for a culture of model auditing, identifying potential failure modes, and data sharing allowing external validation. They suggest two potential ways forward: developing supermodels trained on many datasets and using uncertainty quantification to provide reliable estimates of when to trust a model. There was also a discussion regarding the final step of translation: model approval. The FDA has already established risk-adjusted approval and updating processes that usually require supporting studies and a one-year monitoring. However, the necessary data collection can be costly and only few tools to support these processes exist so far. The roundtable suggested that strategic partnerships between academia and industry can be useful for successful translation and approval of modern ML methods.

Three roundtables were centered around creating better ML models. The topics ranged from injecting domain knowledge, integrating multiple data sources, and using pre-trained foundation models. Different possibilities to inject domain knowledge varying in cost were discussed ranging from using synthetic data to experts-in-the-loop where a human being controls the prediction results. The roundtable also formulated the open question of iterative inclusion of domain knowledge to amortize expensive model re-training. Integrating multiple modalities was considered very relevant to medicine which is multi-modal. Deep learning was highlighted for its flexibility to integrate modalities and different methods to do so were discussed. However, discussants warned about the added complexity, for instance, hampering the detection of a potential bias. The discussion about foundation models for healthcare also considered the question of multimodality, e.g. for images and waveforms. The roundtable also suggested several applications for foundation models for text such as retrospective interpretation of clinical notes and assisted note writing. It was suggested that foundation models would be particularly useful in areas where deep learning already proved superior.

Lastly, two discussions revolved around specific applications of ML for population health and infectious diseases picking up on many topics from previous roundtables. The discussion about population health focused on data quality and dimensionality. These are specifically relevant for population health when dealing with a low-resource setting and heterogeneous populations. The roundtable on infectious diseases discussed different stages of disease outbreaks and potential applications of ML. In particular, they stated that in the early outbreak stages, simplified and imperfect models might be acceptable when they provide some value to frontline workers. For effective long-term applications of ML, they asked for a rapid data-sharing ecosystem, multidisciplinary research teams, and investment in better model development.

5. Lessons Learned

We identified potential topics from the literature, organizers, and speakers. Settling on topic candidates was hard because many suggestions were relatively broad and there were many duplicates. We consider this step crucial because it determines what is included in the poll and what chairs have to prepare for. Using a poll to get feedback from the ML4H community was a simple and useful way to assess the interest in the topics. However, getting enough votes is important for reliable estimates. We obtained 39 results in a period of two months using the official ML4H Twitter account.

Recruiting senior chairs was the most time-consuming task. Especially, for the in-person roundtables we were limited to attendees coming to the venue while trying to achieve fair representation across geographical locations, gender, and levels of expertise. Hence, we started recruiting senior chairs after the registration had been open for some time so that people could decide whether they would attend in person. Also, we were a bit flexible as to which topics to include and how many chairs we recruited. This turned out to be a useful strategy. Recruiting junior chairs was easier since the requirements were lower.

To encourage the use of similar formats across all roundtables, we prepared clear instructions for the chairs to have a consistent experience across roundtables and to make the outcomes usable for the community. Asking chairs for an introduction paragraph and discussion questions in advance encouraged a certain level of preparation.

We held in-person and virtual roundtables concurrently to give virtual attendees the impression of being part of the regular symposium schedule. The number of participants varied a lot between topics, which is especially relevant for the in-person roundtables. All chairs prepared a summary paragraph that we used to compile this document. It could prove useful to have stricter instructions for the summaries to obtain more homogeneous texts. However, this would add an additional burden for the roundtable chairs.

Generally, roundtable sessions seem to encourage active participation as they follow a panel discussion format but with a smaller size and less formality. It also benefited participants by allowing them to network with domain-specific peers. Thus, we encourage similar conferences to employ such roundtable sessions for enhanced participation and networking opportunities.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.130.

Yuri Y M Aung, David C S Wong, and Daniel S W Ting. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *British Medical Bulletin*, 139(1):4–15, 08 2021. ISSN 0007-1420. doi: 10.1093/bmb/ldab016. URL https://doi.org/10.1093/bmb/ldab016.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20 (e2):e253–e259, 2013.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *nature*, 409 (6822):860–921, 2001.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.522.

URL https://aclanthology.org/2021.emnlp-main.522.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3): 283, 2021.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.

Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush Varshney. Out-of-distribution detection in dermatology using input perturbation and subset scanning. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2022.

David Kindig and Greg Stoddart. What is population health? *American Journal of Public Health*, 93:380–383, 3 2003. ISSN 0090-0036. doi: 10.2105/AJPH.93.3.380. URL https://doi.org/10.2105/AJPH.93.3.380. doi: 10.2105/AJPH.93.3.380.

Tsung-Ting Kuo, Xiaoqian Jiang, Haixu Tang, XiaoFeng Wang, Tyler Bath, Diyue Bu, Lei Wang, Arif Harmanci, Shaojie Zhang, Degui Zhi, et al. idash secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on gwas, and dna segment searching, 2020.

Christopher Lance, Malte D. Luecken, Daniel B. Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, Qiao Liu, Nikolay Russkikh, Gleb Ryazantsev, Uwe Ohler, NeurIPS 2021 Multimodal data integration competition participants, Angela Oliveira Pisco, Jonathan Bloom, Smita Krishnaswamy, and Fabian J. Theis. Multimodal single cell data integration challenge: Results and lessons learned. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 162–176. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/lance22a.html.

Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

Haiyi Mao, Hongfu Liu, Jason Xiaotian Dou, and Panayiotis V. Benos. Towards cross-modal causal structure and representation learning. In Antonio Parziale, Monica

Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 120–140. PMLR, 28 Nov 2022.

Matthew BA 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, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Michael P Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G Baxter, Erwin LA Blezer, Thomas Brochier, Aihua Chen, et al. An open resource for non-human primate imaging. *Neuron*, 100(1):61–74, 2018.

Jason Denzil Morgenstern, Emmalin Buajitti, Meghan O'Neill, Thomas Piggott, Vivek Goel, Daniel Fridman, Kathy Kornas, and Laura C Rosella. Predicting population health with machine learning: a scoping review. *BMJ Open*, 10:e037860, 10 2020. doi: 10.1136/bmjopen-2020-037860. URL http://bmjopen.bmj.com/content/10/10/e037860.abstract.

Sendhil Mullainathan and Ziad Obermeyer. On the Inequity of Predicting A While Hoping for B. *AEA Papers and Proceedings*, 111:37–42, May 2021. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20211078. URL https://pubs.aeaweb.org/doi/10.1257/pandp.20211078.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. CHIL '20, page 151–159, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

William Ogallo, Girmaw Abebe Tadesse, Skyler Speakman, and Aisha Walcott-Bryant. Detection of anomalous patterns associated with the impact of medications on 30-day hospital readmission rates in diabetes care. *AMIA Summits on Translational Science Proceedings*, 2021:495, 2021.

Antonio Parziale, Monica Agrawal, Shengpu Tang, Kristen Severson, Luis Oala, Adarsh Subbaswamy, Sayantan Kumar, Elora Schoerverth, Stefan Hegselmann, Helen Zhou, Ghada Zamzmi, Purity Mugambi, Elena Sizikova, Girmaw Abebe Tadesse, Yuyin Zhou, Taylor Killian, Haoran Zhang, Fahad Kamran, Andrea Hobby, Mars Huang, Ahmed Alaa, Harvineet Singh, Irene Y. Chen, and Shalmali Joshi. Machine learning for health (ml 4h) 2022. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 1–11. PMLR, 28

Nov 2022. URL https://proceedings.mlr.press/v193/parziale22a.html.

Nicola Pezzotti, Julian Thijssen, Alexander Mordvintsev, Thomas Höllt, Baldur Van Lew, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova. Gpgpu linear complexity t-sne optimization. *IEEE transactions on visualization and computer graphics*, 26(1):1172–1181, 2019.

Deborah Plana, Dennis L Shung, Alyssa A Grimshaw, Anurag Saraf, Joseph J Y Sung, and Benjamin H Kann. Randomized clinical trials of machine learning interventions in health care: A systematic review. *JAMA Netw Open*, 5(9):e2233946, September 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

Subhrajit Roy, Stephen Pfohl, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer. Machine learning for health (ml4h) 2021. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 1–12. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/roy21a.html.

Martin G Seneviratne, Nigam H Shah, and Larry Chu. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6(2), 2020.

Skyler Speakman, Girmaw Abebe Tadesse, Celia Cintas, William Ogallo, Tanya Akumu, and Adebayo Oshingbesan. Detecting systematic deviations in data and models. *Computer*, 56(2):82–92, 2023.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

Ziwei Wang, Jiajun Liu, Reza Arablouei, Greg Bishop-Hurley, Melissa Matthews, and Paulo Borges. Multi-modal sensing for behaviour recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 900–902, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/

forum?id=yzkSU5zdwD. Survey Certification.

Jack Wilkinson, Kellyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, Andrew Beam, Stefan Konigorski, Christoph Lippert, Mark S Gilthorpe, and Peter W G Tennant. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health*, 2 (12):e677–e680, December 2020.

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

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, and Shalmali Joshi. ”why did the model fail?”: Attributing model performance changes to distribution shifts. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. URL https://openreview.net/forum?id=2RbyKK-l9x.