GLOBALIZING FAIRNESS ATTRIBUTES IN MACHINE LEARNING: A CASE STUDY ON HEALTH IN AFRICA

Anonymous authors
Paper under double-blind review

Abstract

With growing machine learning (ML) applications in healthcare, there have been calls for fairness in ML to understand and mitigate ethical concerns these systems may pose. Fairness has implications for global health in Africa, which already has inequitable power imbalances between the Global North and South. This paper seeks to explore fairness for global health, with Africa as a case study. We propose fairness attributes for consideration in the African context and delineate where they may come into play in different ML-enabled medical modalities. This work serves as a basis and call for action for furthering research into fairness in global health.

1 Introduction

Machine learning (ML) models have the potential for far reaching impact in health. However they also have the potential to propagate biases that reflect real-world historical and current inequities and could lead to unintended, harmful outcomes, particularly for vulnerable populations \[29; 61; 15; 14; 23; 48\]. This has led to the institution of the algorithmic fairness field - various attempts to assess and correct machine learning biases.

Fairness is especially important for global health, which has traditionally been plagued with inequitable power imbalances between high-income countries (HICs) and low- and middle-income countries (LMICs). The recent decolonizing global health movement sheds light on “how knowledge generated from HICs define practices and informs thinking to the detriment of knowledge systems in LMICs”\[20\]. \[20\] explore intersections between colonialism, medicine and global health, and how colonialism continues to impact global health programs and partnerships. They make the point that even the use of the term “Global health” to describe health in LMICs is problematic. Given that most machine learning models are developed with problem formulation, personnel, resources and data from HICs, and may be imported with little regulation to LMICs, there is a risk for algorithmic colonialism \[11; 43\], which would further exacerbate current power imbalances in global health.

To date, fairness in health research has furthered understanding and mitigation of biases along the machine learning development pipeline \[15; 14; 61\]. However most are contextualized to HICs with few studies that have explored contextualizing fairness beyond the west \[55; 21; 10\].

This paper builds on previous work as follows: (1) We identify fairness attributes that should be considered with respect to health in Africa, and (2) discuss implications for machine learning applications and areas for further research.

We acknowledge that Africa is extremely diverse genetically and culturally. Within the continent there are attributes that may apply to some countries and not to others. While there is not a one-size fits all approach, we hope that this serves as a starting point for further research.

2 Fairness Considerations

In this section we discuss axes of disparities for fairness between Africa and non-African countries, and considerations for fairness attributes in the context of Africa. Individuals may have multiple axes of disparities. We detail contextual barriers by health modality in Appendix \[A.1\]. We define key terms in Appendix \[A.2\].
2.1 Africa-specific axes of disparities

Colonial history: All countries in Africa with the exception of Ethiopia and Liberia were colonized by Europeans. There is strong evidence that Africa’s colonial history, and resulting structural challenges continue to create power imbalance between the colonized and colonial rulers and beneficiaries [20]. Colonial history, has been put forward by several scholars as a social determinant of health [63, 52, 17]. We propose colonial history as a fairness attribute that should be considered throughout the ML pipeline - problem formulation, data collection, model development and deployment.

National income level (NIL): Colonialism and its resulting structural issues link to national income level. The number of African countries by NIL are Low-income: 24, lower-middle income:17 upper-middle income: 6, and high-income:1 [45]. Socio-economic disparities resulting from the disproportionately high number of LMICs on the continent imply limited funding from governments for research in ML for health, and limited availability of clinical data generating devices and resources for ML development [44].

Country of origin: Africa is divided into 5 distinct sub regions: North Africa, West Africa, Central Africa, East Africa and South Africa. Countries in these regions have different NIL levels, colonial histories, languages, cultures, ethnicity and racial subgroups. Country lines are colonial constructs, partitions resulting from the scramble for Africa, with the exception of Ethiopia and Liberia. This has implications for ethnicity, resources and development [41]. Different countries may have varying methods of implementing health strategies. Hence machine learning applications that work in one country, may not necessarily transfer to the other.

2.2 Global axes of disparities in the context of Africa

Race: Majority of Africans are racially black and are subjugated to global anti-black rhetorics [51]. North African Arabs are externally racially white, as are Afrikaans in South Africa. There are also people of Asian origin, predominantly in the east and southern Africa. However most Africans do not self-identify by race and are more likely to identify by nationality and ethnicity [4, 37]. Demographic health questionnaires do not typically ask for “race”, though persons of African descent may be at higher risk for certain diseases. One may consider racial disparities that impact Africans on a global scale, or within different countries in Africa for example, between Black and White North Africans [51], or between Black and White South Africans [66]. This has implications for socio-economic disparities in health access, sampling bias and for race-based disparities in ML model performance.

Ethnicity: Africa has over 3000 ethnic groups and is the most genetically diverse continent. Unlike in HICs, most Africans associate with an ethnic group. Ethnicity, in addition to physical traits, language and culture, defines history, aspirations and world view [18]. Socio-economic disparities and approaches to accessing health may run along ethnic lines [22]. This has implications that may skew ML applications towards dominant ethnic groups resulting in ethnic inequities.

Religion: Africans practiced traditional religions pre-colonialism. With the introduction of Western religion, Christianity and Islam have grown to become dominant regions on the continent. Some countries are majority Christian with minority Muslim and traditionalists and vice versa. Religion can be associated with ethnicity and socio-economic status, leading to “ethno-religious disparities” in health [25, 26]. Religion, can also impact access, perceptions and adoption of health practices [68, 56, 47].

Language: There are over 2000 languages in Africa, and most Africans speak more than one [28]. Oral language in Africa typically has the format of a dominant colonial or ethnic language such as English, French, Arabic or Portuguese. There is usually a dominant native language, and several local dialects of less dominance. These dominant languages, particularly colonial languages, facilitate cross-ethnic interaction are typically both written and oral and are utilized in education and professional life [49]. There exist oral and written language barriers for people who speak less known known dialects and for those with lower education and literacy levels, which can impact communication for health delivery [33].

Skin tone: On a global scale, African skin tones, light or dark are associated with blackness and instigate anti-blackness discrimination [51]. Among black Africans, lighter skin tone may be per-
ceived as enabling better social status and favors [46]. As a result 40% of people on the continent, practice cosmetic skin bleaching [5]. Bleaching in itself is a public health concern, increasing risk for severe health problems [34]. Within ML for health this has implications for practitioner attitudes, dermatology, such as skin disease detection, and measurement bias from devices that perform poorly on darker skin.

**Gender:** Historically patriarchal African culture, reinforced by colonial legacies, has led to inequitable distribution of wealth between genders [30]. Female household + child caring duties, and professional duties more recently further burdens and reduce access to basic health amenities for African women [3]. Transgender and gender non-conforming persons in Africa are stigmatized and discriminated against, with few health programs catering for transgender persons [36]. This also has implications for health access, and treatment especially for high sexually transmitted diseases [64].

**Sexual orientation:** Thirty-two countries criminalize homosexuality and it is punishable by death in 3 countries [32]. Homosexual persons are stigmatized and discriminated against which causes exclusion and marginalization within health systems [59] [53].

**Literacy and Education level:** Literacy and education levels impact access to care, health seeking behaviors and understanding of health information, especially when delivered digitally [2]. This has implications not only for one’s own health, but also for the health of their children [13].

**Age:** Age is a global fairness attribute used both in machine learning model development and evaluations for health, due to age-specific incidence rates and co-morbidity. It is essential to demonstrate equitable ML model performance across age groups [40].

**Rural-urban divide:** People living in rural areas may have disproportionate levels of lower socio-economic status, literacy, education and limited access to health facilities. This makes people in rural areas most vulnerable to unethical machine learning practices. However we must ensure just, beneficial ML applications that benefit rural regions, and improve health gaps.

**Socio-economic status:** Individual socio-economic disparities underlie most of the above axes of disparities mentioned. Africa has the second highest wealth distribution gaps [58]. This runs the risk of machine learning models in health perpetuating unfairness in deployment access or inaccuracies from unfair variables towards poorer persons.

**Disability:** Ten to 20% of African populations are affected by disabilities. Disabilities can exacerbate most of the attributes listed due to stigma and inadequate resources and policies at the country level [1][38].

**Health-specific attributes:** Differences in genetic and phenotypic distributions in the population may limit model performance. Underlying pre-existing conditions, and later stages of disease presentation may also impact distribution shift and model performance. may also impact model performance and measurement biases.

## 3 Implications for Machine Learning and Research Opportunities

In light of the axes of disparities above, it is important to understand what implications they have on existing bias mitigation strategies and what particular challenges may arise when building and deploying models within the African context [9][67].

**Contextualization of the fairness criteria:** The initial formalization of popular algorithmic fairness metrics [7], such as demographic parity and equality of opportunity, was intrinsically tied to anti-discrimination laws in the US [8] around disparate impact and disparate treatment. This motivated the development of methodologies to ensure that automated systems did not have discriminatory effects based on contextually-relevant protected attributes such as race and gender. As highlighted in previous sections, the diversity and axes of disparities between protected attributes in HICs and LMICs lead to a fundamental need to study which fairness definitions are pertinent within the local historical context. By virtue of the field still being in its infancy on the African continent, there is a unique opportunity to engage different stakeholders (researchers, policymakers, governance, etc) to define relevant fairness desiderata in accordance with local laws and beliefs. An interdisciplinary approach at this stage is particularly important to avoid potential issues around competing definitions.
of fairness [65, 31, 16] and purely mathematical formulations that lead to unintended performance degradation for all groups [42].

**Practical fairness considerations:** Further studies in the algorithmic fairness field have highlighted the different ways in which bias can manifest in each step of the ML development pipeline namely during (1) problem selection, (2) data collection and processing, (3) algorithm development and evaluation and (4) post deployment considerations [15]. A plethora of debiasing approaches have been proposed in the literature to tackle potential biased during steps (3) through (5) [39, 7]. In addition to better understanding how these mitigation strategies should be adapted to account for varying fairness attributes specific to LMICs, a particular hindrance to the application of machine learning techniques to the health domain in Africa resides in the problem formulation, collection, and utilization of data [50]. Below, we highlight examples of these challenges and important fairness considerations when using machine learning tools in Africa. Details about contextual barriers for different health modality can be found in Appendix A.1.

- **Problem selection:** As noted in [15], biases stemming from unaligned incentives by external organizations can severely affect, for example, which diseases are studied. This has a trickling effect in terms of which data is collected regardless of relevance to the local population. It is particularly important at this step to consider the potential biases of stakeholders and ensure affected communities are included in the decision making step.

- **Scarcity of digitalized health records:** Patients’ health records remain predominantly paper-based due to a lack of infrastructure and computer resources to keep digital copies. This phenomenon is exacerbated in rural areas leading to further inequalities. Awareness of the crucial role of having electronic records in order to utilize machine learning tools ushers in new proposals to digitalize existing handwritten data. Challenging fairness problems in this step include how to develop automated tools that can reliably process underserved African languages as well as how to ensure that such efforts are equally distributed across regions.

- **Prevalence of survey data:** Most application of machine learning to health remain tied to data collected in clinical settings [40]. In the African context, the prevalence of self-reported survey data introduces a number of biases including unbalanced representation of groups based on who answers the survey and potentially incorrect self-reported information about a respondent’s attributes. Application of machine learning methods to such data modality and how to account for measurement biases remain an underexplored area of research.

**Caution around using pretrained models:** Given the scarcity of readily available large training datasets from Africa, a common approach consist of finetuning large pretrained models to downstream tasks. This approach is particularly tempting as it allows to harness the power of large models whose training requires considerable resources and infrastructure within more resource-constrained environments. However, there are several pitfalls in this approach that could lead to unintended biases in the final model. A prevalent concern lies within the inability to adapt fairness properties to distribution shifts. Indeed, demographic shifts (i.e., changes in the distribution of the fairness attribute), covariate shifts (i.e., changes in the distribution of features) or label shifts (i.e., changes in the distribution of outcomes) are likely to occur when deploying models off-the-shelf or finetuning to a target dataset. However, even if bias mitigation techniques are used at training time, there are no guarantees that the fairness property will hold after deployment or after the finetuning process. Ensuring the generalizability and transferability of fairness properties across domains and under distribution shifts is a new and active area of research [57, 24, 62, 6, 54]. Use-cases specific to the African context can serve as motivating examples and drive impactful advances in this area.

### 4 Conclusions

There are unique opportunities for machine learning to make positive impact in Africa and advance global health equity. However, there should be proactive steps taken to prevent harms, reduce biases, and ensure fairness. To develop fair machine learning models, one needs first to understand what the fairness attributes of a given context are and where to apply them. We provide this work as a starting point towards building the foundation for fairness in machine learning in Africa. Future work will involve qualitative and quantitative analysis of machine learning models, how they respond to these attributes, and contextualized mitigation proposals.
REFERENCES


[26] Peter Salama Stanley Gwavuya Ha, Wei and Chifundo Kanjala. *Social science medicine*.


[37] Ariela Schachter Maghbouleh, Neda and René D. Flores. Middle eastern and north african americans may not be perceived, nor perceive themselves, to be white. *Proceedings of the National Academy of Sciences*, 119.7:547–551, 2022.


A APPENDIX

A.1 CONTEXTUAL BARRIERS BY HEALTH MODALITY

The field of ML for health relies on large open-source datasets. However most of these datasets are from countries outside of Africa. Though there are some open-source African health datasets, these are mostly limited to surveys due to limited use of electronic health records. We discuss how these can impacted by fairness attributes across different health modalities.

Radiology: Machine learning has shown the most promise for use in radiology tasks such as mammography, ultrasound, CT, MRIs and X-rays. However radiology imaging devices are limited in
hospitals in Africa, due to their high cost, and maintenance. Even when radiology imaging devices are available, they may be of varying quality, which leads to poor generalization of ML models from HICs. Additionally radiology images may not be digitally stored or connected to other patient health records. This can lead to limited data for retraining models and can generate biases and model errors.

**Visible light imaging:** Visible light imaging, like radiology, impacts several areas of health: microscopy, endoscopy, medical photography for dentistry, dermatology, gynecology and others. Areas of unfairness can arise from differences in types of conditions, and in presentation of conditions. For instance ML-enabled dermatology diagnosis, may have been trained on lighter skin and may not work on presentations of the same disease on dark skin. Additionally there may be different distributions of types of diseases that may not be accounted for in models trained with data from outside the continent.

**Survey data:** Demographic and health survey data present some of the largest, readily available longitudinal sources of health data from the continent. These surveys are used to provide data on disease trends over time by different regions, and demographics. While they allow data representation from African of different backgrounds, when using them for ML model development, care must be taken to ensure that models do not rely on proxy or sensitive attributes that could lead to unfairness towards certain demographic groups.

**Unstructured written health notes:** Machine learning for unstructured health notes utilizes natural language processing models pre-trained on large amounts of text to extract information such as symptoms and action items, and provide disease classifications. However NLP models, pre-trained on large amounts of data already exhibit cultural, gender, and racial biases. Using them without proper evaluation and finetuning may propagate these biases.

**Medical speech:** Automatic speech recognition systems are used in various health facilities by healthcare professionals to dictate notes without having to take time away from patient care. Accent, style of speaking (for instance pidgin) and literacy may impact speech recognition algorithms designed to be used in Africa.

**Optical sensor devices** Optical sensor-based such as pulse oximeters and fitness trackers have been shown to have lower performance on darker skinned persons [60]. Machine learning models developed for these devices may perpetuate measurement bias.

**Facial recognition** Facial recognition algorithms have shown use in health to diagnose medical conditions such as autism [35]. However these algorithms have also been shown to have high performance bias by skin tone and gender [12].

**Health recommenders** A health recommendation system can analyze the status of a patients health to recommend personalized diets, exercise routines, medications, disease diagnoses, or other healthcare services. These may however lack Africa-specific contexts and leading to less relevant recommendations.

### A.2 Definitions

We define terms that are used in this paper. While there is not a standardized definition of fairness, we mainly follow the taxonomy described in [39]. We also provide other terms of general informational value for further understanding of fairness.

**Discrimination:** ‘Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally’.

**Bias:** ‘Bias can be considered as a source for unfairness that is due to the data collection, sampling, and measurement.’

**Fairness:** ‘In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics’

**Types of Fairness**
Under review as a conference paper at ICLR 2023

*Individual Fairness* [19]: ‘Individual fairness is captured by the principle that any two individuals who are similar with respect to a particular task should be classified similarly’.

*Counterfactual Fairness* [?]: ‘The counterfactual fairness definition is based on the “intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group.’

*Group Fairness:*

Broadly, group fairness notions aim to ‘treat different groups equally’. Below are two main statistical group fairness definitions that exist in the literature.

- **Equal Opportunity** [27]: ‘This fairness notion requires that the probability of a person in a positive class being assigned to a positive outcome to be equal for both protected and unprotected (female and male) group members. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates’.

- **Demographic Parity** [19]: ‘requires that the overall proportion of individuals in a protected group predicted as positive (or negative) to be the same as that of the overall population’

*Fairness in Relational Domains:* ‘A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals’

*Subgroup Fairness:*** Subgroup fairness intends to obtain the best properties of the group and individual notions of fairness. It picks a group fairness constraint like equalizing false positive and asks whether this constraint holds over a large collection of subgroups

*Axes of Disparities*

*Africa-specific axes of Disparities:* Disparities that primarily impact Africa

*Global axes of disparities in the context of Africa:* Disparities that have global implications but contextualized for Africa.