Identifying Biases in Clinical Decision Models Designed to Predict Need of Wraparound Services

Suranga N. Kasthurirathne PhD[1,2], Joshua R. Vest PhD[1,2], Shaun J. Grannis MD,MS[1,2]
[1]Regenstrief Institute, Indianapolis, IN, USA; [2]Indiana University, Indianapolis, IN, USA

Introduction. Evidence of systemic biases in Artificial Intelligence (AI) solutions[1,2] have led to calls to rigorously investigate AI models for biases that impact marginalized and vulnerable populations. However, there has been limited efforts to investigate systemic biases present in AI models for the clinical domain. Previously, we developed a series of AI models capable of predicting need of wraparound services, which are defined as additional non-medical services that are provided in conjunction with primary care[3]. We developed AI models predicting need of referrals to wraparound services for behavioral health, social work, and dietitian visits, as well as other services such as respiratory therapy, financial planning, medical-legal partnership assistance, patient navigation, and pharmacist consultations. These models were implemented across nine federally qualified healthcare centers in Indianapolis, IN to predict need of referrals[3]. In this study, we inspect each AI model for evidence of harmful biases across multiple demographic factors.

Materials and methods. We identified a population of adults (>=18 years) with at least one outpatient visit at Eskenazi Health, a county-owned urban safety net provider located in Indianapolis, IN. We extracted a comprehensive list of patient-level demographic, diagnosis, medication, and past visit history data from the Indiana Network for Patient Care (INPC), one of the largest, continuously operated statewide Health Information Exchanges (HIE) in the United States[4]. We used the Gradient Boosting (XGBoost) classification algorithm to develop four AI models capable of predicting need of referrals for behavioral health, social work, dietitian visits, and all other referral types. As with our previous efforts, we restricted the dietitian referral model to a subset of patients with specific risk conditions[3]. For bias detection, we identified three demographic features (race, age, and gender) as 'protected attributes' which present considerable risk of causing biases[5]. We will use these protected attributes to partition the population into two groups, patients who may be advantaged or disadvantaged based on each attribute. We evaluate biases by investigating statistical measures assumed to be equal across groups partitioned using each attribute (Table 1). Fairness and bias measures are context-dependent constructs. A variety of metrics have been proposed to investigate biases across these constructs. We used the fairness tree method[1] to select the most appropriate bias detection metric for our use case and applied this metric to each AI model using the AI Fairness 360 framework[2], which supports a wide variety of well-established bias detection metrics (Figure 1).

Results. We identified a total of 72,484 adult patients from an urban, primary care safety-net population: predominantly female 47,187 (65.1%), ethnically diverse (~25% white, non-Hispanic), and with high chronic disease burdens. Of these, 15,867 (21.9%) were eligible for inclusion in the dietitian model. Need of referrals, which constituted our gold standard reference, were behavioral health (12,162/72,484, 16.8%), social work (4104/72,484, 5.7%), dietitian counseling (4330/15,867, 27.3%), and other services (17,877/72,484, 24.7%). Performance of each AI model, as optimized for F1-score, was high, and compatible to prior modelling efforts[3] (table 2). We selected False Negative Rate (FNR) parity, which characterizes the degree to which model predictions report similar false negatives scores across advantaged and disadvantaged populations defined by each protected attribute. The fairness tree method recommended this metric because our AI models were designed to be assistive in nature, to prioritize predictive equity for patients in need, and because our interventions were designed to be applied to a broader population[1]. We found that FNR parity for each protected attribute and AI model were considerably low (< 0.07), indicating no evidence of biases (table 3).

Discussion. We were able to reproduce AI models with predictive performance metrics which were both high and comparable to our original effort[3]. Investigation using the AI fairness 360 framework found no indication of biases based on patient age, gender or race across any of the models under test. Therefore, we conclude there is a low likelihood that patient age, gender and race are introducing bias into our algorithms. Next steps include expansion of our analysis to investigate biases caused by social determinants such as homelessness, poverty, and unemployment, and individual-level bias metrics, which contrary to group based metrics used in this effort, investigate biases on the principal that similar individuals should be treated similarly irrespective of any protected attributes[6]. Further, our investigation may be further refined by use of additional advantaged and disadvantaged categories for each protected attribute. While our results indicated considerably low FNR parity scores, determining threshold of bias for larger scores requires a broader conversation with a multi-stakeholder group. In the event that models are found to be biased, they can be re-calibrated using a variety of bias mitigation methods also available via the AI Fairness toolkit.

Table 1. Advantaged vs. disadvantaged values for each protected attribute.

Protected attribute	Advantaged vs disadvantaged values
Gender	Advantaged value: male. Disadvantaged value: all others
Race	Advantaged value: non-Hispanic whites. Disadvantaged value: all others
Age	Advantaged value: 18 - 65 years. Disadvantaged value: >= 65 years

Figure 1. The complete study approach from data collection, AI model development to evaluation of biases.

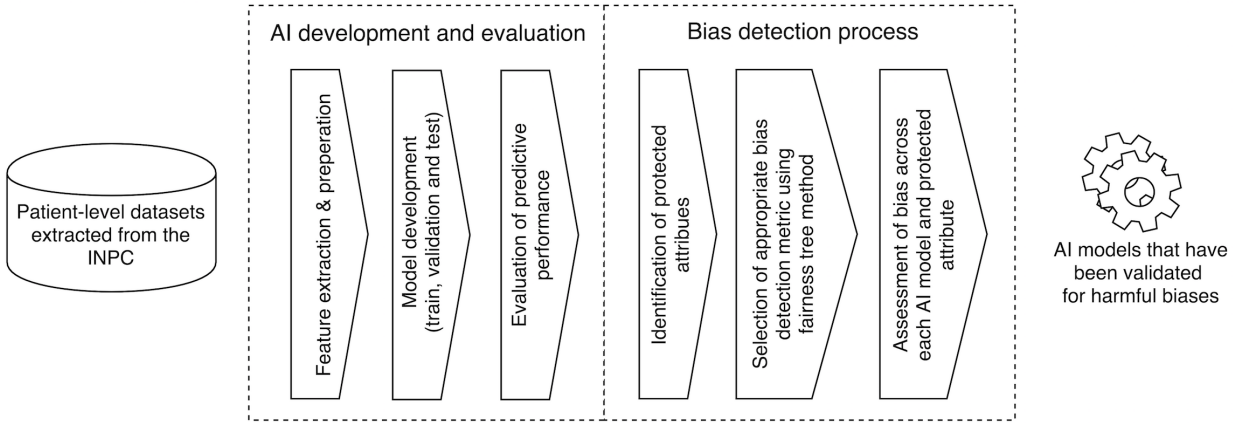

Table 2. AI model performance metrics.

Performance metric	Behavioral health (%)	Social work (%)	Dietitian (%)	Other (%)
Sensitivity	83.5 (83.0, 88.9)	72.5 (69.4, 75.7)	75.13 (70.6, 77.2)	59.1 (56.7, 63.5)
Specificity	99.2 (98.6, 99.8)	99.2 (99.1, 99.5)	93.2 (90.7, 94.4)	92.6 (89.5, 96.2)
F1-score	90.3 (87.5, 93.6)	82.3 (79.5, 85.4)	77.9 (73.2, 80.6)	64.9 (62.7, 67.7)
Precision	95.1 (92.1, 98.2)	95.5 (93.4, 97.5)	79.6 (76.3, 84.1)	73.6 (70.7, 77.3)
AUROC	98.2 (97.5, 98.6)	93.6 (92.7, 95.3)	91.3 (90.2, 92.4)	85.6 (84.5, 86.1)

Table 3. FNR parity for each AI model and protected attribute.

Protected attribute	Behavioral health	Social work	Dietitian	Other services
Gender	0.0504	0.0274	-0.0233	-0.0635
Race	-0.0089	-0.0082	-0.0009	0.0056
Age	0.0334	-0.0320	-0.0139	0.0113

References

1. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:181105577. 2018.
2. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:181001943. 2018.
3. Kasthurirathne S, Grannis S, Halverson P, Morea J, Menachemi N, Vest J. Use of Patient and Population-Level Datasets to Identify Need of Wraparound Social Services: A Precision Health Enabled Machine Learning Approach. JMIR Medical Informatics. 2020.
4. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, et al. The Indiana network for patient care: a working local health information infrastructure. Health affairs. 2005;24(5):1214-20.
5. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, Eng E, Day SH, Coyne-Beasley T. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. American journal of public health. 2015 Dec;105(12):e60-76.
6. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R, editors. Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference; 2012.