# G-computation and machine learning for causal inference

Arthur Chatton[1,2*], Florent Le Borgne[1,2*], Yohann Foucher[1,3]

[1]INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France
[2]IDBC-A2COM, Pacé, France
[3]Nantes University Hospital, Nantes, France
[*]Co-first authors

E-mail for correspondence: arthur.chatton@univ-nantes.fr

**Abstract:** While machine learning approaches are increasingly used in prediction, their applications for causal inference are more recent. We propose an approach combining machine learning and G-computation (Robins, 1986) to estimate the causal effect of a binary exposure on a binary outcome. We evaluated and compared, through a simulation study, the performances of penalized logistic regressions, neural network, support vector machine, boosted classification, regression trees, and an ensemble method called super learner (van der Laan et al., 2007). We proposed six different scenarios including various sample sizes and relationships between covariates, binary exposure, and binary outcome. We reported that, used in a G-computation approach to estimate the individual outcome probabilities, the super learner tended to outperform other approaches both in terms of bias and variance, especially for small sample sizes. Support vector machine also resulted in performant properties, albeit the mean bias was slightly higher compared to the super learner. In conclusion, the use of machine learning approaches can be pertinent to draw causal inference. Contrary to a preconception, this is true even for sample constituted by several hundred subjects, as in the majority of medical studies. The G-computation with the super learner is available in the R package RISCA.

**Key words:** Causal inference; G-computation; Model specification; Super learner; Simulation study.

**Robins JM** (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, 7(9), 1393–1512.

**van der Laan MJ, Polley EC, and Hubbard AE** (2007). Super learner. Stat. Appl. Genet. Mol. Biol., 6(1), Article 25.