Ethnicity-based bias in clinical severity scores

Ethnicity-based bias in clinical severity scores

See Articles page e241

In the age of digital health, a major focus lies on the quantitative analysis of large, high-dimensional patient data sets. The goal is to establish clinical decision making that is informed by evidence from this big medical data. A particularly promising application domain and data source for this task are intensive care units (ICUs), which measure patient characteristics at high frequency for hundreds of variables, and which can benefit enormously from computational approaches that can efficiently summarise this wealth of data and give an accurate representation of a patient's state. As an early example of this development, severity scores for patient risk assessment were introduced decades ago, which correlate markers of illness to relevant clinical outcomes. The family of these scores used in the ICU estimate risk of death by measuring acute physiological disturbances (vital signs, biochemical values, and diagnoses) and pathological processes (admission type), pathological reserves (age and comorbidities), and location factors (before ICU admission).[1] One of the earliest ICU severity scores was Acute Physiology and Chronic Health Evaluation (APACHE) I, dating back to 1981, which is based on a dataset of 805 patients and using expert consensus only.[2] Since then, numerous iterations of the APACHE score and other scoring systems have emerged, building on regression techniques and larger datasets.

In *The Lancet Digital Health*, Rahuldeb Sarkar and colleagues[3] assessed whether clinical severity scores include systematic or implicit ethnicity-based bias. Numerous examples show that clinical decision making is influenced by ethnicity-based, implicit attitudes outside of conscious awareness, often referred to as implicit biases.[4]

To address the question of whether severity scoring systems describe patient status without ethnicity-based bias, Sarkar and colleagues[3] evaluated the most frequently used clinical risk-scoring systems for the intensive care unit (APACHE IVa, Oxford Acute Severity of Illness Score, and Sequential Organ Failure Assessment) on two public data sets (Medical Information Mart for Intensive Care III database and the electronic ICU Collaborative Research Database, with 43 832 and 122 919 admissions, respectively). The investigated ethnicities were Black, Asian, Hispanic, and White. They reported no systematic bias in score discrimination for hospital mortality prediction on individual patients when stratifying their predictions according to ethnic groups ($p>0·01$); however, they observed a statistically significant difference in the calibration of the scores ($p<0·0001$) in Hispanic and Black patients. This result means that patients with the same severity score from different ethnic groups have different standardised mortality ratios (SMRs), rendering the use of such scores for triaging or clinical resource allocation across ethnic groups problematic.

In general, the application of scoring systems to individual patients for outcome prediction has limited benefit and should be applied with caution because the CI of a score is too wide and most patients are scored incorrectly.[5] The SMR of severity scoring systems is mostly used on a cohort level; eg, evaluating ICUs over time and comparing similar hospitals.[1] In the COVID-19 pandemic, instead of severity scores, diagnostic imaging and severity of symptoms are mostly recommended as pragmatic triaging factors because of higher sensitivity when treated at an individual (rather than cohort) level.[6,7] The study by Sarkar and colleagues[3] adds to these concerns by showing that cohorts from different ethnicities should not be subject to the same model.

Building on the findings of Sarkar and colleagues,[3] important directions for future research open up. First, the question of causality can be further investigated; specifically, does the implicit ethnicity-based bias of the health-care system result in a bias in the data that manifests itself in a different calibration of the scores, or is the scoring system itself biased? With Simpson's paradox[8] in mind, this question can be explored by accounting for lurking variables, which could strengthen or weaken the observed statistical differences. An example of such a confounder is ICU length of stay, for which the SMR ratio is closer to 1 for patients staying shorter than 24 h, due to a large number of deaths or discharges occurring in the first 24 h in the ICU.[1] As Sarkar and colleagues note, median lengths of stays are statistically significantly different between ethnic groups. Second, because domain adaptation is an active area of research in machine learning,[9] the hope is to improve machine learning techniques to enhance the usability of established scores, by systematically accounting for the effects of potential confounders such as age distribution,

disease types observed, or gender ratio,[10] when trying to apply severity scores across ethnic groups. To conclude, detecting—as Sarkar and colleagues[3] did in their study—understanding and correcting for ethnicity-based biases in clinical severity scores will remain an important challenge in digital health research.

We declare no competing interests.

*Thomas Gumbsch, *Karsten Borgwardt*
karsten.borgwardt@bsse.ethz.ch

Department of Biosystems Science and Engineering, ETH Zürich, Zürich, Basel 4058, Switzerland (TG, KB); SIB Swiss Institute of Bioinformatics, Switzerland (TG, KB)

1 Barlow CJ, Pilcher D. Severity scoring systems and outcome prediction. In: Bersten AD, Handy JM (eds). Oh's intensive care manual, eighth edn. China: Elsevier, 2019: 1173–75.

2 Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981; **9:** 591–97.

3 Sarkar R, Martin C, Mattie H, et al. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. *Lancet Digit Health* 2021; **3:** e241–49.

4 Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *Am J Public Health* 2015; **105:** e60–76.

5 Booth FV, Short M, Shorr AF, et al. Application of a population-based severity scoring system to individual patients results in frequent misclassification. *Crit Care* 2005; **9:** 1–8.

6 Aziz S, Arabi YM, Alhazzani W, et al. Managing ICU surge during the COVID-19 crisis: rapid guidelines. *Intensive Care Med* 2020; **46:** 1303–25.

7 Barros LM, Pigoga JL, Chea S, et al. Pragmatic recommendations for identification and triage of patients with COVID-19 disease in low-and middle-income countries. *Am J Trop Med Hygiene*, 2021; published online Jan 6. https://doi.org/10.4269/ajtmh.20-1064.

8 Simpson EH. The interpretation of interaction in contingency tables. *J Royal Stat Soc* 1951; **13:** 238–41.

9 Venkataramani R, Ravishankar H, Anamandra S. Towards continuous domain adaptation for healthcare. *arXiv* 2018; published online Dec 4. https://arxiv.org/abs/1812.01281 (preprint).

10 Roessler M, Schmitt J, Schoffer O. Can we trust the standardized mortality ratio? A formal analysis and evaluation based on axiomatic requirements. *arXiv* 2020; published online Sep 8. https://arxiv.org/abs/2009.03650 (preprint).