PROVIDING GUIDELINES FOR THE RESPONSIBLE USE OF AI IN HEALTHCARE

COALITION FOR HEALTH AI VIRTUAL WORKGROUP SESSION: TRANSPARENCY

AUGUST 17, 2022, 2-3:30PM ET

SUMMARY

This Virtual Workgroup Session was convened by the Coalition for Health AI to develop a collective understanding of the definitions, important considerations, and open questions for the concepts of transparency in the development and use of artificial intelligence/machine learning applications for healthcare. With input and participation from a group of subject matter experts from healthcare and other industries, this session included a series of three lightning talk presentations that explored examples of efforts to ensure transparency in health AI, followed by brief group discussions. It also featured a set of breakout sessions that addressed the theme of transparency in the context of the selected use cases. The aim of this and other planned meetings is to develop a practical guide for implementing AI and ML tools in healthcare, one that establishes clear and appropriate guidelines and guardrails for the fair, ethical, and effective application of machine learning in healthcare settings.

OBJECTIVE

The objective for this Health AI Virtual Workgroup Session was to develop our collective understanding of definitions, important considerations, and open questions for the concepts of transparency in health AI.

LIGHTNING TALKS & USE CASES

To articulate key themes and ground discussion in real-world issues affecting healthcare and healthcare delivery, invited experts selected use cases from published reports that examined the development and deployment of algorithmic analytical tools in healthcare and other settings, and examined them in a series of brief lightning talks that were followed by focused discussions. The three use cases, which are being used throughout this series of talks, were selected to inform these discussions with real-world examples. They include:

1. Hospitals, providers, and insurance companies implementing patient-level prediction of all-cause 30-day hospital readmission using claims data or electronic health record (EHR) data[1];
2. A large health system implementing 12-month mortality estimates to support advanced care planning[2]; and
3. A machine learning algorithm being developed to triage, diagnose, and/or monitor for skin cancer using clinical or dermoscopic images of skin disease.[3]

LIGHTNING TALK 1: Z-INSPECTION: A PROCESS TO ASSESS TRUSTWORTHY AI

Presented by Jesmin Jahan Tithi, Research Scientist, Parallel Computing Labs, Intel

Z-Inspection is a practical, participatory assessment process for trustworthy AI that can be applied to multiple domains, including healthcare, the public sector, business, and many others. Z-Inspection is customizable based on use case, domains, and context, and enables the use of existing framework checklists and tools as plugins.

Before assessing trustworthy AI, it must first be defined. Z-Inspection uses the definition given by the European Commission's Independent High-Level Expert Groups on Artificial Intelligence in their Ethics Guidelines for Trustworthy AI. According to these guidelines, trustworthy AI must be:

- Lawful— respecting all applicable laws and regulations;
- Robust— both from technical and social perspective; and
- Ethical— respecting all ethical principles and values.

These guidelines also established seven concrete requirements for trustworthy AI:

- Human agency and oversight–respecting fundamental human rights; humans are in control.
- Technical robustness and safety–resilience to attack, security, a fallback plan, and general safety, accuracy, reliability, and reproducibility.
- Privacy and data governance–respect for privacy, quality, data integrity, and access to data.
- Transparency–includes interoperability, traceability, explainability, and communication about the AI.
- Diversity, nondiscrimination, and fairness–avoidance of unfair bias; accessibility and universal design; and stakeholder participation.
- Societal and environmental well-being–sustainability, environmental friendliness, and consideration of overall impact on society and democracy as a whole.
- Accountability–auditability, minimization of harm, and reporting of negative impact; communication of design tradeoffs; and opportunities for redress.

To evaluate these requirements continuously throughout the entire AI lifecycle and ensure that that the criteria for trustworthy AI are being met, we applied Z-inspection, which has three different phases:

Set-up Phase

Set-up phase consists of validation of preconditions to be verified before assessment begins. Preconditions include questions such as “Who requested the inspection?” and “Why carry out an inspection?” as well as assessing conflict of interests. We establish an interdisciplinary team of experts who work with key stakeholders for the use case to define boundaries and context for the assessment.

Assessment Phase

Next is the *assessment* phase, an iterative process that includes analyzing social and technical scenarios that describe the aim of the AI system, the actors (their expectations and how they interact with the system), the technology used, and the overall process. We then identify ethical issues and tensions in social and technical scenarios and map them to the EU trustworthy AI principles described above (similar requirements can also be applied, such as the UNESCO guidelines for trusted AI). Mapping is followed by validation, in which evidence is sought to validate claims. The process is repeated until consensus is attained.

Resolve Phase

In the *resolve* phase, we address ethical tensions identified during the assessment process where possible tradeoff solutions can be proposed as possible risks and remedies are identified, and recommendations are made to key stakeholders. Throughout the process, document everything that we do in a portable document for later sharing.

To date, four practical healthcare AI systems have been assessed using Z-Inspection:

1. An AI product for predicting cardiovascular risk[4]

This product was already deployed and being marketed as a medical device in Europe and other parts of the world. We found that assessing an already deployed product under NDA or IP constraints is often difficult, because access is limited and sharing of the assessment results for the use case may be constrained.

2. A machine learning tool deployed as a separate supportive component designed to recognize cardiac arrest in emergency calls[5]

In this instance, we found if development does not include all stakeholders, especially the product's end user, the impact in a real-world setting may not match the goals set during experimental settings.

3. A deployed deep-learning-based tool for predicting multi-regional scores that convey the degree of lung compromise in COVID patients[6]

For this use case, we realized that *urgency* during the design of an AI algorithm (e.g., during the COVID pandemic) may lead to important ethical oversight being waived.

4. A co-design phase of a deep learning tool used to classify skin lesions.[7]

In this example, we learned that stakeholders may have different goals (such as reducing overdiagnosis versus reducing mortality rates or having an explainable model). Also, incorporating different viewpoints from domain experts may impact the overall design of the system.

More information about the Z-Inspection process can be found at http://z-inspetion.org and at a preprint article available on arXiv.[8]

Key Discussion Points

- We cannot separate issues of bias and fairness from the people who create, validate, and deploy algorithms. This raises the question of how representative data including "citizen-science" input can be incorporated into the development and validation process. One way to approach this is to ensure that the multidisciplinary teams who evaluate these applications are selected to be representative of all stakeholders, including patients. Specific domain expertise, including legal perspectives and scientific and technological expertise, may also be needed, and these experts must be free of relevant conflicts of interest.
- There may be differing perspectives among domain experts about the ultimate goals of the application. Gathering consensus is therefore an important part of the process. Further, clear communication about and thorough documentation of these disagreements (and any resulting compromises or tradeoffs for a particular use case) are critical to ensuring transparency and clarity for future end users.
- True transparency about how these decisions and tradeoffs are made may require more public engagement and ethical deliberation.

- Consideration of the entire context, including technical scenarios, in which an algorithm is going to be deployed and used is important. A methodology or framework that extends from the definition through development and validation, production, operation, and model retirement. This can be applied throughout the AI's entire lifecycle.

LIGHTNING TALK 2: HEALTHSHEETS – DEVELOPMENT OF A TRANSPARENCY ARTIFACT FOR HEALTH DATASETS

Presented by Negar Rostamzadeh, Google Research

Machine learning approaches are being widely applied across technical fields. Specifically in the domain of healthcare, there is concern about the use of AI in high-stakes patient care scenarios. Existing frameworks governing the use of health data are limited in their applicability to the general use of health datasets for machine learning. The U.S. Health Insurance Portability and Accountability Act (HIPAA) does not mandate ethics review for collection and downstream use of deidentified data, nor does it limit the reuse of de-identified data. The European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) focus on notification, consent, and rights relating to the deletion of consumer data, but do not fully address concerns with ethical collection, documentation, and use of data.

On the other hand, the machine learning (ML) fairness community has identified many issues with ML systems that originate with the training process for these systems. Data documentation can help in surfacing these issues, as exemplified in pioneering work in data transparency through datasheets for datasets9 that was inspired by documentation practices in electrical engineering.

Our first goal was to contextualize datasheets for health applications. The ML community typically likes the idea of generalization – of building a model that works for all tasks and applications. However, there are drawbacks associated with a generalized approach that does not consider context, particularly when assessing systems in high-stakes scenarios.

To contextualize the datasheet, we convened an interdisciplinary team of researchers with expertise in healthcare, ML fairness, applied ethics, and human-centered design. This team co-defined transparency and accountability in healthcare data documentation, with a focus on demographic information, data versioning, accessibility, modalities, and labeling subjectivity, yielding an early version of the HealthSheet called "primary healthsheet." We then interviewed experts and conducted case studies on three publicly available datasets: Medical Information Mart for Intensive Care (MIMIC-III),[10] the Multiple Sclerosis Outcome Assessments Consortium (MSOAC) database,[11] and Floodlight Open, noting processes, limitations, challenges, and gaps in documentation.

Interview participants were selected to represent a wide range of expertise, including areas such as ML for healthcare, legal and regulatory, clinical, product, bioethics, equity, and privacy. These interviews identified documentation shortcomings in health datasets and assessed their impact on research advancements and equitable healthcare practices. The interviews also explored the use of HealthSheet as a dataset diagnostic tool for health status and discussed the potential incentives needed for data curators and extractors to create a HealthSheet for datasets. Multiple themes emerged from the interviews:

- Lack of centralized, comprehensive documentation as well as clear, easily available metadata makes it difficult for experts to select and use datasets.
- Subjectivity across multiple participants, particularly around labeling of data, presents additional challenges.
- Metadata describing inclusion criteria and accessibility are particularly important for understanding nuances and limitations of datasets.
- To ensure efficiency and ease of use, enough associated metadata must be available. Otherwise, datasets, particularly those created over long timespans, could be rendered difficult or impossible to use.
- Potential incentives should be explored to encourage dataset curators and extractors to create HealthSheets for datasets, a process that could potentially involve substantial time and effort.
- There is currently insufficient standardization for dataset curators, but the likelihood of future regulation in this arena should incentivize creators of datasets to think carefully about longer-term curation.

Relying on literature interviews and case studies, we then addressed issues related to the following: dataset versioning; inclusion criteria and accessibility; devices and contextualized attributes in data collection; collection and use of demographic data; labeling (and the subjectivity of that process); and challenge tests and confounding factors.

Although HealthSheet is not intended to be a comprehensive guideline, it may help spark conversations around the issue of health data documentation.

Key Discussion Points

HealthSheets were created based on existing datasheets that offered a generalized framework for data documentation in machine learning. These datasheets were investigated to determine which aspects could be adapted directly for healthcare applications, and what categories needed to be added based on aspects that were specific to healthcare applications.

LIGHTNING TALK 3: AI MODEL CARDS

Presenter: Alissa Graff, Operations Research Analyst, U.S. Internal Revenue Service

The AI model card initiative was completed as part of the Responsible AI focus area within the federal government's AI Community of Practice, which was convened through the General Services Administration (GSA) Office of Technology Transformation Services. The program began with a 6-week sprint that involved research discussions and preparing mockups intended to help government agencies seeking to explore or implement the concept in their own work. This effort represents only the beginning of a minimally viable process for which further iteration and improvement are expected.

Model cards are relatively new, although the idea itself is several years old. Model cards are lightweight documentation for machine learning models that present essential information needed to make an informed decision about the use of a particular model. They are often compared to Nutrition Facts Labels or Material Safety Data Sheets, which do not require specialized knowledge but allow a broad range of users to make an informed decision about whether to use certain products, as well as any precautions that should be taken. At their core, model cards are aimed at improving transparency as part of the model-building process.

Model cards can help an organization reach a broader audience of potential users, especially if they lack the time or expertise to engage with traditional documentation as part of the decision-making process. They are also useful for facilitating conversations among those developing or using the models, or to help understand or compare vendor offerings. Model cards may also be useful to decision makers for defining the intended use of models and any ethical implications.

Despite variations across different models, there are several common components. First are details about the model itself, including who is responsible for developing the model and information that can associate a model card with a specific model version. Across all model card versions, there will also be the intended uses of models. This is critical to emphasize and cannot be overstated: A model card's information is specific to the intended use of the model at hand. This information is essential because designers must ensure that potential model users do not over generalize the context and apply a model to an area for which it was never intended; or if they do, they should perform additional testing to ensure the model is performing as intended. Additional similarities across model cards include metrics, key information about the data, and ethical considerations.

It is important to differentiate model cards from other forms of documentation and doing so requires answering the question of what values the model cards must provide beyond detailed "traditional" documentation. For example, IBM's AI FactSheets[12] share some similarities with model cards; however, they include interactions between models and other aspects of software engineering. Model cards, on the other hand, are specific to a given model. This means that for an application that incorporates multiple models, a model card must be created for each one.

An extensive literature describes summary documentation for datasets to discuss factors related to quality and use of data for AI-related tasks. This provides a good supplement for model cards, but they are specific to the datasets that they cover and thus have a narrower focus, while algorithmic impact assessments cover the outcomes associated with the application holistically within the given use. Model card documentation does not replace high-quality traditional documentation.

There are several points to consider when thinking about when and how to implement model cards. First, model cards should be created for any model that may have a direct effect on people. Second, they should be created in the interval between a model's finalization and its deployment, at which time the target audience should be determined. A single card that suits both technical and nontechnical audiences may be optimal.

Ideally, a model card should be no more than two pages; however, there are several elements it should contain, as reflected in the list below:

- A "top matter" section that includes information about the individual or group that created and maintains the model, as well as information leaking the model cards, with specific model versions.
- An overview section that provides the text description of the model, how it fits into the overall application, the model's intuitive uses, and user incentives.
- An ethical considerations section that describes possible risks associated with

using the model and any strategies that can be used mitigate that risk.

- A data section that addresses data the model is trained and evaluated on, emphasizing details that could affect the generalizability of the model to other use cases.
- A model architecture section that emphasizes inputs and outputs.
- A quantitative analysis section that provides performance metrics for the model and support for understanding the quality of the predictions in the given use case.
- An algorithmic fairness section that includes subgroup analysis of model performance or downstream outcome metrics.
- A caveats & recommendations section that comprises anything not covered elsewhere that stakeholders should know when deciding on the model's use.
- A related documentation section that includes links to any other documentation for elements of the application (e.g., model cards).
- A references section with links to papers or reports that may have influenced the design of the current model.

Key Discussion Points

- Although these model cards were not specifically designed for healthcare applications, they are intended to be useful for those who are building the model to think through the different components that are important for someone who will use the model. For example, a data scientist who is not an expert in a healthcare field should be able to explain and document the potential challenges, uses, limitations, or caveats that should be known to inform decision-making.
- The GSA Technology Transformation Services convened a group across various agencies to develop examples of model cards and research recommendations that can be implemented at various agencies. For example, at the IRS there is a group working on applications for model cards in the analytic process and research.
- More broadly, it may be helpful to consider how resources might be implemented in a variety of use cases and agencies to affect transparency, trustworthiness, and fairness in a variety of content and contexts.

BREAKOUT SESSIONS

Following the conclusion of the lightning talks, conference attendees were divided into groups to participate in breakout sessions that addressed topics related to transparency in healthcare AI applications. Each breakout session included a series of key topical questions intended to focus the resulting discussions.

Transparency & Relevant Definitions

There are several important dimensions when discussing transparency in AI, including differences in the concepts of interpretability and explainability:

- **Interpretability**
  - Ability to understand/comprehend system's inner workings
  - Allows any user or stakeholder to assess credibility
  - Open to public scrutiny
- **Explainability**
  - Ability to provide explanations of AI models and decisions to others
  - Ability to answer questions such as why did a model output a particular

decision? Does the output make sense? Is it reproducible?

Key Questions

- Why is transparency important for trustworthiness?
- Why is transparency important for quality?
- What are the drivers of trustworthiness?
- What are the tradeoffs of transparency (intellectual property, privacy, accuracy)?

Discussion Points

- How can we translate concepts of interpretability and accountability into language that clinicians can understand and act upon? Incorporating output from ML algorithms into clinical decision-making requires filling a gap in terms of how clinicians are trained to think and evaluate information and apply it to clinical scenarios.
- Transparency may also include aspects such as traceability of data and communication with users who are impacted by an algorithm. Explanations should be tailored to the audience in whom the algorithm is being applied. If an explanation is not clear or credible, it might be ignored, and the system might not be used.
- There is a tension between understanding the elements that go into the model (the information and data used) vs understanding how the model works (the inner techniques and decision trees). It might be easier to explain what information is being used rather than how a decision has been reached.
- Physicians are trained to accept findings produced by randomized controlled findings while not necessarily understanding the rationales behind randomization. A robust external validation that establishes correlations should be the standard for evaluating these ML models. Over-reliance on transparency can give a false sense of understanding and control. If we are not comfortable using correlations, then transparency is irrelevant. The type of modeling is more important than a layer of "seeming" transparency.
- Relatability may be an important consideration for communicating with patients. If an end user does not understand how a machine learning application is impacting their care, they may not trust it or want to use it.
- Unless there is a validation/verification component, voluntary reporting in healthcare is subject to toxic incentives and unconscious bias, and it is hard to be completely confident on voluntarily reported data.
- There is a need to consider transparency in terms of outcome measurements and establishing a clear understanding of whether that improves or worsens the health outcome(s) that are important to the patient. Explainability may be less important for building trust than adaptability to feedback.
- It is difficult to know how to measure the impact of different communication techniques on different user groups. What is the most effective way to communicate decisions and increase trust/transparency? In the absence of established best practices, what specific, generalizable approaches for user testing could we try? And how could we test them?
- Detailed and careful process documentation is a potentially important component of building trust.

Use Case: 30-Day Hospital Readmission

Key Questions

- Where might it be hard to provide the level of transparency in that specific use case?
- What information needs to be shared to promote transparency to the end users - clinical users?
- What information needs to be shared to promote transparency to the patients?
- What information needs to be shared to promote transparency to the regulators and governing bodies?
- Should there be transparency when there is uncertainty of data and accuracy in different populations?

Discussion Points

- The definition of transparency might be different for a model creator versus someone who is a user of this data.
- Transparency, from the patient's perspective, might include questions such as: what information is being collected, where is that information being used, what assumptions are we making, what tools are being used? What tools are being used on *me*? Who is liable for their use? These questions are difficult to "future-proof," because of the changing array of tools that might be applied over time.
- In the case of a predictive algorithm for 30-day readmission: although this might seem relatively "safe," there is no requirement for FDA review or approval, no third-party evaluation.
- What constitutes transparency is subjective. However, transparency is not a goal in itself – it is a means to achieve accountability.
- When talking about transparency, it's critical to identify who the stakeholder is. This includes not just the users themselves, but all those upstream and downstream of the application's use.
- We need to meet stakeholders where they are and understand differences among patients vs providers vs data scientists. How is a model translated into human-readable format for someone who is not a data scientist? What data were used, and are there biases present? Further, the outcome of interest itself may reflect a bias (for example, outcomes that are primarily of interest to hospital administration as opposed to patients). These kinds of information should be available to all stakeholders.
- The incentive structures around transparency may pose challenges. There is broad agreement that transparency is desirable, but how can we get the right information to the right constituencies?
- Approaching incentives for transparency from the perspective of providing guidelines for development and deployment might be useful. If transparency is not required as part of the process, the product should not be used.
- Different levels of information are needed for different users - what do patients need to know, versus other stakeholders?
- General education and awareness for the public around these issues ideally should take place before these technologies are implemented.
- The underlying data raises issues as well – where it comes from, how it's collected, how it's used. Is it from the electronic health record (EHR) or the insurer? The total volume is massive, but it may or may not be accurate, and that affects inferences drawn from that data.

It's important to recognize and communicate these limitations.

- As we move into an era of data collection from personal devices without the presence of a physician, questions about the quality of data being collected and what those data are being compared to are key issues to address.
- Having some understanding of what constitutes an "evil use case.", and how patients or end users can be protected are important considerations. "Evil use cases" could result in biased or disparate financial outcomes, bad actors or adversaries working to disrupt operations and/or cause poor outcomes, or using data or algorithms to impose punitive actions on a historically disadvantaged or discriminated group.
- There are orders of magnitude more consumer-generated data than EHR data. The amounts of data being transacted are enormous and growing. We need continuous transparency.

Use Case: Diagnostic Imaging for Skin Cancer

Key Questions

- Where might it be hard to provide the level of transparency in that specific use case?
- What information needs to be shared to promote transparency to the end users - clinical users?
- What information needs to be shared to promote transparency to the patients?
- What information needs to be shared to promote transparency to the regulators and governing bodies?
- Should there be transparency when there is uncertainty of data and accuracy in different populations?

Discussion Points

- It is often impossible to know how an AI algorithm arrives at decision , even when those decisions are matters of life and death. Human wisdom and nuance are key. There are trade-offs between interpretability and accuracy.
- There is a spectrum regarding transparency or lack thereof in terms of data quality. Data from EHRs or wearable devices may be on the messier end of the spectrum, as many decisions have been made as part of the process. However, with diagnostic imaging, there might be a relatively short distance between ground truth and model, making it potentially easier to describe.
- A key question about transparency concerns the intention of the tool. Was it to decrease costs? To validate a finding? The intention must be clear from the outset. The aspect of transparency likely to cause distrust is the perception of hiding information or refusing to share it. Every AI tool may need a "report card" that explains the levels of evidence upon which the tool is based.
- Regarding the use case of diagnostic imaging for skin cancer, we might want to know how many skin tones were used and what types of cancer were looked at. What is the false negative rate? What is the threshold of a false negative? How does the algorithm manage the level of uncertainty? Different cancers might induce different thresholds of false alarms.
- There is a tradeoff between validation by end users and scalability. It is unrealistic and perhaps unfair to assume that end users like clinicians will examine and validate every output (or have the technical expertise to do so). A clinical end user should be assured by some responsible authority in their organization that the responsible

authority has validated the tool. Such information may be important but may be most useful for regulators and governing bodies, with clinical end users adding a degree of fine-tuning.

- When a model is offered as an asset, transparency should include an indication of what “generation” the model belongs to – in other words, an index of the performance and development history of the model. Audit trails help us see the iterations and changes to the models and who it is competing against.
- Regarding the use case of skin cancer, we need models that go beyond “yes” or “no” to be able to express levels of confidence or uncertainty. An algorithm that assesses skin cancer in diagnostic imaging should provide an estimate of error if the image quality or lack of relevant training data make assessment difficult.
- The use of synthetic data – and the potential for bias and impacts on fairness – is an area that may deserve further discussion and consideration.

REFERENCES

1. Wang HE, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, Saria S. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc. 2022 Jul 12;29(8):1323-1333. doi: 10.1093/jamia/ocac065. Erratum in: J Am Med Inform Assoc. 2022 Jun 17;: PMID: 35579328; PMCID: PMC9277650.
2. Ron C. Li, MD, Margaret Smith, MBA, Jonathan Lu, MS, Anand Avati, MS, Samantha Wang, MD, MHS, Winifred G. Teuteberg, MD, Kenny Shum, PhD, Grace Hong, Briththa Seevaratnam, MS, Jerri Westphal, MSN, RN, CNML, et al.NEJM Catalyst Innovations in Care Delivery 2022; DOI:https://doi.org/10.1056/CAT.21.0457
3. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol. 2021 Nov 1;157(11):1362-1369. doi: 10.1001/jamadermatol.2021.3129. PMID: 34550305.
4. Zicari RV, et al. Z-Inspection®: A Process to Assess Trustworthy AI. IEEE Transactions on Technology and Society, 2021;2(2):83-97. doi: 10.1109/TTS.2021.3066209.
5. Zicari RV, Brusseau J, Blomberg SN, et al. On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Front Hum Dyn. 2021. https://doi.org/10.3389/fhumd.2021.673104
6. Allahabadi H, Amann J, Balot I, et al. Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients. IEEE Transactions on Technology and Society. 2022. doi: 10.1109/TTS.2022.3195114.
7. Zicari RV, Ahmed S, Amann J, et al. Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. Front Hum Dyn. 2021. https://doi.org/10.3389/fhumd.2021.688152
8. Zicari RV, Amann J, Bruneault F, et al. How to assess trustworthy AI in practice. Preprint available from arXiv. June 26, 2022. https://doi.org/10.48550/arXiv.2206.09887
9. Gebru T, Morgenstern J, Vecchione B, Wortman Vaughan J, Wallach H, Daumé H III, Crawford K. Datasheets for Datasets. Communications of the ACM. 2021; 64(12):86-92. DOI: 10.1145/3458723
10. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127; PMCID: PMC4878278.
11. LaRocca NG, Hudson LD, Rudick R, Amtmann D, Balcer L, Benedict R, Bermel R, Chang I, Chiaravalloti ND, Chin P, Cohen JA, Cutter GR, Davis MD, DeLuca J, Feys P, Francis G, Goldman MD, Hartley E, Kapoor R, Lublin F, Lundstrom G, Matthews PM, Mayo N, Meibach R, Miller DM, Motl RW, Mowry EM, Naismith R, Neville J, Panagoulias J, Panzara M, Phillips G, Robbins A, Sidovar MF, Smith KE, Sperling B, Uitdehaag BM, Weaver J; Multiple Sclerosis Outcome Assessments Consortium (MSOAC). The MSOAC approach to developing performance outcomes to measure and monitor multiple sclerosis

disability. Mult Scler. 2018 Oct;24(11):1469-1484. doi: 10.1177/1352458517723718. Epub 2017 Aug 11. PMID: 28799444; PMCID: PMC6174619.

12. Arnold M, Bellamy RKE, Hind M, et al. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. Preprint available from arXiv. https://doi.org/10.48550/arXiv.1808.07261

SELECTED READING & RESOURCES

- Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Shah NH. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. medRxiv 2021.07.21.21260282; doi: https://doi.org/10.1101/2021.07.21.21260282
- Ratti E, Graves M. Explainable machine learning practices: opening another black box for reliable medical AI. Ai and Ethics. 2022. https://link.springer.com/article/10.1007/s43681-022-00141-z
- Sendak M, Sirdeshmukh G, Ochoa T, et al. Development and Validation of ML-DQA -- a Machine Learning Data Quality Assurance Framework for Healthcare. Preprint available from arXiv. https://doi.org/10.48550/arXiv.2208.02670
- Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020 Mar 23;3:41. doi: 10.1038/s41746-020-0253-3. PMID: 32219182; PMCID: PMC7090057.
- Sendak M, Elish MC, Gao M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. January 2020 Pages 99–109. https://doi.org/10.1145/3351095.3372827
- Malik MM. A Hierarchy of Limitations in Machine Learning. Preprint available from arXiv. https://doi.org/10.48550/arXiv.2002.05193
- Malik MM. Interpretability is a red herring: grappling with “prediction policy problems.” https://www.mominmalik.com/ier2019.pdf
- Karunagaran S. Making it easier to compare the tools for explainable AI. June 30, 2022. https://partnershiponai.org/making-it-easier-to-compare-the-tools-for-explainable-ai/. Accessed August 31, 2022.
- US Food and Drug Administration. Good machine learning practice for medical device development: Guiding principles. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. Accessed August 31, 2022.
- Robbins R. An experiment in end-of-life care: Tapping AI’s cold calculus to nudge the most human of conversations. STAT News. July 1, 2020. https://www.statnews.com/2020/07/01/end-of-life-artificial-intelligence/. Accessed August 31, 2022.