

International Journal of Population Data Science

Journal Website: www.ijpds.org



Evaluating Accuracy and Fairness of Machine-Learning Osteoporosis Case Definitions in Linked Administrative Data

Hassan Maleki Golandouz¹, Malcolm Doupe¹, Zhiyang Zhou², and Lisa M. Lix¹

¹College of Community and Global Health, University of Manitoba, Winnipeg, Canada

²Joseph J. Zilber College of Public Health, University of Wisconsin-Milwaukee, Milwaukee, USA

Background

Background: Machine-learning case definitions are increasingly adopted in administrative data to accurately identify individuals with health conditions such as osteoporosis for surveillance and research. While overall model accuracy is reported, little is known about model fairness—whether prediction errors are consistent across demographic groups. Differences in these errors can bias disease estimates.

Objective

To develop and validate machine-learning case definitions for osteoporosis and assess their accuracy and fairness across demographic groups.

Methods

We conducted a retrospective cohort study using Manitoba (Canada) administrative data for women aged ≥ 50 with clinically-confirmed osteoporosis. Model features included demographic, comorbidity, and healthcare utilization from hospital, physician, and pharmacy records. Overall accuracy was assessed using area under ROC curve (AUC). Fairness was assessed from age-specific (≤ 60 vs. > 60) false negative and false positive rates (FNR, FPR) with 95% confidence intervals (95% CIs). We compared logistic regression (LR; baseline) with XGBoost, a tree-based boosting method chosen for its strong predictive performance.

Results

The cohort included 58,410 women, 27.7% with clinically-confirmed osteoporosis (2006–2018). XGBoost demonstrated higher AUC (0.86; 95%CI: 0.85–0.87) than LR (0.84; 95%CI: 0.83–0.85). For fairness, XGBoost yielded lower FNRs for both younger (0.28) and older women (0.18) than LR (0.37 and 0.23), meaning XGBoost missed fewer cases. In both models, FPRs were similar in younger women and higher in older women, indicating more false positives in > 60 age-group.

Conclusion

Despite similar accuracy, the models differed in fairness across age groups, underscoring the need to assess fairness when validating case definitions for surveillance and research.

