*Research in Health & Medical Sciences* *Editorial Article*

*2025 January; 4(2)*

# Integrating machine learning into epidemiologic research and its implications for non-response bias

**Zahra Maleki[1], Reza Khatibi[1], Zahra Yazdansetad[2], Mohebat Vali[*3]**

[1] Health Sciences Research Center, Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran
[2] Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran
[3] Non-Communicable Diseases Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

***Corresponding Author Email: mohebatvali@gmail.com**
**Received: 2026/2; Revised: 2026/6; Accepted: 2026/6**

The integration of Machine Learning (ML) and Artificial Intelligence (AI) into epidemiological research offers unprecedented opportunities for data analysis, prediction, and enhanced inference. However, these opportunities come with methodological challenges—particularly when addressing non-response bias, a longstanding threat to validity in observational studies (1).

## Importance of ML in Epidemiologic Context

Machine learning enhances epidemiologic analyses by handling high-dimensional data, capturing complex nonlinear relationships, and enabling predictive accuracy beyond that of traditional regression models. In causal inference, ML can complement parametric models by reducing model specification bias when functional forms are unknown. However, it introduces the risk of plug-in bias when ML estimates are used directly in effect estimation formulas without careful methodological calibration (2).

## Non-Response Bias in Epidemiologic Research

Non-response bias occurs when individuals selected for a study (e.g., longitudinal cohort or cross-sectional survey) do not provide complete data, and their absence is associated with exposures or outcomes of interest. This leads to non-random missingness, compromising generalizability and potentially skewing estimates. For instance, in patient-reported outcomes research, socioeconomically deprived and non-white respondents had significantly lower response rates, altering the representativeness of survey results (3).

## ML for Predicting and Mitigating Non-Response

Recent evidence shows the utility of ML to predict survey response propensity, thereby supporting targeted retention strategies. In the Millennium Cohort Study, for example, researchers developed supervised ML classifiers to predict response to follow-up surveys, achieving improved predictive performance compared with standard models. These findings highlight that ML can identify patterns of non-response, enabling epidemiologists to implement targeted outreach or adjust analytic weights to address differential participation (4).

## Methodological Considerations and Best Practices

To ensure robust epidemiologic application of ML in the context of non-response bias, the following practices are critical:

Explicit Bias Assessment: Employ fairness metrics and bias detection tools during model development to measure disparities across subgroups.

Weighted and Calibrated Models: Integrate survey weights or inverse probability weighting within ML training to adjust for known nonresponse mechanisms.

Transparent Reporting: Report model inputs, assumptions, and potential limitations according to established guidelines (e.g., STROBE with ML extensions).

Sensitivity Analyses: Perform sensitivity analyses to evaluate how predictions and effect estimates change under different assumptions about missing data mechanisms.

## Conclusion

The application of ML in epidemiology can substantially improve predictive performance and operational efficiency. Nonetheless, the threat of non-response bias remains profound when ML models unknowingly reflect underlying participation disparities. Thus, a commitment to bias mitigation, rigorous methodological scrutiny, and transparent reporting is imperative to uphold the validity and equity of epidemiologic research in the era of AI and ML.

**Keywords:** Machine Learning, Non-Response Bias, Predictive Modeling

## References

1. WIEMKEN, Timothy L.; KELLEY, Robert R. Machine learning in epidemiology and health outcomes research. Annual review of public health, 2020, 41: 21-36. https://doi.org/10.1146/annurev-publhealth-040119-094437
2. MOCCIA, Chiara, et al. Machine learning in causal inference for epidemiology. European journal of epidemiology, 2024, 39.10: 1097-1108. https://doi.org/10.1007/s10654-024-01173-x
3. STEPHENS, Andrew R., et al. Evidence of non-response bias in patient reported outcome measurement information system surveys. Interventional Pain Medicine, 2025, 4.2: 100588. https://doi.org/10.1016/j.inpm.2025.100588
4. BARKHO, Wisam, et al. Utilizing machine learning to predict participant response to follow-up health surveys in the Millennium Cohort Study. Scientific Reports, 2024, 14.1: 25764. https://doi.org/10.1038/s41598-024-77563-8