**Assessing sex-based fairness across patient-level prediction models**

**Aniek F. Markus**[1]

[1]**Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands**

## Background

The Patient-Level Prediction (PLP) framework enables the development and validation of models according to best practices [1]. It is long agreed upon that performance should (at least) be assessed in terms of discrimination (e.g. area under the receiver operating characteristic curve, AUROC) and calibration (e.g. observed versus predicted probabilities) to determine whether predicted probabilities are accurate [2]. However, it is recognized that fairness should be assessed across specific groups (e.g. based on sex, age or other characteristics) as there might be differences in model performance [3]. This work aims to 1) assess model fairness between groups based on birth sex for two prediction tasks, and 2) investigate whether the model choice or the use of group-specific models might lead to fairer outcomes.

## Methods

The study design to evaluate fairness between males and female is summarized in Figure 1. The experiments were conducted on the Integrated Primary Care Information (IPCI) database using two prognostic prediction tasks [4]: (1) prediction of 5-year risk of dementia in elderly individuals aged 55-84, and (2) prediction of 30-day hospital readmission risk following an inpatient visit.

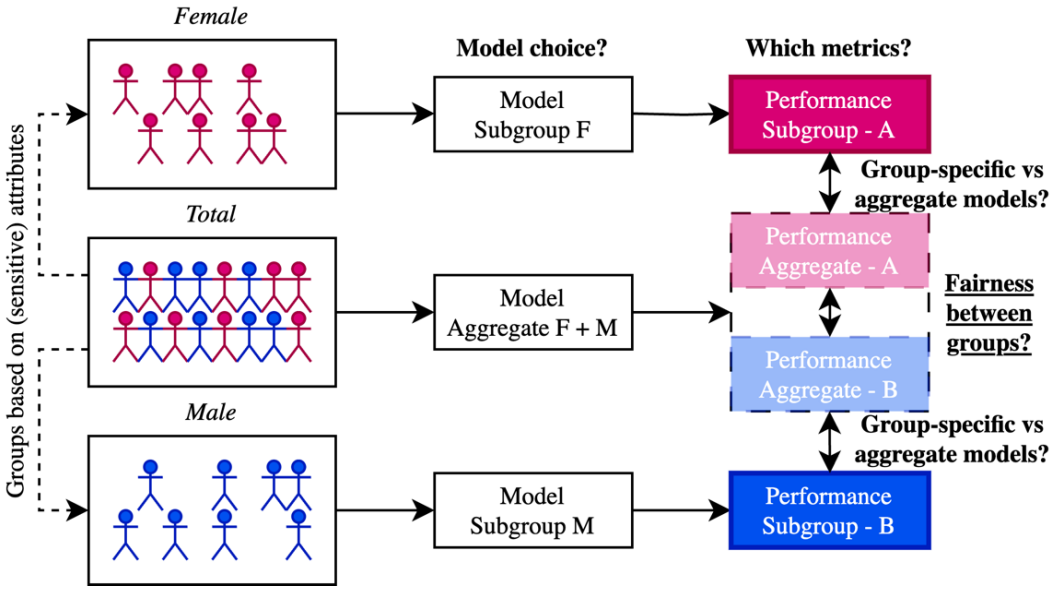


**Figure 1. Overview of study design.**

- ***Model training.*** Models were developed using the PLP package (version 6.4.1) with both LASSO logistic regression and XGBoost algorithms. We use a random 75-25% sample split for model training and performance evaluation, respectively. Hyperparameters were tuned using 3-fold cross-validation with default settings. Candidate predictors include patient demographics (e.g. age and sex), along with a set of common medical conditions and drug exposures based on [5].

- ***Population.*** In addition to models trained on the total population, separate models were trained for the subgroups defined by sex to investigate whether group-specific models could improve performance and fairness.

- ***Performance evaluation.*** Predictive performance was assessed using internal validation, both by group (female/male) and across the total population, with conventional metrics for discrimination (AUROC) and calibration (Brier score). In addition, we analyzed error rates across groups to assess the fairness of the clinical prediction models. Group-level fairness was evaluated across risk thresholds by considering the balance of false positive rates (FPR) and true positive rates (TPR) across groups. Equalized odds is more appropriate when outcome rates differ between groups, as opposed to demographic parity, which considers the balance of positive prediction rates across groups.

**Results**

***Sex differences across prediction tasks (Table 1).*** We identified 44,938 patients for the hospital readmission task and 30,985 patients for the dementia task. We found the samples where relatively balanced in terms of female and male representation. The dementia sample included about 15% more females, while hospital readmission sample had only 1.6% more females included. However, we found both outcomes rates were higher in males, with the outcome being 23.3% more frequently observed in the hospital readmission task and 25.0% more in the dementia task.

***Model evaluation using conventional metrics in subgroups (Table 1).*** The best models demonstrated a moderate performance for hospital readmission (AUROC = 0.632, Brier score = 0.095) and good performance for dementia (AUROC = 0.841, Brier score = 0.021) in the total populations. We observed no clear differences in AUROC and Brier score between the subgroups when using the standard approach to develop PLP models (i.e. creating one aggregate model). For females compared to males, performance was slightly better for the hospital readmission task (+0.018), but worse for dementia task (-0.011).

***Group-fairness using equalized odds (Figure 2).*** To further evaluate fairness we investigate whether errors are similarly distributed across groups. We investigate equal opportunity (TRP) and mistreatment (FPR). In a best case scenario, the TPR is as high as possible (no missed outcomes) whereas FPR is as low as possible (no false alarms). Although AUROC performance across groups is similar, we find that the types

of errors made differ as shown by the large negative (positive) values for the difference in rates. For females, FPR/TRP are both lower for the hospital readmission task, while they are both higher for the dementia task. Depending on the how different type of errors are valued, these differences have implications for the equity across groups.

***Model choice matters (Table 1 & Figure 2).*** When comparing LASSO and XGBoost we found the AUROC does not differ substantially. XGBoost is marginally better to predict hospital readmission (+0.005), LASSO to predict dementia (-0.007). However, differences are observed when evaluating equalized odds: the XGBoost models have slightly fairer error distributions with more balanced FPRs between groups for both tasks and more similar TPRs in the dementia task.

***Group-specific models are not more fair (Table 1 & Figure 2).*** For both evaluation approaches, we consistently find that models trained by group do not lead to improved performance. The differences in AUROC and Brier score for the group-specific versus aggregate models are zero or (slightly) negative. When evaluating the equalized odds we find the group-specific models have larger differences in FPR/TPR – especially between for the dementia task – indicating reduced fairness.

**Table 1. Descriptive statistics and performance (in test set). Letters indicate w.r.t. which group absolute difference is measured: M = males, A = aggregate, L = LASSO.**

| | | | | AUROC value | | | | Brier score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Population size | Outcome rate | LASSO logistic regression | | XGBoost | | LASSO logistic regression | | XGBoost | |
| | | | | Aggregate | Subgroup | Aggregate | Subgroup | Aggregate | Subgroup | Aggregate | Subgroup |
| **Hospital Readmission** | **Total** | 44,938 | 11.0% | 0.627 | 0.626 (A: -0.001) | 0.632 (L: +0.005) | 0.630 (A: -0.002) | 0.095 | 0.096 (A: 0.000) | 0.095 | 0.095 (A: 0.000) |
| | **Male** | 22,290 (49.6%) | 12.2% | 0.616 | 0.611 (A: -0.004) | 0.623 | 0.627 (A: +0.004) | 0.105 | 0.105 (A: 0.000) | 0.104 | 0.104 (0.000) |
| | **Female** | 22,648 (50.4%) | 9.8% | 0.626 (M: +0.018) | 0.629 (A: +0.003) | 0.631 (M: +0.008) | 0.628 (A: -0.003) | 0.086 | 0.086 (A: 0.000) | 0.086 | 0.086 (0.000) |
| **Dementia** | **Total** | 30,985 | 2.3% | 0.841 | 0.833 (A: -0.007) | 0.833 (L: -0.007) | 0.808 (A: -0.026) | 0.021 | 0.021 (A: 0.000) | 0.021 | 0.022 (A: +0.001) |
| | **Male** | 14,411 (46.5%) | 2.5% | 0.849 | 0.840 (A: -0.009) | 0.835 | 0.814 (A: -0.022) | 0.023 | 0.024 (A: 0.000) | 0.024 | 0.024 (A: + 0.001) |
| | **Female** | 16,574 (53.5%) | 2.1% | 0.838 (M: -0.011) | 0.837 (A: -0.001) | 0.833 (M: -0.002) | 0.828 (A: -0.005) | 0.020 | 0.019 (A: 0.000) | 0.019 | 0.019 (A: 0.000) |

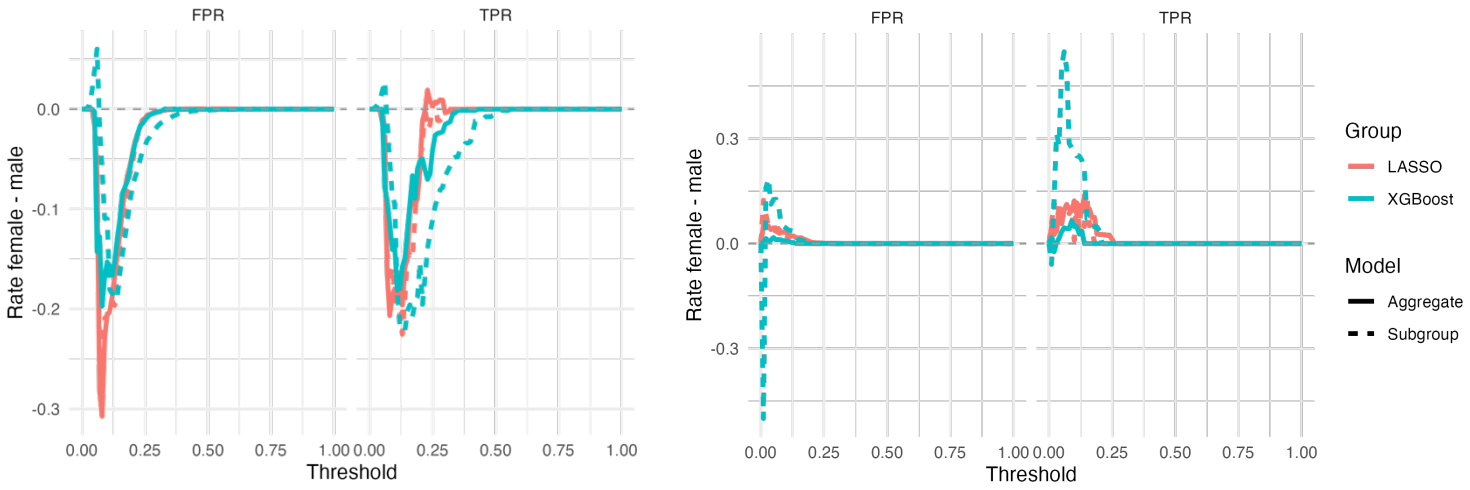


**Figure 2. Differences in equalized odds across thresholds for predicting hospital readmission (left) and dementia (right). Values close to zero indicate no difference in error rates between females and males.**

## Conclusion

Evaluating disparities in performance is important. While conventional metrics are essential, they do not fully capture fairness. For example, we observe no differences between subgroups when examining AUROC and Brier scores; however, error rates reveal distinct patterns between males and females. We conclude that discrimination and calibration alone are not sufficient to assess disparities in model performance across specific groups (e.g. based on sex). Fairness assessments should be integrated into the PLP framework to ensure equitable models, predictions, and decisions that benefit all patients. Establishing best practices for evaluating and improving fairness remains an open challenge in medicine [3]. Therefore, in the future, we will explore additional metrics (e.g. multi-calibration), attributes (e.g. age or other characteristics,) and bias mitigation strategies.

## References

1. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018;25(8):969-75.
2. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
3. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, et al. A translational perspective towards clinical AI fairness. npj Digital Medicine. 2023;6(1):172.
4. de Ridder MAJ, de Wilde M, de Ben C, Leyba AR, Mosseveld BMT, Verhamme KMC, et al. Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands. International Journal of Epidemiology. 2022;51(6):e314-e23.
5. Jenna MR, Jenna W, Egill AF, Chungsoo K, Luis HJ, Ross DW, et al. Finding a constrained number of predictor phenotypes for multiple outcome prediction. BMJ Health & Care Informatics.2025;32(1):e101227.