Algorithmic Bias in Healthcare Prediction Systems

Tresna Dey

Abstract

With Bangladesh progressing faster into its vision of Smart Bangladesh 2041, the stethoscope is more and more joined by the algorithm. Diagnostic tools developed on the basis of Artificial Intelligence (AI) and used to screen tuberculosis, monitor maternal health, detect diabetic retinopathy, or predict risk factors of non-communicable diseases are no longer abstract research prototypes. Pilot projects are being done in community clinics, tertiary hospitals and digital health platforms nationally.

In most instances, such systems are projected to deliver quicker triage, lower diagnostic back logs as well as provide cost effective decision assistance in a resource-deficient healthcare system.

However, there is a deeper and more subtle threat to consider below the banner of efficiency, and that is algorithmic bias. When AI systems are operated as black boxes that are able to learn on the basis of previously recorded information, then they might learn and possibly extend the social inequalities inherent in this data by default. In a nation where gender inequalities continue to exist, so does the disparity between cities and rural areas, socioeconomic stratification, etc., the challenge of establishing algorithmic fairness is not quite a technical improvement. It is an ethical and social medical necessity.

The machine learning models are essentially data-driven. They lack context awareness and moral reasoning; they identify trends in the data they receive. However, in Bangladesh, data in healthcare is usually scattered, partial, and distorted. The concentration of Electronic Health Records (EHRs) is in urban private hospitals. The community clinics are highly dependent on the paper-based reporting systems. Mobile health applications win their users mainly among the digitally connected populations. This lumpy data ecosystem generates what it is called dataset imbalance by researchers. Training systems based on such data can be systematically poor along the lines of such groups when those groups are underrepresented or misrepresented in the training data. Underperformance in a healthcare setting is not merely reduced accuracy but it can be one or more of the following: delayed treatment, misdiagnosis, or inadequate distribution of limited resources.

Among the most urgent ways of the dataset imbalance in Bangladesh, the gender data gap deserves to be mentioned. Recent studies have shown that, women do not own smartphones as compared to men and their use of mobile internet is still years behind. This digital shadow has far-reaching consequences to AI based health tools. When the cardiovascular disease AI diagnostic application is trained using data that is largely obtained in men (the user group more prone to owning smartphones, health apps, and seeking formal clinical care) the model will capture masculine symptom patterns.

However, cardiovascular disease among women tend to manifest differently, and women have less evident or non-stereotypical symptoms in comparison with typical manifestations in men. Unless the training dataset is representative of female cases, the algorithm will not be able to identify these differences. The outcome may be an overall system that seems very accurate but systematically inaccurate in risk profiling of women.

This is not some imaginary issue. The evidence of the world-wide experience shows that AI diagnostic systems can suffer much elevated error rates in the case of women and people with darker skin complexion when they are trained on non-diverse data sets. In dermatology, e.g. models trained on lighter skin images have been unable to detect diseases on darker skin. In cardiology, models that have been trained on male data have been less sensitive to female patients. In Bangladesh, which is characterized by a structural disadvantage to care access, such as cultural constraints, financial dependency and a mobility constraint, algorithmic bias would exacerbate other inequities. Such a biased system can not only misdiagnose, but it can also support the neglect of a system.

Another dimension of invisibility brought by socioeconomic disparity is the visible poor. Bangladesh healthcare data is skewed towards an urban population. Tertiary hospitals in the urban areas produce comprehensive digital documents whereas in rural areas and in informal settlements there is a lack of uniformity in the use of electronic records. In the char lands, haor regions, and hill tracts, the patients might never be found in organized hospital databases since they are unable to spare funds to travel, consult, and test. When predictive models are trained mostly on hospital records of cities, they will understand patterns that are related to those individuals that commonly use formal healthcare. This produces what may be termed as the black hole effect. Prediction algorithms that help identify high risk in terms of historical spending habits, frequency of visits to the hospital, or the intensity of past diagnostic procedures may tag wealthy urban patients as high priority just because they can afford to interact with the healthcare system more frequently. At the same time, a day worker who seldom visits a clinic (although there might be serious underlaying conditions to consider) will keep being statistically invisible to the model.

In these cases, algorithmic bias will not be caused by ill intent but structural data gaps. When such models play a role in policy making, such as the distribution of diagnostic equipment, workforce in the community clinic, or the provision of subsidies, the effect can be profound. A predictive system that conflates

healthcare utilization with healthcare need can divert resources to those populations that are already well-serviced. In such a way, the presence of algorithmic bias can convert the existing disparities into automated decision-making processes, introducing the structural disadvantage to the digital infrastructure.

To overcome this issue, it is necessary to go beyond aggregate measures of accuracy. Conventionally, the model evaluation will be on the general performance measures like accuracy, precision, recall, or area under the curve (AUC). Nevertheless, a model may have high overall accuracy and to do so badly on a subgroup. To reduce this threat, scholars are employing more and more fairness measures to assess the performance of models on demographic groups. An example of this is demographic parity which evaluates the likelihood of being considered high risk as the same among different groups of people e.g. men and women or urban and rural populations. Equal opportunity extends by investigating the consistency of the true positive rate, that is, the likelihood that a model correctly detects a disease when it is present, between subgroups. This may apply in the Bangladeshi environment and ensure that a model of tuberculosis detection is equally sensitive to detect a positive case in a villager in Kurigram and a resident of Banani.

In addition to measurement, mitigation measures are required. In cases where the datasets are imbalanced e.g. 1000 records of males with only 200 records of females then naturally the algorithms that are used tend to optimize towards the majority group which is the 1000 records of males. Re-sampling methods are artificially imposed to overrepresent the minority groups in the training data either by replicating existing samples or creating artificial ones. In re-weighting, the underrepresented groups are given a higher weight in training the model and, therefore, the inaccuracies associated with such groups are more heavily penalized. In Bangladesh researchers who have tried to work with rural health data sets are investigating how to boost signals of poorly-represented groups without biasing underlying epidemiological trends. Although they are not the best methods, they are significant milestones in achieving fair model performance.

However, fairness cannot be ensured through technical adjustments only. The other necessary dimension is explainability. Clinicians working in high-stakes areas like the healthcare sector cannot afford to base predictions on opaque forecasting. The explainable AI (XAI) methods are also being incorporated into the diagnostic systems to demonstrate what logic the model is using. Some tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can be used to determine which features had most impact on a prediction. As an example, when an AI system signals that a given chest X-ray is the evidence of tuberculosis, it must point out the problematic areas of the lung or the opacities that caused the diagnosis. When, rather, the model seems to derive its prediction on the basis of a non-relevant artifact, e.g., a hospital watermark or embedded text, then this is an indication of a bias in the training data. Transparency allows clinicians to identify the spurious correlations and discard faulty outputs prior to their impact on patient care.

The trust is also promoted by explainability. Physicians in Bangladesh are already subjected to high

workload pressures in the public hospitals of the country. Unless AIs are viewed as transparent and trustworthy, they will not be adopted. Opening the black box allows the developers to offer a collaborative validation between the clinicians and the technologists. This way, AI will not be an authority but a decision support tool.

In the future, the way to ethical AI in Bangladeshi healthcare needs to be made on a structural level on many levels. To begin with, creation of local datasets should be a national priority. The use of training data in the Western or high-income country may be subject to importing demographic assumptions, which are non-congruent to local epidemiology, genetics, or social determinants of health. The biological and social diversity of the population in Bangladesh could be captured in a National Health Data Repository that has strong privacy protections and comprehensive sampling strategies.

This type of repository must also combine rural clinics, urban slums, and marginalized communities, making them representative in terms of age, gender, ethnicity and socioeconomic status.

Second, institutionalization is required before bias audit deployment. Any AI tool that is presented in a Government Upazila Health Complex ought to be evaluated stringently to prove that it can display fair performance in all demographic groups. These audits cannot be formalities but ongoing processes of monitoring because data distributions keep on changing with time. An algorithm that is trained on the 2023 data might not operate on the same basis in the year 2028 because the patterns of diseases and digital behaviors will have changed.

Third, the design should be inclusive in nature and should influence development. Once the algorithms are deployed, it is difficult to retrofit algorithmic fairness. Not only software engineers and data scientists, but also female researchers, rural health workers, sociologists and bioethicists should be part of development teams. Local participation may reveal local variations in the context, like cultural constraints in reporting symptoms or local differences in the manifestation of a particular disease, that technical teams may not grasp. Design inclusive integrates fairness as a statistical limitation into an activity.

Finally, AI can transform the healthcare system in Bangladesh. It may take diagnostics of the level of specialists to distant areas, streamline the distribution of resources, and allow the early identification of diseases that can be prevented. Technology is not in a social vacuum, however. History teaches learners algorithms, and history in Bangladesh, as in most societies, harbors strategies of exclusion and inequality. AI systems have a chance of capturing these patterns in cyberspace, which is dangerous without any conscious intervention.

AI as a challengeers of equality is based on the intentional management, moral sightedness, and the long-term adherence to equity. Algorithms bias is a not an unavoidable effect, it is just a risk to be managed, however, it should be recognized, quantified and mitigated. This is because as we move into the digital era

in Bangladesh, we should not be by letting the impact of technological innovation to increase the disparity between the urban and the rural, the wealthy and the impoverished, the men and women. Rather it must highlight underlying inequalities and overcome remedial action.

Every life of Bangladesh has equal intrinsic value in the era of predictive algorithms and automated diagnostics. It is our duty to create systems that are indicative of that truth. When justice is made a priority and not an option, AI can enhance justice and not inequality. It is time to make sure that when algorithms are part of the stethoscope, the invisible pulse of every citizen, irrespective of the sex, location, or wealth, can be audible equally well and responsibly.