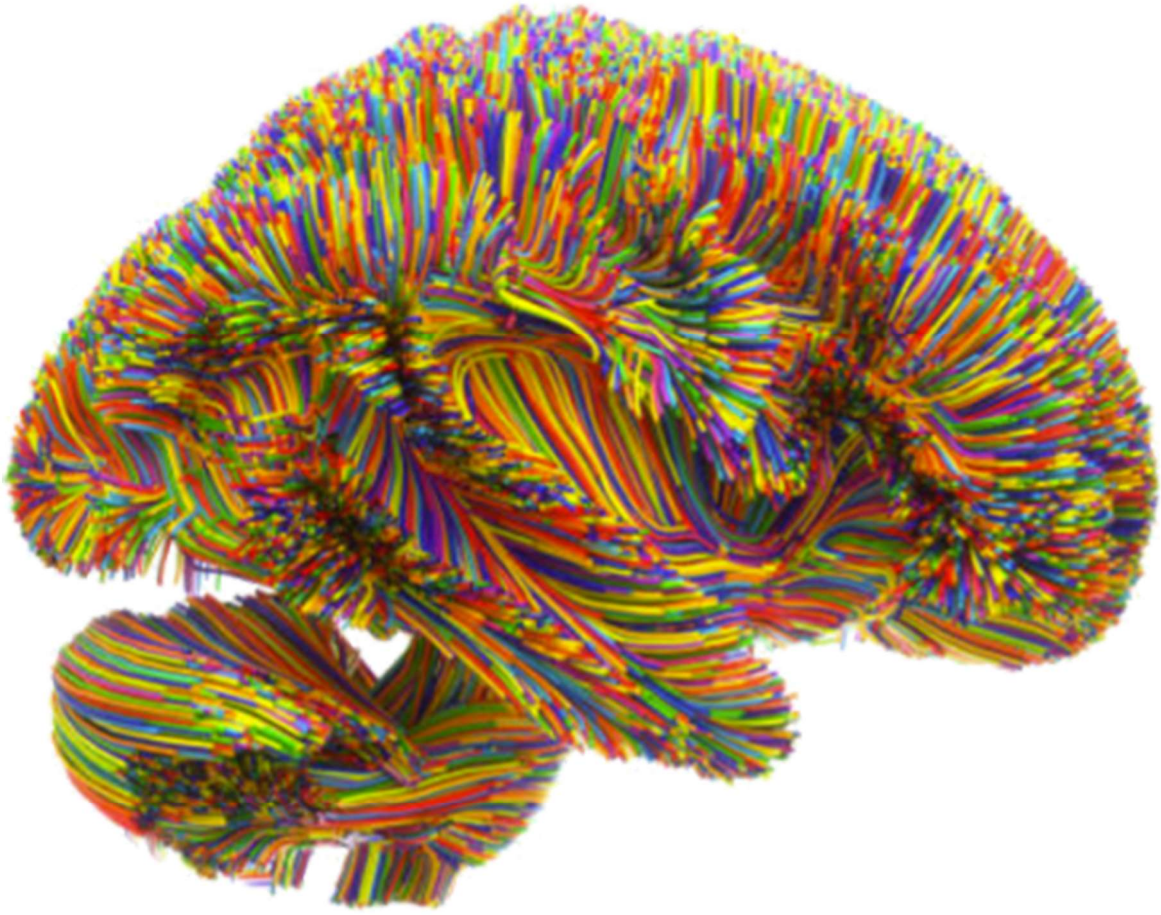

PROVIDING GUIDELINES FOR THE RESPONSIBLE USE OF AI IN HEALTHCARE

COALITION FOR HEALTH AI
VIRTUAL WORKGROUP SESSION #1: BIAS, EQUITY, AND FAIRNESS

JUNE 21, 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 equity, bias, and fairness in health. 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 and group discussions centered on preselected use cases. It also featured a set of breakout sessions that addressed the themes of Health Equity by Design; Bias and Fairness Processes and Metrics; and Impacting Marginalized Groups: Mitigation Strategies for Data, Model, and Application Bias. 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 useful application of machine learning in healthcare settings.

INPUT AND FEEDBACK

We welcome feedback and input on the ideas presented here, on additional ideas and concepts, and on the future direction of work pertaining to bias, equity, and fairness in health AI.

Input and feedback are requested via submission form on our website during a 30-day comment period, ending September 15, 2022.

INTRODUCTION

The use of artificial intelligence (AI) and machine learning (ML) applications in healthcare offers enormous potential for accelerating clinical research and for improving the quality and delivery of healthcare. However, a growing body of evidence shows that such tools can perpetuate and increase harmful bias.

Responding to concerns about bias, fairness, and equity in the use of AI applications in healthcare, the Coalition for Health AI in collaboration with Duke AI Health, Mayo Clinic, The MITRE Corporation, and Partnership on AI, and with support from The Gordon & Betty Moore Foundation, convened a group of subject matter experts to work together across healthcare and other industries. By bringing together government, academia, and industry for constructive dialogue, the group aims to develop a practical guide for implementing AI and ML tools in healthcare that establishes clear and appropriate guidelines and guardrails for the fair, ethical, and useful application of machine learning in healthcare settings.

With this overarching goal in mind, the objective of this Virtual Session was to develop our collective understanding of the definitions, important considerations, and open questions for the concepts of equity, bias, and fairness in health. This Virtual Session included a series of three lightning talk presentations and group discussions centered on preselected use cases, and a set of breakout sessions that addressed the following themes: Health Equity by Design; Bias and Fairness Processes and Metrics; and Impacting Marginalized Groups: Mitigation Strategies for Data, Model, and Application Bias.

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.

LIGHTNING TALK 1: 30-DAY HOSPITAL READMISSION

Presented by Suchi Saria, PhD (Johns Hopkins University)

Hospitals, providers, and payers have implemented ML models that predict 30-day all-cause patient readmission to hospitals. These models typically rely on data drawn from patient electronic health records (EHRs) or from insurance claim databases. For this session, discussion centered on a published report by Echo Wang and colleagues that describes an "end-to-end" bias checklist for ML models and its application to models designed to predict outcomes such as 30-day readmission.[1] Hundreds of metrics have been used to develop predictive models but end-to-end assessments that evaluate the performance of the model across all phases, from development through deployment in the clinical setting, have been lacking, as have investigations of the potential interactions between these different phases.

We specifically examined disparate performance, in which an algorithm performs well for one population or setting but not another. One way to mitigate is to adjust thresholds for action in the different populations/settings to enable equitable allocation of resources.

In healthcare, the ultimate marker of bias are the presence of disparities in outcomes and unequal allocation of resources. The checklist walks through how the model is defined, developed, and used.

Key Discussion Points

- Regardless of whether an algorithm is based on clinical heuristics or ML, users need to know two things: 1) how will this algorithm be applied, and 2) will it create an unequitable distribution of resources? More complex models are not necessarily subject to more bias and conversations about whether something is or is not AI are unproductive. The focus should be on identifying potential sources of bias in the applied setting through checklists and other types of tools.
- Accountability for bias is a shared responsibility for all—not just developers, but also the health system deploying the model and the end user interpreting and acting on its output. Checklists can help by letting users assess model deployment in local, real-world contexts and settings.

Sources of Bias in Predictive Models at Different Stages	
Model definition & design	
Label bias	Biased proxy variables used in place of ideal predictive variable during model training
Modeling bias	Model's design yields inequitable outcomes
Data collection & acquisition	
Population bias	Model performs poorly in some populations because those groups were not adequately represented in training data
Measurement bias	Bias introduced by differences in quality or ways that features are selected and calculated across groups
Validation	
Missing validation bias	Lack of validation studies to examine performance in subgroups
Deployment & use	
Human use bias	Inconsistencies introduced by human operators acting on model outputs

Adapted from: Echo 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 May 17:ocac065. doi: 10.1093/jamia/ocac065.

LIGHTNING TALK 2: 12-MONTH MORTALITY ESTIMATES FOR ADVANCED CARE PLANNING

Presented by Nigam H. Shah, MBBS, PhD (Stanford University)

The use case for this talk is drawn from a report, published in *NEJM Catalyst* in 2022,[2] of a large academic health system's experiences with implementing a model that provides 12-month mortality estimates to inform advanced care planning for patients.[3] Appropriate and effective AI-guided care rests on three factors: 1) the model and its output; 2) governing policies and the capacity to act upon the model's output; and 3) the properties of the intervention. These elements together define the degree to which AI-guided care is useful, reliable, and fair. Achieving the desired state of useful, reliable, and fair AI-guided care requires not only pushing the frontier in terms of theoretical development, and implementing processes to routinely assess these attributes, as well as the business value of data and algorithms. It also requires ensuring that the system's IT organization and infrastruture are AI-ready. There is currently a mismatch in the large number of recommendations—more than 200—on what to report for model development, and the number of recommendations—only 10—devoted to assessing fairness in predictive models.[4] Furthermore, these checklists and practices are primarily focused on an approach to fairness that seeks to ensure an absence of systematic differences between groups of people, but that in itself is not sufficient to create equity. Equally important is the *absence of systematic differences* in how the benefits of using a given model accrue to different groups or populations. Ensuring equality of resource allocation between groups does not automatically result in equity with regard to outcomes. In fact, a recent analysis[5] suggests that most attempts to fix the fairness of a given algorithm or model's outputs render it less accurate for the majority of the people to whom it is being applied.

Key Discussion Points

- How does one measure and track the accrual of benefit related to the use of a particular model? Can structural and historical discrimination be factored into that calculation? A consequentialist approach will look at how many individual persons receive benefit as the result of the model output informing human judgement, but ultimately there is no way to ensure that "equality" in the model output will yield to equity without a human in the decision-making loop.
- Other possible approaches to addressing bias and fairness at the development stage include increasing the capacity of the model or adding data representing a minority group of patients. There is also the possibility of the end user adjusting the allocation of resources; however, applying different standards for different groups or subgroups may raise several sensitive issues.

LIGHTNING TALK 3: DATA BIAS

Presented by Hong Qu (Harvard Kennedy School)

Deployment of socio-technical systems can "bend the arc of justice", and not always in desirable directions. Algorithms themselves cannot do this by themselves, and engineers are not trained to deal with structural biases.[6,7] As we develop and deploy algorithmic models, we as developers are responsible for assessing risk, but are also accountable to the people affected by the model.

Examples of different kinds of bias abound, ranging from data and computational biases to human and structural biases. The Massachusetts Institute of Technology's AI Blindspot Project was undertaken to illustrate and counter the various kinds of bias that can creep into all stages of model development and deployment with an emphasis on communicating concepts to a nontechnical audience. Another example of unintended bias affecting data analysis can be seen in Boston's crowdsourced "Street Bump" project, which featured a smartphone app that gathered accelerometer data to identify potholes and other problems with Boston road maintenance. However, because certain groups of Boston residents were less likely to have smartphones, the app disproportionately collected data from neighborhoods that tended to be younger and/or wealthier.[8]

Historical power structures and stigma are always present. Data by itself cannot counter these forces; those who work with it must be intentional in accounting for systemic and institutional biases. The best and most direct way to accomplish this is by working directly with the communities affected. It is important to acknowledge that data are rarely fully complete or wholly representative of a population. Developers must ask who is being counted, remembering that sampling methods are often affected by sampling bias.

Key Discussion Points

- Data governance, including risk assessments prior to launch and impact assessments after deployment, is essential for detecting and addressing systemic, computational, and human biases. Continuous monitoring over time is key as the contexts of data use evolve.
- Data from the past often guides current and future efforts. But if those data were gathered without respect to appropriate ethics and/or representation, what happens when those data are used going forward? Historical data does not in itself provide solutions. Other disciplines, like social scientists who are trained in the social determinants of health, may provide a critical perspective. We need to examine differential impact as opposed to discriminatory impact, and precisely measure it.

BREAKOUT SESSIONS

Following the conclusion of the lightning talks, participants were divided into breakout sessions that addressed the following topics: 1) health equity by design; 2) processes and metrics for assessing bias and fairness; and 3) impact on marginalized groups: mitigation strategies for data, model, and application bias. Each breakout session included a series of key topical questions intended to focus the discussions.

Health Equity by Design

Discussion Questions

- What does health equity by design mean (For you? For all? For marginalized and under-represented groups?) What are the important considerations? What are the open questions?
- Are there ways to contribute to addressing and increasing diversity in data, model, and applications (with respect to the use cases discussed)?

Key Points

Health Equity by Design

Health equity by design is a concept analogous to "quality by design" in quality assurance and quality control. It requires intentionality, especially when addressing potentially painful issues related to structural and historical bias. Tools for creating health equity by design include data governance, asking questions at the appropriate times, and encouraging developers to question their work. The process of ensuring health equity by design encompasses the entire pipeline lifecycle and includes the data we choose to collect, the ways we collect it, the methods we use to develop our models, and how we deploy the results into practice.

In practical terms, there is a "knowing/doing gap": in other words, there is a disconnect between promoting best practices and current real-world practice in healthcare. There is also a mismatch in the need to build certain kinds of tools to support equity and the incentives that affect developers and the companies for which they work. We need to seek ways to influence a developer community in which ethical considerations are viewed as an afterthought and not a primary design consideration.

The field is still at the "data collection" step. However, relatively few people are invested in fixing the data, but instead want to work on developing and deploying models. Nevertheless, until we collect the proper data in the proper ways—by viewing data through the lens of equity—existing biases will be perpetuated or even amplified.

Best Practices

Cybersecurity spaces are familiar with the concept of the threat model—a systematic approach to identifying potential threats to a system, thinking about what could possibly go wrong, what level of skilled advisory is needed, what precautions or mitigations are in place to prevent a bad outcome from happening. This approach would be useful for pursuing equity in predictive modeling, a space characterized by complex systems that must be protected from threats arising from bias and unethical motivations and behaviors.

Governance is critically important, as is ensuring that these principles are disseminated within an organization through documentation, pathways of accountability, and internal audit systems. AI tools are relatively new, and some organizations are still struggling to understand how to work in this space. Furthermore, we don't yet have objective measures for what counts as "good enough" when it comes to explainable AI or transparency. We also need to continue working on regulation and guidance (such as the National Institute of Standards and Technology [NIST] risk management framework for AI), and articulate ways to operationalize guidance internally and establish benchmarks and measures for AI tools that everybody in the community can understand. This may include defining end-to-end what a model looks like, including audit trails, data governance, real-world validation, open-source models, and independent expert evaluation.

Increasing Diversity in the Dataset

A salient question to ask is why a given dataset lacks diversity in the first place. Does the population trust you when you collect data? Why not? How can you connect with them in ways that foster trust? Sometimes people are reluctant or unwilling to provide their data—if people from marginalized groups are opting out, why is that?

Earned trust is key— people must be confident that researchers or health systems will not misuse or weaponize health data. In marginalized communities, trust has been compromised by a generational history of bias, suffering, and exploitation. Trust must be built in at the design stage, in part by considering how people make choices and asking permission (for example, defaulting to an "opt-in" model for sharing data). In the past, too much focus has been placed on trying to gain trust rather than on *being trustworthy*. The healthcare and data industries would benefit from some careful and candid introspection on the topic of trustworthiness.

We must systematically rethink and redesign data collection, labeling, and accessibility to overcome current shortcomings. This includes transparency about data definitions and the decisions that affect what elements are included or excluded. Cultural and patient preferences should be considered when developing outcomes. Presently, healthcare organizations do not have the legal and technical infrastructure to collect, aggregate and share data openly, especially in low-resource settings.

"Earned trust is key—people must be confident that researchers or health systems will not misuse or weaponize health data."

Linking Outcomes to Equity by Design

Designing for equity requires measuring the downstream effects of AI predictions and developing policies that move towards more equitable outcomes. We must invest in processes at the intersection of care pathways, populations, and environments that lead to the most desirable outcomes. Policies and outcomes are sensitive to

demographic and equity issues, but we have very little data about these issues to inform end-to-end systems as they are designed.

Bias and Fairness Processes and Metrics

Discussion Questions

- What are the key processes and metrics for evaluating algorithmic models for bias and fairness processes and metrics that come to mind?
- What are the most important considerations and tradeoffs when using these processes and metrics?
- What are the appropriate processes and metrics to use when quantifying marginalized groups?
- How do you select appropriate processes and metrics to fit a given use case?

Key Points

Considerations when Addressing Causes of Bias

When assessing and addressing the causes of bias, simply including representative data in a training dataset does not guarantee that bias will not be present; for example, in a mammogram dataset, black patients are typically more likely to receive more invasive biopsies than patients of other races/ethnicities, but at the same time, Black women are more likely to have more advanced cancer. These kinds of intersectionalities become important when assessing bias and hidden stratification of populations can be difficult to evaluate. We can also examine the marginal change in bias that occurs with the introduction of the algorithm. Even if the algorithm is implemented and demonstrates the extent of bias in the status quo, that itself is valuable information, contrasted with the previous state of being ignorant of the existence of bias.

Managing Tradeoffs

Addressing bias requires the user/organization to make tradeoffs thoughtfully.

> *"Ignorance of bias is our current state, so with increased measurement, we are less ignorant but more uncomfortable."*

No single metric or checklist can solve the problem of bias. We must identify the skillsets and training curricula that will allow a given team at a hospital or health system to deal thoughtfully with these issues. Organizations should discuss the tradeoffs that are inevitable in the development and deployment of predictive models and end users must be equipped to make decisions about tradeoffs with complex ethical impacts. If organizations are incapable of working through the ethical issues surrounding the use AI-assisted tools in specific settings, then the tools should not be deployed.

Making the Implicit Visible

The process of making implicit decisions and rendering them explicit is a necessary but time-consuming and labor-intensive project. Ultimately, the definitions of bias and fairness will depend on each setting. Such decisions require the engagement of all

stakeholders, including patients, in a co-design process built from the outset.

Data Expiration Dates

Underlying data contribute to the bias present in an algorithm. We need more investment in the creation and curation of datasets. Historical data should not by default be considered ground truth and datasets may need expiration dates.

Tradeoffs and Domain-Specific Fairness Criteria

Equal patient outcomes refer to the assurance that protected groups will benefit equally in terms of patient outcomes affected by the deployment of machine-learning models. Equal performance refers to the assurance that a model is equally accurate for patients in protected and nonprotected groups. Equal allocation (also known as demographic parity) ensures that the resources are proportionately allocated to patients in the protected group. Policy makers, including those in the hospital and health system deploying the algorithm, should select which criteria should be maximized.

Community Engagement and Knowledge

More effort and resources should be devoted to understanding how we can work with and empower communities to make the most of the data they own. As we do so, it is essential that we consider local context and historical nuances, including social determinants of health, structural issues, housing, environment, and more. This means engaging with and listening to community voices. Community voices can inform every aspect of the development pipeline from the problems we choose to work on to the development of outcomes labels that reflect community preferences and needs. It's equally important that we understand community engagement as a continuous process, not a one-time event, and that we develop ways to measure impact.

Impacting Marginalized Groups: Mitigation Strategies for Data, Model, and Application Bias

Discussion Questions

- Are there ways to address bias and fairness for data bias? For model bias? For application bias?
- What are the important considerations for impacting marginalized groups regarding these kinds of bias?

Key Points

Expanding Solutions for Bias

Checklists and training individuals on handling certain datasets are not sufficient to counter the problems of bias. An organization reports that comparing data to representation in the underlying populations is primarily driven by variation in racial categories, economic stability, and health literacy. There will always be at least one attribute for which the data is biased; we need to define what attributes are important to the context to assess bias and the degree of bias to report. Depending on what variables are being assessed, one can run the risk of missing assessment of important subgroups.

Everything Starts with Data

We need data about data in healthcare. This means incorporating measurement models and psychometrics, drawing upon a wealth of experience in other fields such as engineering and actuarial science. Everything starts with data. Data are often restricted and context-dependent, which becomes a critical issue that needs to be addressed.

> *"If the status quo is inaccurate or imbalanced, an AI system, even if biased, may be an improvement."*

We must also think about how the data is being used. We should ask what it is we are trying to do or fix by applying a model. A theory of power may be useful for thinking critically about datasets and issues that we should consider.

"Nothing About Us Without Us"

There are always limitations that affect a given dataset, but it is important to remain mindful of where the model will be deployed, and who will be impacted by its use. Those people who are affected by the model must be understood as key stakeholders, and individuals and populations have the right not to have their data used it they so decide.

Validation and Measurement

AI applications have the potential to uncover or even mitigate bias. However, due to worries about patient harms from the use of algorithms, adoption of algorithmic tools by healthcare providers remains low. Currently, there exist 364 separate equations designed to estimate cardiovascular risk, none of which have examined the calibration of subgroups of large datasets. No one knows what would happen if we examined these through the lens of age, gender, or race. In short, the current conceptualization of the "AI problem" is not broad enough and is missing other important issues. If the status quo is inaccurate or is imbalanced, an AI system, even if biased, may be an improvement.

REFERENCES

1. Echo 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 May 17:ocac065. doi: 10.1093/jamia/ocac065. Epub ahead of print. Erratum in: J Am Med Inform Assoc. 2022 Jun 17;: PMID: 35579328.
2. Li RC, Smith M, Lu J, Avati A, Wang S, Teuteberg WG, Shum K, Hong G, Seevaratnam B, Westphal J, et al. NEJM Catalyst Innovations in Care Delivery 2022; 3(04): doi: https://doi.org/10.1056/CAT.21.0457
3. Lu J, Sattler A, Wang, S. et al. (2022). Considerations in the Reliability and Fairness Audits of Predictive Models for Advance Care Planning. Preprint available from medRxiv. July 12, 2022. doi: https://doi.org/10.1101/2022.07.10.22275967
4. 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. Preprint available from medRxiv. July 23, 2021. doi: https://doi.org/10.1101/2021.07.21.21260282
5. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform. 2021 Jan;113:103621. doi: 10.1016/j.jbi.2020.103621. Epub 2020 Nov 18. PMID: 33220494; PMCID: PMC7871979.
6. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. National Institute of Standards and Technology (NIST). NIST Special Publication 1270. March 2022. Available at: https://doi.org/10.6028/NIST.SP.1270 Accessed June 30, 2022.
7. Benjamin R. Assessing risk, automating racism. Science. 2019 Oct 25;366(6464):421-422. doi: 10.1126/science.aaz3873. PMID: 31649182.
8. Crawford K. The Hidden Biases in Big Data. Harvard Business Review. April 1, 2013. Available at: https://hbr.org/2013/04/the-hidden-biases-in-big-data. Accessed July 1, 2022.

SELECTED READINGS

- Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn (2020). 2020 Apr;2020:151-159. doi: 10.1145/3368555.3384468. PMID: 33196064; PMCID: PMC7665161.
- Gaddy M, Scott K. Principles for Advancing Equitable Data Practice. Urban Institute. Available at: https://www.urban.org/research/publication/principles-advancing-equitable-data-practice. Accessed June 30, 2022.
- MITRE Corporation. A National Strategy for Digital Health. Technical Paper. March 2022. Available at: https://www.mitre.org/publications/technical-papers/a-national-strategy-for-digital-health. Accessed June 20, 2022.
- Nushi B. Responsible Machine Learning with Error Analysis. Microsoft Corporation website. Artificial Intelligence and Machine Learning. February 18, 2021. Available at: https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/responsible-machine-learning-with-error-analysis/ba-p/2141774. Accessed June 20, 2022.
- Wattenberg M, Viégas F, Hardt M. Attacking discrimination with smarter machine learning. Google Research. Available at: https://research.google.com/bigpicture/attacking-discrimination-in-ml/. Accessed June 20, 2022.
- Partnership on AI. Fairer Algorithmic Decision-Making and Its Consequences: Interoogating the Risks and Benefits of Demographic Data Collection, Use, and Non-Use. December 2, 2021. Available at: https://partnershiponai.org/paper/fairer-algorithmic-decision-making-and-its-consequences/. Accessed June 20, 2022.
- Costanza-Chock S, Raji ID, Buolamwini J. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533213