



# What Epidemiologists Can Do in the Era of Machine Learning and Artificial Intelligence

Akihiro Nishi<sup>1,2</sup> and Kosuke Inoue<sup>3,4</sup><sup>1</sup>Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, California, United States<sup>2</sup>California Center for Population Research, University of California, Los Angeles, California, United States<sup>3</sup>Department of Social Epidemiology, Graduate School of Medicine, Kyoto University, Kyoto, Japan<sup>4</sup>Hakubi Center for Advanced Research, Kyoto University, Kyoto, Japan

Received November 11, 2024; accepted December 1, 2024; released online January 11, 2025

Copyright © 2025 Akihiro Nishi et al. This is an open access article distributed under the terms of Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## To the Editor:

Epidemiologists have qualitatively and quantitatively examined and provided knowledge that helps understand the influence of risk or protective factors, ranging from genetic to socio-political factors, for health and illness outcome variables at the population level. In the 2020s, modern epidemiologists have witnessed the era of machine learning, deep learning, and, more broadly, artificial intelligence (AI), which emulates human intelligence using computers.<sup>1</sup> Many epidemiologists have tried applying AI in their research. However, how epidemiologists can interact with AI has yet to be summarized or itemized holistically. Therefore, in the present letter, we brainstorm and discuss seven perspectives. Although some perspectives have already been hot research areas,<sup>1,2</sup> many remain on the epidemiologists' future to-do list.

First, AI-based prediction models can qualitatively indicate how epidemiologists should theorize the relationship of risk or protective factors for an outcome variable. For example, suppose epidemiologists initially theorize and hypothesize two factors independently cause an outcome ( $X_1 \rightarrow Y \leftarrow X_2$ ), and a more flexible AI-based prediction model exhibits a higher predictive ability than a conventional linear regression model. While the better prediction does not necessarily suggest its usefulness in causal modeling, it may encourage them to update their model to AI-based flexible models (eg, SuperLearner<sup>2</sup>), accounting for the role of a non-linear contribution and potential synergy of  $X_1$  and  $X_2$ , which the conventional model might not incorporate initially. Such contribution can be observed in genetic and multi-omics epidemiology.<sup>3</sup>

Second, AI allows epidemiologists to quantify what was not quantifiable previously. For example, wearable devices have brought large spatiotemporal data; texts and images in electric health records can be convertible to analyzable data using natural language and image processing. Such new data types can help epidemiologists identify new risk or protective factors and detect diseases or illnesses earlier. For example, a previous study applied AI to electrocardiograms to build the prediction model for chemotherapy-induced cardiotoxicity.<sup>4</sup> The other example is the application of deep learning to cardiac magnetic resonance images to assess the size of the ascending and descending

thoracic aorta.<sup>5</sup> Researchers then conducted genome-wide association studies and found a genetic basis for variations in the size of the ascending and descending thoracic aorta, which helps us better understand the disease processes underlying aneurysms and dissections.

Third, AI can enhance causal inference. For example, AI has been widely applied to causal modeling, such as propensity score matching, g-formula, and doubly robust estimators.<sup>2</sup> Recently, AI has also contributed to estimating heterogeneous treatment effects, accounting for the complex interplay of multiple individual characteristics.<sup>6</sup> Such an approach allows researchers to estimate the treatment effect for each individual based on their observed characteristics and target individuals who are estimated to have a large treatment effect (ie, "high-benefit approach").<sup>7</sup> Because conventional subgroup analyses were not able to assess such high-dimensional interaction, the AI application in this field has the potential to promote the discussion about personalized medicine and public health.

Fourth, in addition to conventional association/causation studies involving independent and dependent variables, epidemiologists also work in systems science, such as infectious disease modeling and prediction, where more dynamic and non-acyclic flows of multiple factors are assumed over time.<sup>8</sup> For example, reinforcement learning was used to optimize the dynamic treatment strategies for adult patients with sepsis in the intensive care unit.<sup>9</sup>

Fifth, it is also important to note that AI can be a risk or protective factor for health and illnesses. Human-AI interactions, including using generative AIs (eg, ChatGPT and Gemini), may enhance or impair people's health. For example, AI-driven care robots may enhance care recipients' health by early detection of injury risks, while they may impair mental health by accelerating social isolation from other humans. Thus, AI, using it, and AI-human interactions can be social determinants of health.

Sixth, since it is known that AIs sometimes produce algorithmic unfairness (dependency of the model results on variables representing race, gender, and others),<sup>1,10</sup> epidemiologists should be interested in monitoring the unintended negative consequences of the use of "black-box" AI-based models, especially among vulnerable populations. For example, AI-based prediction models

Address for correspondence: Akihiro Nishi, MD, DrPH, Department of Epidemiology, Fielding School of Public Health, 650 Charles E Young Dr. S, Los Angeles, CA 90095 (e-mail: akihironishi@ucla.edu).

trained by publicly available chest x-ray datasets incorrectly predict “no finding” for females and Black and Hispanic patients.<sup>11</sup> To mitigate such a bias, bias-mitigated techniques (eg, *fairlearn* in Python and *aif360* [IBM’s AI Fairness 360] in R) can be used.<sup>10</sup>

Finally, AIs may help and redefine the work of epidemiologists in the future. For example, *Pcalg* is an AI package in R for causal discovery, which can deliver potential causal structures for provided data (constraint-based search methods) or assign a relevance score for causal diagrams (search-based methods).<sup>12</sup> This implies that the theoretical work that epidemiologists have done—drawing causal diagrams based on their theory and prior knowledge—can be supported by AIs and data. Developing these AI-based epidemiologic methods is also included in the work of epidemiologists.

## ACKNOWLEDGEMENTS

Conflicts of interest: AN is a consultant to Vacan, Inc. and obtained an honorarium from Taisho Pharmaceutical Co., Ltd. and Schmidt Futures. AN is supported by the National Institutes of Health (K01AI166347), the National Science Foundation (#2230125), and the Japan Science and Technology Agency (JST; JPMJPR21R8), which have no role. KI is supported by the Japan Society for the Promotion of Science (23KK0240), JST (JPMJPR23R2), the Japan Agency for Medical Research and Development (AMED; JP22rea522107), and the Program for the Development of Next-generation Leading Scientists with Global Insight (L-INSIGHT) sponsored by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. KI received a speaking fee for the pre-seminar at the 33rd and 34<sup>th</sup> Annual Scientific Meeting of the Japan Epidemiological Association. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

## REFERENCES

1. Sung J, Hopper JL. Co-evolution of epidemiology and artificial intelligence: challenges and opportunities. *Int J Epidemiol*. 2023;52:969–973.
2. Moccia C, Moirano G, Popovic M, et al. Machine learning in causal inference for epidemiology. *Eur J Epidemiol*. 2024;39(10):1097–1108.
3. Pingault JB, O’Reilly PF, Schoeler T, Ploubidis GB, Rijdsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet*. 2018;19:566–580.
4. Yagi R, Goto S, Himeno Y, et al. Artificial intelligence-enabled prediction of chemotherapy-induced cardiotoxicity from baseline electrocardiograms. *Nat Commun*. 2024;15(1):2536.
5. Pirruccello JP, Chaffin MD, Chou EL, et al. Deep learning enables genetic analysis of the human thoracic aorta. *Nat Genet*. 2022;54(1):40–51.
6. Inoue K, Adomi M, Efthimiou O, et al. Machine learning approaches to evaluate heterogeneous treatment effects in randomized controlled trials: a scoping review. *J Clin Epidemiol*. 2024;176:111538.
7. Inoue K, Athey S, Tsugawa Y. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *Int J Epidemiol*. 2023;52:1243–1256.
8. Kompella V, Capobianco R, Jong S, et al. Reinforcement Learning for Optimization of COVID-19 Mitigation policies. arXiv. 2020; 2010.10560.
9. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716–1720.
10. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7:719–742.
11. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176–2182.
12. Naser MZ. Causality and causal inference for engineers: beyond correlation, regression, prediction and artificial intelligence. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2024;14.