Abstract: Stop training your model, read me first: The importance of dataset choice. Lessons learned from COVID-19 X-ray imaging models.

Beatriz Garcia Santa Cruz[1,2], Matias Nicolas Bossa[3], Jan Soelter[2], Frank Hertel[1,2], Andreas Husch[2]

[1]Centre Hospitalier de Luxembourg, Luxembourg
[2]University of Luxembourg, Luxembourg
[3]Vrije Universiteit Brussel, Belgium
garciasantacruz.beatriz@gmail.com

The robust translation of medical imaging-based models from research to real clinical settings opens new challenges. A prominent recent case is the development of models for the prediction of COVID-19 pneumonia from planar X-Ray imaging. Hundreds of models, intended for clinical use, were published within the last months. In a critical appraisal, most published models were characterised for having a high risk of bias, hampering their safe clinical use [1]. One of the main causes for such risks might be the use of rapidly collected and poorly described datasets during the model development.

Attempting to address this, we conducted the first systematic review of publicly available COVID-19 X-ray datasets focusing on the datasets characteristics that potentially induce bias and/or uncontrolled confounders into the models [2]. Using PRISMA, we identified 112 unique datasets and only 11 were found suitable for further analysis. We evaluated the risk of bias aspects using adapted CHARMS and BIAS tools. To quantify the impact of such issues, we conducted a dataset frequency analysis of published papers over 12 months. We found that in general, the dataset description was poor calling for better documentation. Also, dataset remixes combining several source datasets might lead to accidental data leakage. Temporal analysis revealed researchers tended to choose the datasets based on their size and availability instead of their quality. These results highlight the necessity of paying more attention to the dataset selection process. Our work concludes with a set of general practical advice for modellers to reduce these common pitfalls. We anticipate our article is not only useful for COVID-19 researchers but a representative case of the current state of the area.

References

1. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj 74 (2020).
2. Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem. Med Image Anal 74 (2021), p. 102225.