Resources for AMIA 2024 Artificial Intelligence Evaluation Showcase

I. Stage specific comments

Stage I - Technical Performance Studies

AMIA also strongly recommends adherence to published reporting guidelines, including the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Checklist. The TRIPOD Checklist is a 22-item checklist that aims to improve the transparency and reporting of a prediction model, whether for diagnostic or prognostic purposes. Below are recommendations to help satisfy adherence to Statistical Analysis Methods (10d) within the TRIPOD Checklist.

- Specify all measures used to assess model prediction performance and, if relevant, to compare multiple models.
    - Area under the Receiver Operating Characteristic curve (AUROC)
    - Area under the Precision-Recall Curve (AUPRC)
    - Sensitivity, specificity, and positive predictive value (PPV)
- Specify measures for model calibration, if available
    - Hosmer-Lemeshow (HL) goodness-of-fit test
    - Brier Skill Score (BSS)
- Specify measures for model fairness (bias), if available
    - AUROC, AUPRC, sensitivity, specificity, and PPV for each demographic subgroup, e.g., race subgroups

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015 Jan 6;162(1):55-63. doi: 10.7326/M14-0697. Erratum in: Ann Intern Med. 2015 Apr 21;162(8):600. PMID: 25560714.

Stage II - Usability and Workflow Studies

AMIA …

II. References

1. AMIA 2021 panel: Panel: Making Health AI Work in the Real World: Strategies, innovations, and best practices for using AI to improve care delivery https://docs.google.com/presentation/d/1RHML0xYUcTOm5wZJK5rWQCfabr2aKpYC/edit#slide=id.p28

2. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015 Jan 6;162(1):55-63. doi: 10.7326/M14-0697. Erratum in: Ann Intern Med. 2015 Apr 21;162(8):600. PMID: 25560714.

3. Echo Wang, H., Landers, M., Adams, R., Subbaswamy, A., Kharrazi, H., Gaskin, D. J., … (2022). A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *Journal of the American Medical Informatics Association*, *29*(8), 1323–1333. https://doi.org/10.1093/JAMIA/OCAC065

4. Elm, E. von, Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines

for reporting observational studies. *BMJ*, *335*(7624), 806–808. https://doi.org/10.1136/BMJ.39335.541782.AD

5. Estiri, H., Strasser, Z. H., Rashidian, S., Klann, J. G., Wagholikar, K. B., McCoy, T. H., & Murphy, S. N. (2022). An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *Journal of the American Medical Informatics Association*, *2022*(0), 1–8. https://doi.org/10.1093/JAMIA/OCAC070

6. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27 (4): 621–33.

7. Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth, Tina Seto, Steven C Bagley, Nigam H Shah, A framework for making predictive models useful in practice, *Journal of the American Medical Informatics Association*, Volume 28, Issue 6, June 2021, Pages 1149–1158, https://doi.org/10.1093/jamia/ocaa318

8. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health 1, e271–e297 (2019)

9. Park, Y., Hu, J., Singh, M., Sylla, I., Dankwa-Mullan, I., Koski, E., & Das, A. K. (2021). Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. JAMA Network Open, 4(4), e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909

10. Park, Y., Jackson, G. P., Foreman, M. A., Gruen, D., Hu, J., & Das, A. K. (2020). Evaluating artificial intelligence in medicine: Phases of clinical research. *JAMIA Open*, *3*(3), 326–331. https://doi.org/10.1093/JAMIAOPEN/OOAA033

11. Patel VL. Cognitive science in the evaluation of medical AI systems. BMJ Health Care Inform. 2023 Dec 14;30(1):e100929. doi: 10.1136/bmjhci-2023-100929. PMID: 38101808; PMCID: PMC10728996.

12. RKE, B., A, M., & S, N. (n.d.). AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev*, *63*(4/5), 4:1-4:15.

13. Salwei, M. E., & Carayon, P. (2022). A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery. Journal of Cognitive Engineering and Decision Making, 16(4), 194-206. https://doi.org/10.1177/15553434221097357

14. Megan E. Salwei, Pascale Carayon, Peter L.T. Hoonakker, Ann Schoofs Hundt, Douglas Wiegmann, Michael Pulia, Brian W. Patterson, Workflow integration analysis of a human factors-based clinical decision support in the emergency department Applied Ergonomics, Volume 97, 2021, 103498, ISSN 0003-6870, https://doi.org/10.1016/j.apergo.2021.103498.

15. Salwei, M. E., & Carayon, P. (2022). A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery. Journal of Cognitive Engineering and Decision Making, 16(4), 194-206. https://doi.org/10.1177/15553434221097357

16. Shortliffe EH. Role of evaluation throughout the life cycle of biomedical and health AI applications. BMJ Health Care Inform. 2023 Dec 11;30(1):e100925. doi: 10.1136/bmjhci-2023-100925. PMID: 38081766; PMCID: PMC10729087.

17. Shortliffe, E.H, Sepùlveda,M-J, and Patel, V.L, Framework for the Evaluation of Clinical AI Systems. In: Cohen TA, Patel VL, and Shortliffe EH (Eds)), *Intelligent Systems in Medicine and Health: The Role of AI*. 2022, Springer, London, UK http://www.shortliffe.net/AIM-textbook.html

18. Skivington, K. et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *Br. Med. J.* 374, n2061 (2021).

19. Vasey, B., Clifton, D. A., Collins, G. S., Denniston, A. K., Faes, L., Geerts, B. F., … McCulloch, P. (2021). DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.*, *27*(2), 186–187. https://doi.org/10.1038/s41591-021-01229-5

20. von Elm, E. et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Br. Med. J.* 335, 806–808 (2007).

21.