Title: Comparative Performance of Logistic Regression and Random Forest in Propensity Score Methods: *a simulation study*

Authors: M. Sanni Ali[1], Sara Khalid[1], Daniel Prieto-Alhambra[1,2]

1. Musculoskeletal Pharmaco- and Device Epidemiology, Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK
2. GREMPAL Research Group , Idiap Jordi Gol Primary Care Research Institute and CIBERFes, Universitat Autonoma de Barcelona and Instituto de Salud Carlos III, Barcelona, Spain

Background: Propensity scores (PS) are typically estimated using logistic regression (LR). Machine learning techniques such as random forests (RF) have been suggested as promising alternatives for variable selection and PS estimation.

Objective: To evaluate the comparative performance of LR and RF for PS estimation with respect to covariate balance, bias and precision after PS matching.

Methods: Simulation studies were conducted of binary covariates, treatment and outcome data. In several scenarios, different sample sizes, matching callipers and covariates were created. Treatment effects estimates (relative risk) were derived after PS matching and inverse probability of treatment weighting (IPTW) using Poisson models; covariate balance was checked before and after matching using absolute standardized differences (ASD). Percentage bias (PB) was calculated for each of the proposed models. Mean squared error (MSR) and coverage probabilities were calculated for the different methods.

Results: The performance of LR was comparable to RF in terms of both covariate balance and bias when LR models including interactions and non-linearities included are used: 5% versus 4% ASD and 3% versus 1% PB (n=5000, calliper=0.05). Compared to RF, simpler LR model specification (including only main terms) resulted in suboptimal covariate balance (ASD 4% versus 10%, respectively) and bias (PB X% vs Y%). PB increased in such LR models with the strength of existing interactions/non-linearity up to X% in worst case scenario. Both methods resulted in similar MSR and 95% CI coverage probabilities for PS matching and IPTW when interactions and non-linear terms were included in LR.

Conclusions: PS matching using LR should involve an iterative approach to include interactions and non-linear terms. Alternatively, the use of data driven (without *a priori* model specification) machine learning methods like RF improves covariate balance and results in similar PB in the absence of unmeasured confounding.