ADDICTION
SSA SOCIETY FOR THE STUDY OF ADDICTION
COMMENTARY

Commentary on Afzali *et al.* (2019): Two data sets are better than one

The use of large data sets in addiction research is welcome, because statistical power is increased. When applied to large data sets, machine learning can help with interpreting variable importance and with quantifying reproducibility. However, application of machine learning in the real world requires consideration of several factors, such as economic cost.

Afzali and colleagues [1] utilized machine learning on two large data sets from different continents and reported reproducible findings on predictors of adolescent alcohol use. Large data sets have the statistical power to yield reliable insights into variables associated with addiction, but they present problems when interpreting results. That is, how can we gauge the relative merits of variables, and how do we know if the result is meaningful or trivial? Null hypothesis statistical testing may not help when the sample is large, because even very small effect sizes correspond to significant *P*-values. Afzali *et al.*'s application of machine learning addresses these issues. However, more development is needed to bridge the gap from research models to practical, cost-effective interventions.

Afzali *et al.* employed two methods to identify the most important variables associated with adolescent alcohol use. First, the best-performing machine was one that employed regularized regression using the Elastic Net [2], a method that automatically selects out the most predictive variables in a data set, performing a similar role to stepwise regression but with distinct advantages. The Elastic Net attenuates overfitting when selecting variables, whereas stepwise regression is especially prone to this [3], and Elastic Net regularization accepts or rejects groups of correlated variables: this is important in addiction research, where measurement variables are typically correlated. Secondly, Afzali *et al.* also systematically included or excluded variables associated with various domains.

Notably, for both data sets, inclusion of all domains produced the most accurate predictions. These results provide guidance for variables that we should assay, given time and budgetary constraints, if our goal is to predict alcohol-related behaviour with high accuracy: measuring psychopathology and personality should be a priority, but it is also worth obtaining some data from a wide variety of other domains.

Using two large independent data sets, Afzali *et al.* were able to use external cross-validation to evaluate the performance of their model. External cross-validation directly tests a model's ability to generalize to previously unseen data because the model is trained on one data set and then tested on a separate data set. External cross-validation therefore speaks directly to replication issues in science [4,5]. The use of external cross-validation in Afzali *et al.*'s work highlights an interesting aspect of this validation method when applied to addiction data. Unlike other data-driven fields (e.g. internet search), addiction researchers cannot easily add more data for validation purposes. Data-driven addiction research is therefore likely to be advanced by interactions among scientists practising 'team science', with research distributed throughout sites to increase statistical power and for testing generalizability [6]. Variables do not have to be identical throughout sites, as was the case in Afzali *et al.*, who used different questions to measure alcohol use in the Australian and Canadian samples. Indeed, the external validation of models despite the use of slightly different variables is a strength—scientific findings should be robust to reasonable deviations in methodology between sites. An avenue for future research should be to validate Afzali *et al.*'s findings in additional samples, particularly in a wider range of cultures.

Afzali *et al.*'s results tell us what the most important predictors of adolescent alcohol use are, but there is a sizable gap between research-focused models and their application on a population level [7]. For example, economic analyses are needed to quantify the relative efficacy of machine learning versus traditional methods to identify adolescents at high risk of alcohol initiation. Machine learning in the wild must accommodate a host of other factors not typically examined in a research setting. For example, the cost of misclassification depends on the nature of any subsequent intervention. If false positives (incorrectly classifying as high risk) result in allocation to a resource-intensive intervention programme, then the machine should be trained to avoid false positives. Alternatively, if the intervention is low-cost and benign (e.g. delivering information online), then the machine should avoid false negatives: it is better to intervene for someone at low risk rather than miss someone at high risk. Furthermore, not all variables cost the same to obtain. It is plausible that a weaker, but cheaper, predictor could be more cost-effective than a stronger, more expensive, one at the population level.

Afzali *et al.* have made a valuable contribution to the addiction literature—a reproducible set of findings

produced by the combination of sophisticated methods and cross-country collaboration. We are still some distance away from the era of personalized interventions at the earliest stage of substance use disorder, but studies such as those by Afzali *et al*. certainly represent encouraging initial steps.

Declaration of interests

None.

Keywords Addiction, adolescence, alcohol, machine learning, reproducibility, team science.

ROBERT WHELAN[1,2]

School of Psychology, Trinity College Dublin, Dublin, Ireland[1] *and Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland*[2]

E-mail: robert.whelan@tcd.ie

Submitted 22 January 2019; final version accepted 31 January 2019

References

1. Afzali M., Sunderland M., Steward S., Masse B., Seguin J., Newton N., *et al*. Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. *Addiction* 2019; **114**: 662–71.
2. Zou H., Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology* 2005; **67**: 301–20.
3. Babyak M. A. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004; **66**: 411–21.
4. Ioannidis J. P. Why most published research findings are false. *PLOS Med* 2005; **2**: e124.
5. Open Science Collaboration Estimating the reproducibility of psychological science. *Science* 2015; **349**: aac4716.
6. Munafò M. R., Nosek B. A., Bishop D. V., Button K. S., Chambers C. D., Du Sert N. P., *et al*. A manifesto for reproducible science. *Nat Hum Behav* 2017; **1**: 0021.
7. Fernandez-Moure J. S. Lost in translation: the gap in scientific advancements and clinical application. *Front Bioeng Biotechnol* 2016; **4**: 43.