Biases in EHR Databases; a Medical vs Statistical Approach through the ICU Readmission Case

Konstantina Remoundou, Institute of Communication and Computer Systems, NTUA, Greece &

Emanuele Koumantakis, Department of Clinical and Biological Sciences, University of Turin, Italy

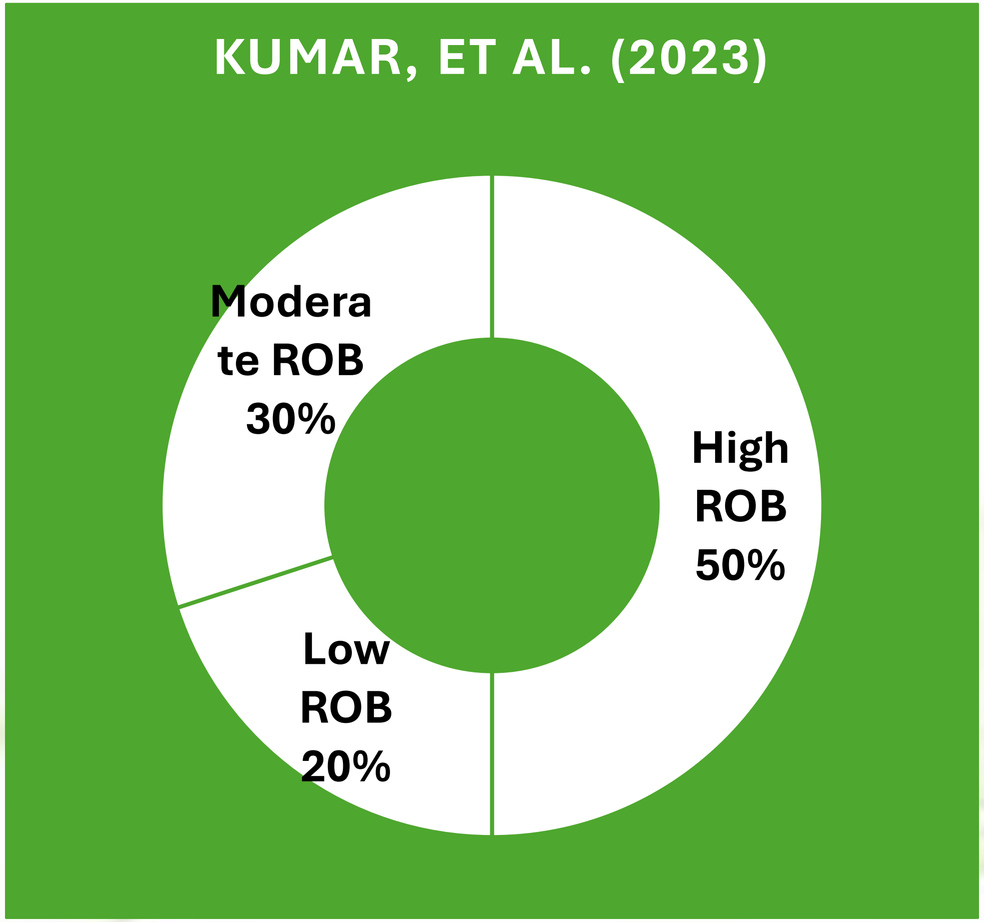
The ICU case

Exchange Scheme: Risk Assessment and Decision Support for ICU Readmission Prediction

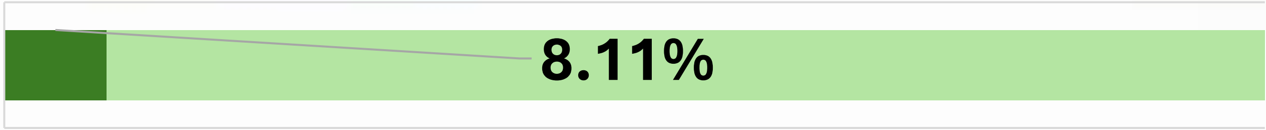
- Intensive care unit (ICU) readmission prediction
- Adults (≥ 16 yo) admitted in ICU
- Deep learning (DL) models
- Studies' publication cut day: 4 March 2025
- Only English language studies

5.41%

Medical and Health AI biases in publication

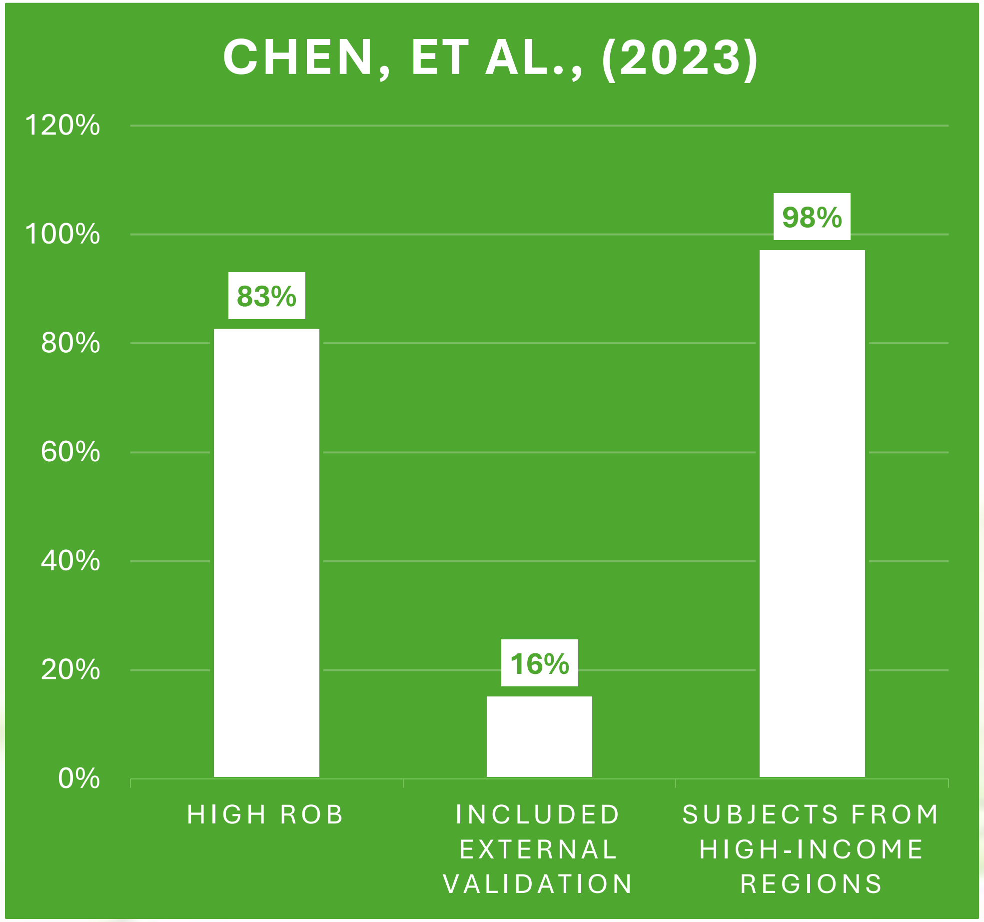


- Using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and a standardized methodology to estimate risk of bias (ROB)
- 48 studies distributed across tabular, imaging, and hybrid data models
- Often related to absent sociodemographic data, imbalanced or incomplete datasets, or weak algorithm design

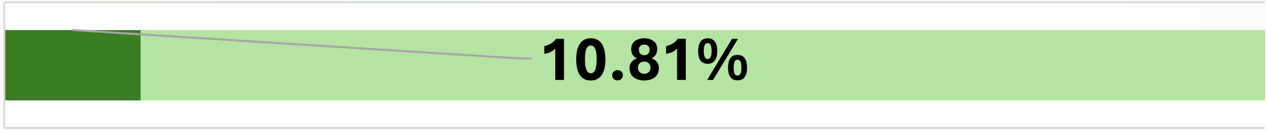


UNIVERSITÀ DI TORINO

Medical and Health AI biases in publication



- Using the PROBAST (Prediction model Risk Of Bias ASsessment Tool) framework
- 555 published neuroimaging-based AI models for psychiatric diagnosis



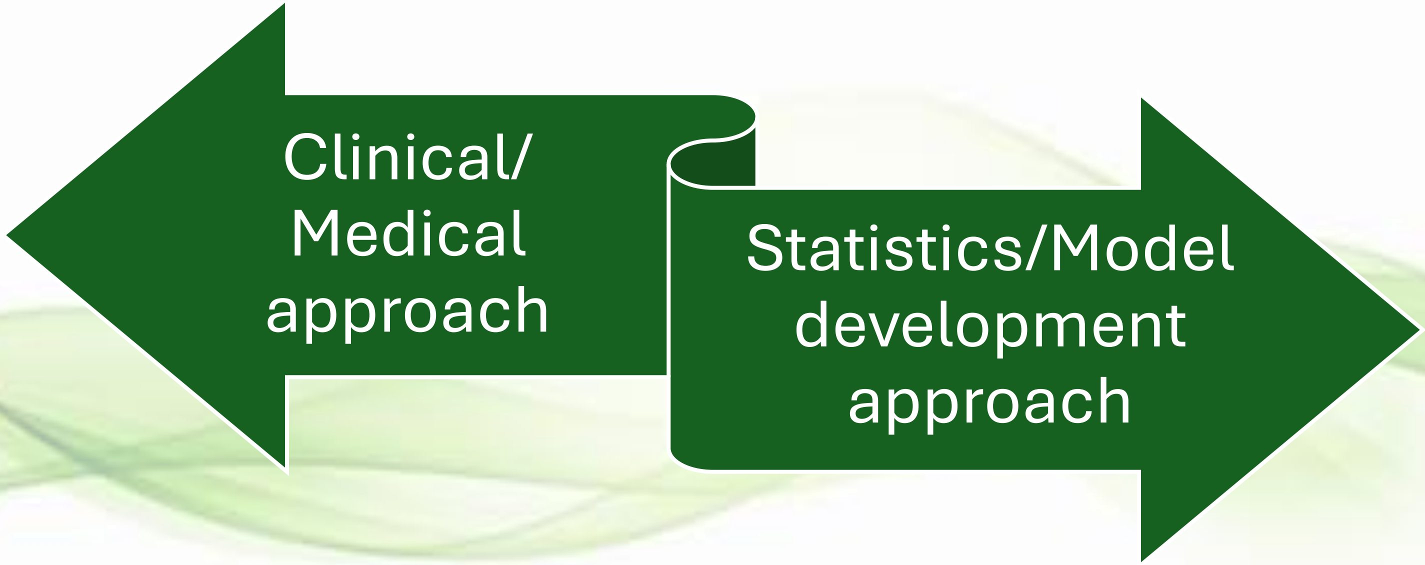
Metrics for biases used in studies

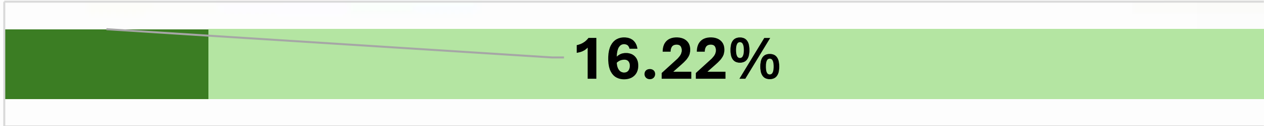
The systematic review from Chen F et al. (2024), focusing on AI models used for HER datasets, concluded that, out of the 20 selected studies:

- 8 studies (40%) applied only performance metrics as sensitivity, specificity, accuracy, mean squared error (MSE) and AUROC.
- 12 (60%) employed fairness metrics and all of them focused on group fairness which tests for some form of statistical parity (eg, between positive outcomes, or errors) for members of different protected groups.

Objective

These studies emphasize a critical need for improved awareness of bias in healthcare AI, and the routine adoption of mitigation strategies capable of bridging model conception through to fair and equitable clinical adoption.





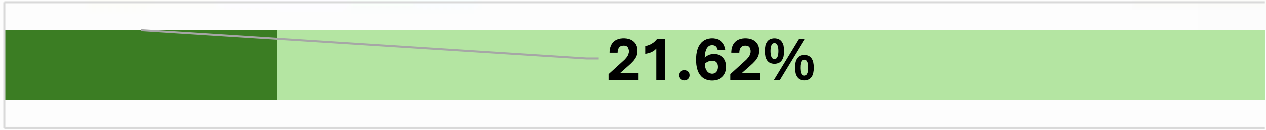
UNIVERSITÀ DI TORINO

Medical Approach: Reproducibility

Reproducibility

DL models must be reproducible to be reliable for clinical application. To achieve full reproducibility the following criteria should be satisfied:

1. Technical reproducibility
2. Statistical reproducibility
3. Conceptual reproducibility (= replicability)



Technical reproducibility

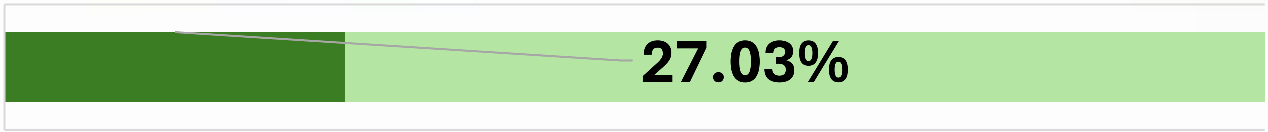
= the ability of an independent research team to produce the same results using the same DL method based on the documentation made by the original research team. To achieve this, the following should be shared:

- Dataset
- Code (preprocessing and model)

Technical reproducibility

= the ability of an independent research team to produce the same results using the same DL method based on the documentation made by the original research team. To achieve this, the following should be shared:

- Dataset **…few are public, de-identification/usefulness balance**
- Code (preprocessing and model) **…may not run correctly**

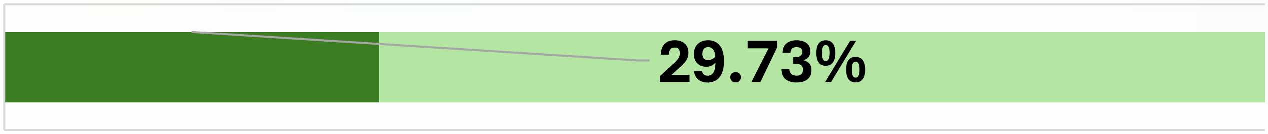


Technical reproducibility

= the ability of an independent research team to produce the same results using the same DL method based on the documentation made by the original research team. To achieve this, the following should be shared:

In our review

- Dataset → **All studies (theoretically)**
- Code (preprocessing and model) → **Only four studies (17%)**

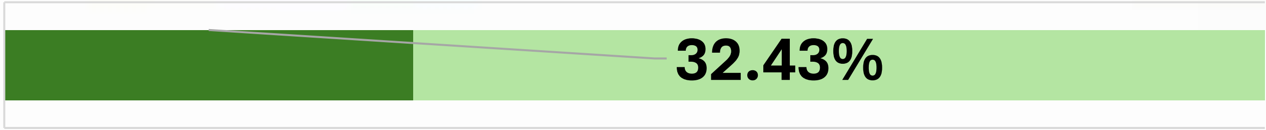


Statistical reproducibility

= the ability to obtain statistically equivalent results under resampled conditions (= internal validity). Generally addressed by DL model development studies, but how to assess?

- K-fold cross-validation and/or other data splits

- Variance (e.g., SD) of performance metrics is reported



Statistical reproducibility

= the ability to obtain statistically equivalent results under resampled conditions (= internal validity). Generally addressed by DL model development studies, but how to assess?

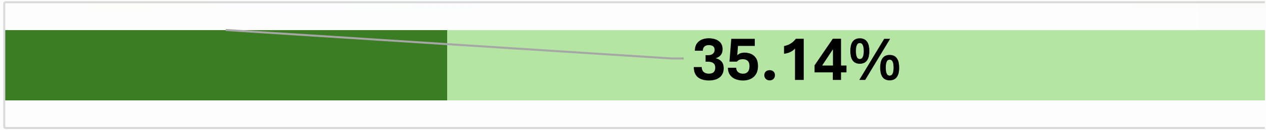
In our review

- K-fold cross-validation and/or other data splits.

→ Only one study did not report internal validation method

- Variance (e.g., SD) of performance metrics is reported.

→ 11 of 22 (50%) studies reporting AUROC



UNIVERSITÀ DI TORINO

Conceptual reproducibility (= replicability)

= the ability to reproduce the desired results under conceptually similar conditions (= external validation). Task-definition dependent. Issues:

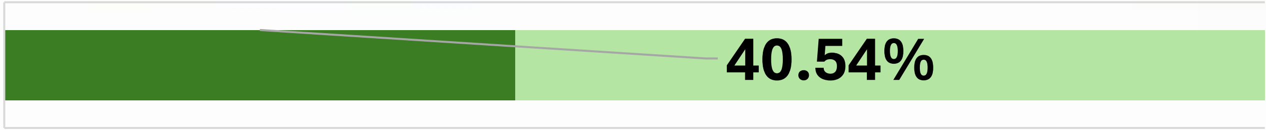
- External validation rarely performed
- Multi-institution datasets

Conceptual reproducibility (= replicability)

= the ability to reproduce the desired results under conceptually similar conditions (= external validation). Task-definition dependent. Issues:

In our review

- External validation rarely performed → **2 studies (8%)**
- Multi-institution datasets → **3 studies (13%). No studies integrated multiple datasets**



Statistical Approach: Generalization

Predictive modelling

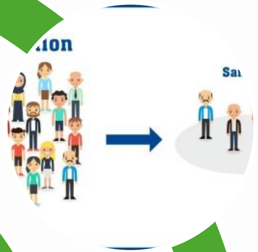
- Predictive modeling → multiple factors/predictors for accurate prediction (causal inference/reasoning)

- The main focus of prediction model studies is the overall predictive or diagnostic performance of the model which should also be **assessed in new patients (validation) →** ***generalization***

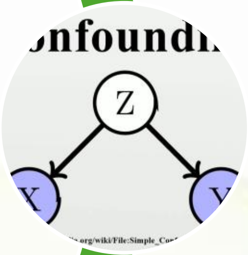
45.95%

Common Types of statistical Biases from AI

Statistical/ Development Biases



Sampling Bias



Confounding Bias

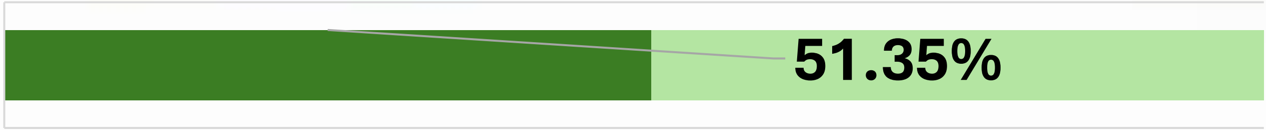
Algorithmic Bias

48.65%

Sampling bias

- This type of bias, aka population bias, occurs when the way of the selected objects is leading to results representing specific groups of data and not the targeted population.
  - Asking or selecting the wrong population/characteristics
  - Missing the necessary response

- Question in focus “*why do some patients have complete data and others do not?*”



Common Types of Sampling/Population bias

ENFIELD

Selection bias

- Systematically selected population is the not correct one based on the *inclusion* and *exclusion* criteria for the specific problem
- Sampling Frame Bias - the sampling frame used to collect data does not cover the entire population of interest

Survivorship Bias

- How many participants “survived” during the duration of the study
- “Non-survivors” means losing participants from any cause (e.g. death, leaving the study, injury, etc.) not related to the study objectives, at any point of the study.
- Is death/mortality an event in the ICU readmission case?

Non-Response Bias

- Missing values and the way to manage those.
- Participants not responding due to ethical/psychological reasons
- Reporting problems from the medical team (textual and nontextual, forgetfulness of reporting, etc.)

54.05%

UNIVERSITÀ DI TORINO

Causes of sampling bias

- Inadequate Data Collection
  - Reporting issues → such as low-quality data coming from low-income countries e.g. the study from Tolera A. et al. (2024) that showed a shortage of data entry formats and/or delays in supplies that affected healthcare data quality.
- Data Preprocessing
  - Various data types in machine-learning approaches employed within the predicted process → 3 most frequent data types include *image, text, and tabular/numerical* making data cleaning and data transformation a difficult task. (Albahra et al., 2023)
  - Missing values handling → different statistical approaches for each case and scope
- Data Imbalance/ fairness
  - when not all observed characteristics are equally represented in the dataset
  - Representative results through clearly defining the target population. All characteristics being equally represented.

UNIVERSITÀ DI TORINO

56.76%

Confounding factor

- Confounding factors may mask an actual association or, more commonly, falsely demonstrate an apparent association between the treatment and outcome when no real association between them exists.
- For confounding bias, the relevant question is: *why did a patient receive one particular drug over any other?*
- *The effects of confounding may result in:*
  - *An observed association when no real association exists.*
  - *No observed association when a true association does exist.*
  - *An underestimate of the association (negative confounding).*
  - *An overestimate of the association (positive confounding).*

Confounding outcomes

- n the study by Ramspek et al. (2021), 30% of the prediction studies reviewed interpreted included predictors in a causal manner, by suggesting that modifying a predictor could improve a patient's prognosis..
- **Misinterpretation**: since predictors in prediction models *do not need to be causally associated* with the outcome, such studies cannot validly conclude that an individual's prognosis would change if these predictors were modified.
- Another study presenting a dementia risk score, concluded that a high BMI is protective. These conclusions may mislead readers into thinking obesity has health benefits (Li J, et al. 2018)
- Confounding also constitute to poor research methodology if not be adjusted properly.
- Unfortunately, machine learning algorithms cannot distinguish mediators from confounders or recognize bias (Lin S-H, et al. 2020)

UNIVERSITÀ DI TORINO

Dealing with Confounders

Identifying and manage the cofounding factor, generally focusing on controlling it or remove it entirely.

Assessing predictive models

Predictive models are assessed by their prediction accuracy. Cross-validation through k-fold.

(Chyzhyk D. et al. 2022; Soneson et al., 2014)

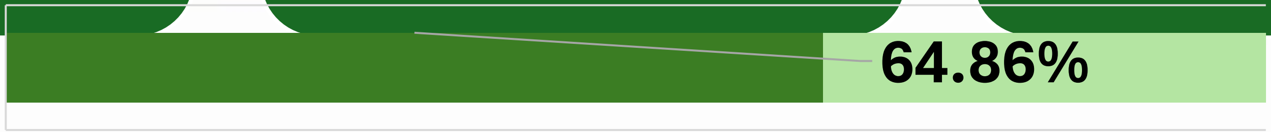
Deconfounding

Removing the confounding factor after discovering it.

Many papers propose different approaches of doing this, depended on the targeted problem. (Zhao et al., 2020; Zhang et al., 2019)

Controlling the confounding factor

Minimize or stabilize the factor after identifying it (D. Chyzhyk, et al., 2018; Chyzhyk D. et al. 2022)



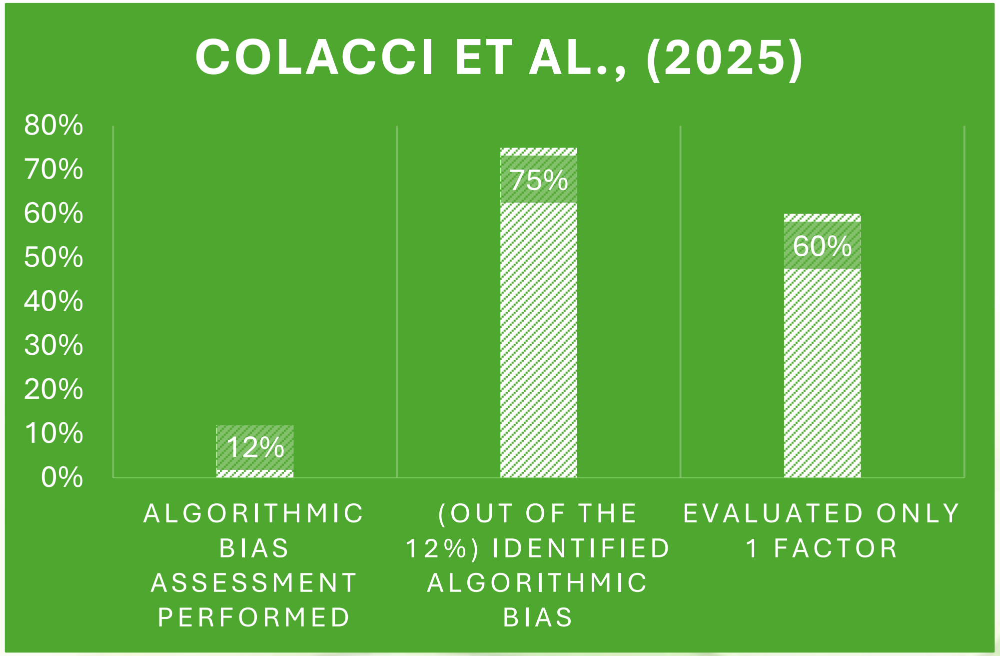
Algorithmic bias

- Algorithmic bias emerges because of wrong assumptions made during the training of prediction models, frequently mirroring biases present in the real world or originating from incorrect or insufficient datasets leading to bad trained algorithms.
- Within the realm of healthcare, this bias has the potential to result in **inaccurate diagnoses/ decision making** or **suboptimal interventions.**
- All parties should focus on **fairness in the data & equal treatment of the patients.**

Algorithmic bias

Colacci et al., (2025) through a study of 760 articles reporting on a clinical ML model, concluded that

- Algorithmic bias assessments was only performed in roughly 12% of the articles, 75% of which identified a bias
- The efficaciousness of bias mitigation techniques ranged from 25% to 89%, with participant reweighting and varying model type being the most effective methods



70.27%

The black-box nature of deep learning

Lack of a clear standard of fairness

Challenges in Health Systems

Lack of contextual specificity

72.97%

Mitigating the Algorithmic bias

Data approach

- Correct framing of the problem – target population
- Data diversification and representation (equality)
- Managing bias in data preprocessing (missing data, confounders, etc.)

Technical approach

- Eliminating bias during model development and validation
- Equitable model implementation (following reporting guidelines and updates in data) (Kolbinger et al., 2024)
- Create an explainable framework of the model (SHAP, LIME, visual explanation, etc.)

Explainability

Explainability

The adoption of DL models as advanced decision-making tools in healthcare is limited by their lack of transparecy and interpretability → «black box» problem

A potential solution are Explainable Artificifial Intelligence (XAI) methods

The most commonly used is SHAP, followed by LIME and GradCAM (Aziz et al., 2025)

UNIVERSITÀ DI TORINO

81.08%

XAI methods

Can be defined based on:

1. Stage → post hoc vs ante hoc
2. Applicability → model-agnostic vs model-specific
3. Scope → global vs local
4. Form → rule-based vs. visual representation
5. Type → e.g., feature important rankings

UNIVERSITÀ DI TORINO

83.78%

XAI methods – an example

Shapley Additive Explanations (**SHAP**). Assigns each feature an importance value for a particular prediction by calculating Shapley values derived from cooperative game theory.

1. Stage → post hoc
2. Applicability → model-agnostic
3. Scope → global and local
4. Form → rule-based and visual representation
5. Type → feature important rankings

XAI methods - SHAP

In our review

Two studies employed SHAP:

- Lim et al. (2025) found that peripheral oxygen saturation, respiratory rate and heart rate were key predictors for ICU readmission.
- Pishgar et al. (2022) highlighted the importance of severity scores for the specific ICU subpopulation of patients with heart failure diagnosis.

UNIVERSITÀ DI TORINO

Take-home messages: should we trust AI predictions?

- DL model development articles still showing high RoB.
- Reporting issues reduce fairness and reliability of AI models.
- Reproducible and generalizable results are fundamental for clinical applicability.
- Even if the above are achieved, explainability still remains a major concern and should be adequately addressed.
- Medical staff, researchers and developers should work in alignment and under common conception of the results aimed.

Thank You!

Bibliography

- Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S. V. S., Ockunzzi, S., Lallo, D., Hu, B., & Rashidi, H. H. (2023). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. Seminars in Diagnostic Pathology, 40(2). https://doi.org/10.1053/j.semdp.2023.02.002
- Aziz NA et al, 2025. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. doi: 10.1101/2024.08.10.24311735
- Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. J Am Med Inform Assoc. 2024 Apr 19;31(5):1172-1183. doi: 10.1093/jamia/ocae060. PMID: 38520723; PMCID: PMC11031231.
- Chen Z, Liu X, Yang Q, et al. Evaluation of Risk of Bias in Neuroimaging-Based Artificial Intelligence Models for Psychiatric Diagnosis: A Systematic Review. JAMA Netw Open. 2023;6(3):e231671. doi:10.1001/jamanetworkopen.2023.1671
- Chyzhyk D, Varoquaux G, Milham M, Thirion B. How to remove or control confounds in predictive models, with applications to brain biomarkers. Gigascience. 2022 Mar 12;11:giac014. doi: 10.1093/gigascience/giac014. PMID: 35277962; PMCID: PMC8917515.
- Colacci, M., Huang, Y. Q., Postill, G., Zhelnov, P., Fennelly, O., Verma, A., Straus, S., & Tricco, A. C. (2024). Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. Journal of Clinical Epidemiology, 178, 111606. https://doi.org/10.1016/j.jclinepi.2024.111606
- D. Chyzhyk, G. Varoquaux, B. Thirion and M. Milham, "Controlling a confound in predictive models with a test set minimizing its effect," 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), Singapore, 2018, pp. 1-4, doi: 10.1109/PRNI.2018.8423961.
- Gundersen OE and Kjensmo S, 2018. State of the Art: Reproducibility in Artificial Intelligence. doi.org: 10.1609/aaai.v32i1.11503.
- Haneuse, S. (2016). Distinguishing Selection Bias and Confounding Bias in Comparative Effectiveness Research. Medical Care, 54(4), e23–e29. https://doi.org/10.1097/mlr.0000000000000011
- Kolbinger, F.R., Veldhuizen, G.P., Zhu, J. et al. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med 4, 71 (2024). https://doi.org/10.1038/s43856-024-00492-0
- Kumar, A. et al. Artificial intelligence bias in medical system designs: a systematic review. Multimed. Tools Appl 83, 18005–18057 (2024).
- Li J, et al. Practical risk score for 5-, 10-, and 20-year prediction of dementia in elderly persons: Framingham Heart Study. Alzheimers Dement. 2018;14(1):35–42
- Lim, Leerang et al., Multicenter validation of a machine learning model to predict intensive care unit readmission within 48 hours after discharge. eClinicalMedicine, 2025. Volume 81, 103112

Bibliography

- Lin S-H, Ikram MA. On the relationship of machine learning with causal inference. Eur J Epidemiol. 2020;35(2):183–5.)
- McDermott MBA et al, 2021. Reproducibility in machine learning for health research: Still a ways to go. doi: 10.1126/scitranslmed.abb1655.
- Nazer, L., Zatarah, R., Waldrip, S., Janny, X. C. K., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2(6), e0000278–e0000278. https://doi.org/10.1371/journal.pdig.0000278
- Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health. 2019 Dec;9(2):010318. doi: 10.7189/jogh.09.020318. PMID: 31788229; PMCID: PMC6875681.
- Pishgar M, Theis J, Del Rios M, Ardati A, Anahideh H, Darabi H. Prediction of unplanned 30-day readmission for ICU patients with heart failure. BMC Med Inform Decis Mak. 2022 May 2;22(1):117. doi: 10.1186/s12911-022-01857-y. PMID: 35501789; PMCID: PMC9063206.
- Ramspek, C. L., Steyerberg, E. W., Riley, R. D., Rosendaal, F. R., Dekkers, O. M., Dekker, F. W., & van Diepen, M. (2021). Prediction or causality? A scoping review of their conflation within current observational research. European Journal of Epidemiology, 36(9), 889–898. https://doi.org/10.1007/s10654-021-00794-w
- Ratwani, R. M., Sutton, K., & Galarraga, J. E. (2024). Addressing AI Algorithmic Bias in Health Care. JAMA, 332(13). https://doi.org/10.1001/jama.2024.13486
- Skelly, A., Dettori, J., & Brodt, E. (2012). Assessing bias: The importance of considering confounding. Evidence-Based Spine-Care Journal, 3(1), 9–12. https://doi.org/10.1055/s-0031-1298595
- Smith, J., Holder, A., Kamaleswaran, R., & Xie, Y. (2023). Detecting algorithmic bias in medical AI-models. ArXiv.org. https://arxiv.org/abs/2312.02959v3
- Soneson, C., Gerster, S., & Delorenzi, M. (2014). Batch Effect Confounding Leads to Strong Bias in Performance Estimates Obtained by Cross-Validation. PLoS ONE, 9(6), e100335. https://doi.org/10.1371/journal.pone.0100335
- Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies. Mobile Information Systems, 2022, 1–20. https://doi.org/10.1155/2022/8167821
- Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., & Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Medicine, 10(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381
- Tolera A, Firdisa D, Roba HS, Motuma A, Kitesa M, Abaerei AA. Barriers to healthcare data quality and recommendations in public health facilities in Dire Dawa city administration, eastern Ethiopia: a qualitative study. Front Digit Health. 2024 Mar 14;6:1261031. doi: 10.3389/fdgth.2024.1261031. PMID: 38550717; PMCID: PMC10972939.
- Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, Matsui Y, Nozaki T, Nakaura T, Fujima N, Tatsugami F, Yanagawa M, Hirata K, Yamada A, Tsuboyama T, Kawamura M, Fujioka T, Naganawa S. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2024 Jan;42(1):3-15. doi: 10.1007/s11604-023-01474-3. Epub 2023 Aug 4. PMID: 37540463; PMCID: PMC10764412.
- Zhang, L., Wang, Y., Ostropolets, A., Mulgrave, J., Blei, D., Hripcsak, G., Zhang, Y., Wang, A., Ostropolets, J., Mulgrave, D., Blei, G., & Hripcsak, M. D. (2019). The Medical Deconfounder: Assessing Treatment Effects with Electronic Health Records. Proceedings of Machine Learning Research, 106, 1–22. https://proceedings.mlr.press/v106/zhang19a/zhang19a.pdf
- Zhao, Q., Adeli, E. & Pohl, K.M. Training confounder-free deep learning models for medical applications. Nat Commun 11, 6010 (2020). https://doi.org/10.1038/s41467-020-19784-9

UNIVERSITÀ DI TORINO