*Journal of the American Medical Informatics Association*, 29(7), 2022, 1129–1130
https://doi.org/10.1093/jamia/ocac074
Editorial

Editorial

# Consideration of bias in data sources and digital services to advance health equity

**Suzanne Bakken**

School of Nursing, Department of Biomedical Informatics, Data Science Institute, Columbia University, New York, New York, USA

Corresponding Author: Suzanne Bakken, PhD, RN, FAAN, FACMI, FIAHSI, School of Nursing, Department of Biomedical Informatics, Data Science Institute, Columbia University, 630 W. 168th Street, New York, NY 10032, USA; sbh22@cumc.columbia.edu

Received 2 May 2022; Editorial Decision 2 May 2022; Accepted 3 May 2022

In this editorial, I highlight 5 papers that address expanded data sources and services to understand, contextualize, promote, and predict individual health with careful consideration of bias. Coiera et al[1] promote the idea of family informatics to create a set of digital services to support the family network. Two studies in this issue examine the role of social determinants of health (SDOH) in predictive model performance as a strategy for identifying and mitigating bias.[2,3] Lastly, two papers are about data sharing for a variety of purposes. One explores willingness to share as a potential bias in analytic data sets and algorithms[4] while the other evaluates a privacy-protecting framework for one type of data.[5] As a group, these papers provide additional foundation to advance health equity.

In a perspective, Coiera et al[1] highlight the importance of understanding individuals in the context of their family and argue that this may require new classes of digital services (ie, family informatics) to address important chronic health challenges such as obesity, mental health, and substance abuse, and to support acute health challenges, and promote self-management capacity. They conceptualize the family network as a multiagent system with distributed cognition. They propose that digital tools can address family needs in four key areas: (1) sensing and monitoring; (2) communicating and sharing; (3) deciding and acting; and (4) treating and preventing illness.

Juhn et al[2] applied machine learning models for predicting asthma exacerbation in children with asthma. They measured one SDOH, socioeconomic status (SES), using the HOUsing-based SocioEconomic Status measure (HOUSES) index, to assess its influence on predictive model performance. They also compared incompleteness of EHR information relevant to asthma care by SES. Those with lower SES had a higher proportion of missing information relevant to asthma care (eg, asthma severity). The HOUSES index enables assessment of SES bias in predictive model performance.

Amrollahi et al[3] compared the performance of sepsis readmission prediction models with and without inclusion of SDOH. They used data from the All of Us Research Program participants across 35 hospitals ($n = 8935$ septic index encounters) to develop a multicenter validated sepsis-related unplanned 30-day readmission models with and without SDOH to predict 30-day unplanned readmissions. Incorporation of SDOH factors (eg, economic stability) into the model of clinical and demographic features improved area under the receiver operating characteristic curve significantly (from 0.75 to 0.80; $P < .001$).

Research participant willingness to share types of data sources can influence the representativeness of samples in analytic datasets. Joseph et al[4] examined the willingness of participants in the National Institutes of Health All of Us Research Program to share EHR information. In a sample of 25 852 participants (White—66.5%, Black—18.7%, Hispanic—7.7%, female—32.5%), 2.3% declined to share EHR data. Younger age (1.26 [1.19–1.33]), female sex (1.74 [1.42–2.14]), and education >high school (2.44 [1.86–3.21]), but not race or ethnicity, were significantly associated with decline to share EHR data.

Concerns about privacy may limit willingness to share data. Bonomi et al[5] propose a privacy-protecting method for sharing one type of data, individual-level electrocardiography (ECG) time-series data. Their approach leverages dimensional reduction technique and random sampling to achieve privacy protection against an informed adversarial model while enabling useful aggregate-level analysis while maintaining the usability for data analytics. Their evaluation of the approach on two real-world ECG data sets demonstrated significant reduction in privacy risks while retaining data usability for tasks such as predictive modeling and clustering.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

1. Coiera E, Yin K, Sharan RV, *et al.* Family informatics. *J Am Med Inform Assoc* 2022; 29 (7):1310–15.
2. Juhn YJ, Ryu E, Wi CI, *et al.* Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *J Am Med Inform Assoc* 2022; 29 (7):1142–51.
3. Amrollahi F, Shasikumar S, Meier A, Ohno-Machado L, Nemati S, Wardi G. Inclusion of social determinants of health improves sepsis prediction models. *J Am Med Inform Assoc* 2022; 29 (7):1263–70.
4. Joseph CLM, Tang A, Chesla DW, *et al.* Demographic differences in willingness to share electronic health records in the All of Us Research Program. *J Am Med Inform Assoc* 2022; 29 (7):1271–78.
5. Bonomi L, Wu Z, Fan L. Sharing personal ECG time-series data privately. *J Am Med Inform Assoc* 2022; 29 (7):1152–60.