Essential principles towards improving clinical risk assessment tools: A conversation with Uri Kartoun, PhD

Daniela Kamir
Medical Writer, Bioforum Group

doi: 10.56012/hivc3918

Correspondence
Daniela Kamir
Daniela.kamir@bioforumgroup.com

Abstract

Uri Kartoun (PhD in robotics, Ben Gurion University of the Negev, Israel) is a Staff Research Scientist and an IBM Master Inventor, co-developer of technologies such as MELD-Plus, EMRBots, Memory-memory (M2) Authentication, and Subpopulation-based Feature Selection. Prior to joining IBM Research in 2016, Kartoun worked at Microsoft Health Solutions Group and at Massachusetts General Hospital.

EMWA Guest Editor Daniela Kamir, PhD, interviewed Kartoun about clinical risk assessment tools, organ transplant allocation disparities, and how the Model for End-Stage Liver Disease (MELD) score is used to allocate livers for transplantation. The conversation has been edited for brevity and clarity.

Daniela Kamir: What factors are potentially predictive in developing effective clinical risk assessment tools, and how can unbiased feature selection techniques help in this regard?

Uri Kartoun: To develop effective clinical risk assessment tools to help better manage a disease, clinicians and data scientists must select patient characteristics that are potentially predictive, such as a subset of laboratory values, co-morbidities, medications, and genetic profiles. This selection process should incorporate both practical experience and knowledge acquired from scientific manuscripts. With the advancement of machine learning–based technologies,

unbiased feature selection techniques can help recommend which characteristics should be incorporated into these tools.[1,2] Additionally, novel metrics, such as those related to fairness, can aid in designing the next generation of risk assessment tools, beyond just assessing the tools by using traditional metrics such as prediction performance and calibration.

DK: The MELD score is used to prioritise patients on the liver transplant waiting list, with higher scores indicating greater illness severity and thus greater urgency for transplant. What guided the development of the Model for End-Stage Liver Disease 3.0 score?

UK: In a recent announcement, the Organ Procurement and Transplant Network (OPTN) Board has decided to replace the MELD-Na (MELD + serum sodium) with the Model for End-Stage Liver Disease 3.0 (MELD 3.0) score for determining organ allocation priorities in the United States.[3,4] This move comes after the co-creators of the MELD 3.0 score were congratulated for their efforts.[5] The cocreators outlined several principles that guided the development of the new score. These principles included the requirement that all features included in the score must be measurable in an objective fashion, generalisable, devoid of unnecessary volatility without biological significance, and reportable to the OPTN without causing an undue burden. OPTN's decision is expected to have a significant impact on organ allocation in the United States as the MELD 3.0 score is a more refined and accurate way of determining organ allocation priorities and is expected to result in better outcomes for patients in need of liver transplants.

DK: Why does the MELD 3.0 score incorporate sex as a variable?

UK: The MELD 3.0 score has incorporated sex as a variable for two reasons: mitigating sex disparity in access to transplantation and improving prediction performance. The inclusion of sex differences in the MELD 3.0 score corrects for sex disparity caused by creatinine and differences in risk of death, among other factors.[3,6] The primary objective of adding the new sex variable, as well as revising the creatinine coefficient, was to improve fairness across the sexes. Note, however, that assessing fairness quantitatively was not thoroughly discussed in related manuscripts.

DK: What is the significance of using fairness-related metrics to assess the performance of risk assessment tools as used in organ allocation?

UK: As a more modern score that accounts for fairness, it is crucial to assess the performance of the MELD 3.0 using metrics specific to fairness.[5] If performance of fairness-related metrics may be found unsatisfactory then a revised version must be developed urgently (i.e., MELD 4.0). Standard metrics such as discrimination and calibration have been used to assess the performance of the new score, but it is also important to use measures such as statistical parity difference, true positive rate difference, and true negative rate difference to assess fairness within the context of patient characteristics such as sex, race, and age.[7] Overall, the incorporation of sex into the MELD 3.0 score is a step towards improving access and fairness in organ allocation. As further assessments of its performance continue to emerge, it will be interesting to see how this new approach to liver disease assessment and transplantation impacts patients and medical professionals alike.

DK: Can you give an example of how fairness should be assessed?

UK: IBM Research and the Broad Institute of MIT and Harvard have collaborated on a recent study that assessed the performance of widely used risk scores in cardiology, namely the Cohorts for Heart and Ageing in Genomic Epidemiology Atrial Fibrillation (CHARGE-AF) score for AF and the Pooled Cohort Equations (PCE) score for Atherosclerotic Cardiovascular Disease (ASCVD).[8] The study evaluated performance by using standard metrics such as discrimination, calibration, and standard hazard ratios, as well as fairness-related metrics considering sex, race, and age ranges.

Evidence was found of potentially unfair performance, with significant differences in fairness metrics for sex and race in both scores. The study considered three large independent datasets, including the Explorys Life Sciences Dataset, Mass General Brigham, and the UK Biobank.[9] Notably, the sensitivity difference of both scores was much lower for females than males in the intermediate-age subgroups, suggesting that current scores may miss more females at high risk for events, potentially worsening existing sex-related treatment gaps.[10] The findings underscore the importance of evaluating prognostic models across specific subpopulations to better understand the accuracy and potential unfairness of the prognostic information used to drive clinical decisions at the point of care.

This study highlights the importance of assessing the performance of prognostic models using metrics specific to fairness and calls for continued evaluation of widely used risk scores to better understand their impact on patient outcomes across various subpopulations. The collaboration between IBM Research and the Broad Institute of MIT and Harvard provides important insights into the limitations of current risk prediction models and paves the way for more equitable and effective approaches to cardiology risk assessment. Similarly, future versions of the MELD score must exhibit small to non-existent bias across all age ranges and characteristics such as sex and race. These findings underscore the importance of developing healthcare scores that are not biased and that accurately reflect the severity of patients' conditions.

DK: Can you give an example of a method that you developed that could aid in identifying additional features for risk assessment tools and reduce bias?

UK: Subpopulation-based feature selection that was developed as part of another collaboration (between IBM Research and MIT) is an iterative machine learning–based technique used to identify the most important features for risk assessment in specific subgroups of patients and was proved to be superior compared to notable widely used feature selection methods.[1] Incorporating novel covariates that improve performance and fairness is expected to provide clinicians with more accurate and unbiased patient risk assessments. Within the context of liver, new versions of MELD are expected to better fairly rank patients on the liver allocation list, once they incorporate novel features that are also adjusted to optimise fairness-related metrics. Combining these principles with the principles specified is expected to yield better performing and more equitable risk assessment tools in heart, liver, and beyond.[4]

Disclaimer

The views and opinions expressed in this interview are those of Uri Kartoun and do not necessarily reflect the views or positions of IBM.

References

1. Feature selection based on subpopulations and propensity score matching: A coronary artery disease use case using the UK Biobank [cited 2023 Feb 27]. Available from: https://knowledge.amia.org/76677-amia-1.4637602/f007-1.4641746/f007-1.4641747/580-1.4642000/269-1.4641997?qr=1
2. Agrawal S, Klarqvist MDR, Emdin C, et al. Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction. Patterns (N Y). 2021;2(12):100364. doi: 10.1016/j.patter.2021.100364
3. Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: The Model for End-Stage Liver Disease updated for the modern era. Gastroenterology. 2021;161(6):1887–95.e4. doi:10.1053/j.gastro.2021.08.050
4. Ge J, Kim WR, Lai JC, et al. Response to: "Towards optimally replacing the current version of MELD." J Hepatol. 2023;78(3):e100–01. doi:10.1016/j.jhep.2022.11.008
5. Kartoun U. Fairness metrics: Additional principles to consider for improving MELD. J Hepatol. 2023;79(2):e80–1. doi: 10.1016/j.jhep.2023.01.028
6. Ge J, Kim WR, Lai JC, et al. "Beyond MELD" – Emerging strategies and technologies for improving mortality prediction, organ allocation and outcomes in liver transplantation. JHepatol 2022;76(6):1318–29. doi:10.1016/j.jhep.2022.03.003
7. Bellamy RKE, Dey K, Hind M, et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. doi:10.48550/arXiv.1810.01943
8. Kartoun U, Khurshid S, Kwon BC, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. Sci Rep. 2022;12(1):12542. doi:10.1038/s41598-022-16615-33
9. Mehran R, Vogel B, Ortega R, et al. The Lancet Commission on women and cardiovascular disease: time for a shift in women's health. Lancet. 2019;393(10175):967–68. doi:10.1016/S0140-6736(19)30315-0

Author information

Daniela Kamir, PhD, has been a medical writer with the Bioforum Group, a global data-focused clinical research organisation, since 2020. Daniela has an extensive international research background, with an emphasis on molecular biology and new technologies in the sciences and medicine. She is experienced in writing pre-approval regulatory documents and scientific writing.