

Contribution of machine learning methods to propensity score estimation

A Real World Comparative Analysis of Two Treatment Routes

INTRODUCTION AND OBJECTIVES

Estimating treatment effects in observational studies requires handling confounding factors and correctly modelling causal mechanisms. Traditional propensity score estimation relies on logistic regression, which may perform less well when treatment assignment depends on complex interactions or higher-order relationships between covariates. Machine learning (ML) models can in principle better capture these non-linear patterns, potentially improving covariate balance.

We used a methodology combining causal inference approaches with machine-learning-based propensity score (PS) matching to assess the impact of the route of administration - **Route A** versus **Route B** - on treatment persistence.

METHOD

1. Exploration of confounding factors

Two treatment groups (**Route A** and **Route B**) were matched using propensity scores (PS) derived from 17 covariates including sociodemographic variables and medical history.

Relationships between covariates, treatment and persistence → Correlations and directed acyclic graphs (DAGs).

2. Propensity score (PS) computation

Benchmark of different **ML models** to compute PS: Logistic regression (LR), Decision trees (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), XGBoost (XGB).

→ Estimation of the probability of receiving **Route A** given the covariates.

3. Matching

Propensity score matching was performed without replacement using the **K-nearest-neighbours** algorithm, with a caliper of 0.2 times the variance of the score logit. Covariate balance was assessed using the average standardized absolute mean (ASAM) and standardized mean differences (SMD).

→ Selected ML model: minimises ASAM while maximising the matching rate (MR).

4. Study of the effect of the route of administration on persistence

Cox model: Estimation of the average treatment effect on the treated (ATT).

→ Hazard ratio (HR 95% CI)

CONCLUSION

Integrating causal inference with propensity score matching based on ML models offers a robust framework for evaluating treatment effects in real-world data. In our analysis, more complex ML models tended to produce less stable propensity scores, probably due to overfitting, a limited prevalence of treatment route A, and reduced overlap of propensity scores, which translated into higher ASAM values and a reduced number of matched pairs in small cohorts. Logistic regression achieved better covariate balance across the populations. These results underline the need for careful model selection and suggest that hybrid strategies could improve matching performance and reliability, although their applicability remains context-dependent and must be confirmed by future larger-scale studies.

Abbreviations
AB: AdaBoost; **ASAM:** average standardized absolute mean; **DAG:** directed acyclic graph; **DT:** Decision tree; **GB:** Gradient Boosting; **HR:** Hazard ratio; **LR:** Logistic regression; **ML:** Machine learning; **MR:** matching rate; **PS:** propensity score; **RF:** Random forest; **SMD:** standardized mean differences; **XGB:** XGBoost

RESULTS

Analysis populations

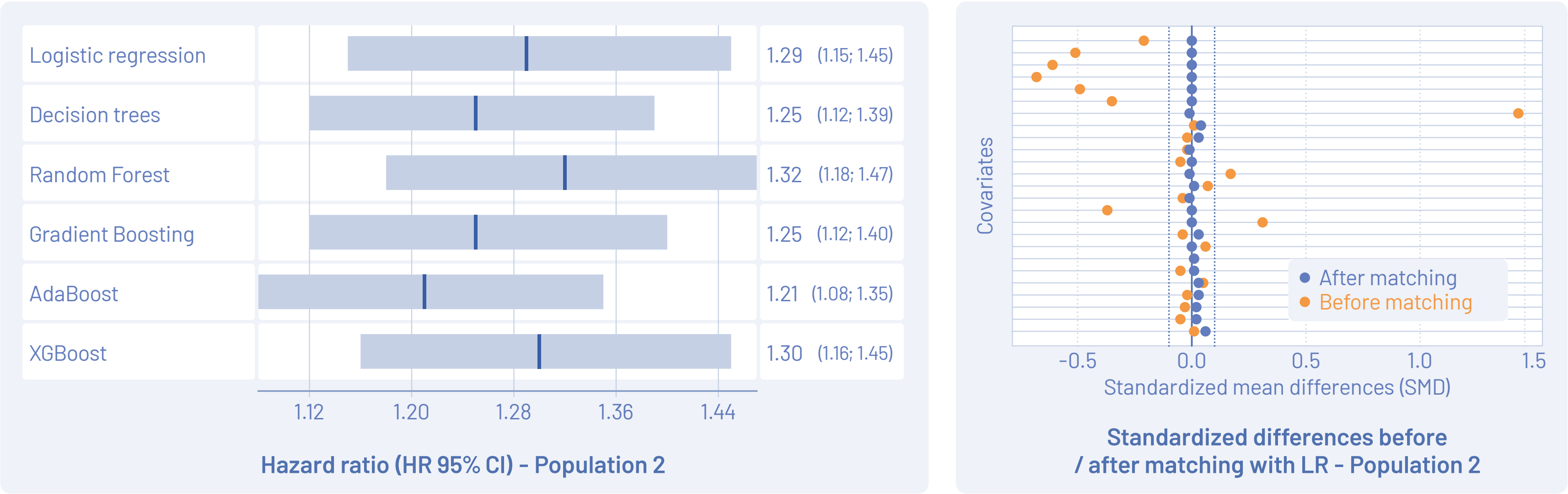
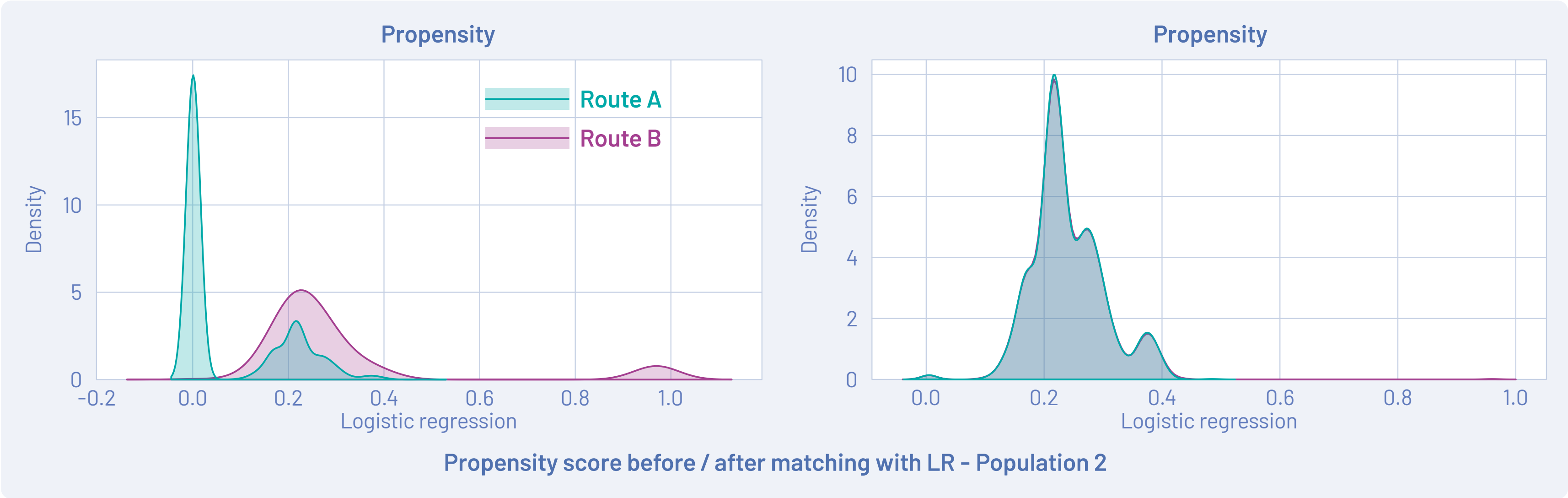
A total of 52,264 individuals were included in the analysis, split into 6 analysis populations of interest:

Population subgroup	Route A	Route B	Total
Overall	4,859 (10.2%)	47,405 (89.8%)	52,264
Population 1	1,086 (11.5%)	9,477 (88.5%)	10,563
Population 2	2,356 (11.1%)	21,209 (88.9%)	23,565
Population 3	823 (11.2%)	7,328 (88.8%)	8,151
Population 4	242 (8.8%)	2,747 (91.2%)	2,989
Population 5	84 (7.6%)	1,104 (92.4%)	1,188
Population 6	42 (3.3%)	1,283 (96.7%)	1,325

The populations of interest comprised between 1,188 and 23,565 individuals, with a **Route A / Route B** ratio between 3.3% and 11.0%. Following the confounding-factor exploration analysis, the 17 variables (sociodemographic characteristics, comorbidities of interest, medical history ...) initially considered were retained to compute the propensity scores.

Main results

For the largest population (2), logistic regression (LR) achieved the best trade-off between covariate balance (ASAM = 0.01) and matching rate (MR = 91%), followed by gradient boosting (ASAM = 0.03; MR = 0.91).



Treatment effects (ATT) on persistence were consistent across models, with HR varying between 1.21 (1.08; 1.35) and 1.32 (1.18; 1.47).

In the other populations (1, 3, 4 and 5), logistic regression achieved the best covariate balance, but not systematically the best matching rate (MR < 65% for the smaller populations).

The other ML models showed reduced performance in the small populations, with ASAM values above 0.1.

Decision trees (DT) were the worst-performing model for all populations, with systematically the highest ASAM and MR. Propensity score estimation computing times were similar across models (max = 1.5 seconds).

Comparative results by population

Population	Number of matched patients	Matching rate (MR)	Best ML model*	ASAM after matching	HR (95% CI)	P-value
1	1,004	92%	AB	0.02	1.39 (1.19; 1.62)	<0.001
2	2,141	91%	LR	0.01	1.29 (1.15; 1.45)	<0.001
3	329	40%	LR	0.03	0.74 (0.57; 0.94)	0.020
4	53	22%	LR	0.07	1.01 (0.59; 1.73)	0.970
5	37	44%	GB	0.10	0.77 (0.32; 1.87)	0.570
6	25	60%	AB	0.07	0.94 (0.42; 2.03)	0.840

*Model minimising ASAM. In case of a tie, the model with the highest MR was selected

