

SUPPLEMENTAL FILE 1

1. Least absolute shrinkage and selection operator (lasso) technique

Candidate predictors underwent a variable selection procedure using the least absolute shrinkage and selection operator (lasso) technique [5]. This is a variation from the procedures described in the published study protocol [1]. Penalized regression is the preferred variable selection procedure and decreases the likelihood of overfitting [2]. First, continuous variables across all imputed datasets were scaled to standard scores (z-scores). Next, the optimal lasso penalty λ was chosen for each imputed data set using ten-fold cross-validation based on the deviance. To obtain parsimonious models, λ was chosen to be one standard error larger than the optimal λ [4]. This procedure shrinks some coefficients to zero, selecting a subset of predictors in each imputed dataset. The frequency with which each predictor was selected across all datasets was calculated. A series of candidate models were generated with increasing numbers of included predictors, starting with the most commonly selected predictor. For these candidates, model fit was assessed using Akaike's Information Criterion (AIC). The model producing the smallest combined AIC across all imputed datasets was selected as the final model.

2. Cross-validation

Models predicting dichotomized outcomes perform optimistically within their derived sample [3]. We subjected the multivariable model to a ten-fold cross-validation procedure to internally validate our findings [2;3]. This involved splitting the sample ten-fold, with all previously specified imputation and variable selection procedures repeated in 90% of the sample. The model derived from 90% of the sample is then tested on the remaining 10%.

3. Software Availability

An R package, miPredict, implementing the model selection procedure described here is available from GitHub: <https://github.com/humburg/miPredict>

REFERENCES

- [1] Jenkins LC, Chang W-J, Buscemi V, Liston M, Toson B, Nicholas M, Graven-Nielsen T, Ridding M, Hodges PW, McAuley JH. Do sensorimotor cortex activity, an individual's capacity for neuroplasticity, and psychological features during an episode of acute low back pain predict outcome at 6 months: a protocol for an Australian, multisite prospective, longitudinal cohort study. *BMJ open* 2019;9(5):e029027.
- [2] Retel Helmrich IR, van Klaveren D, Steyerberg EW. Research Note: Prognostic model research: overfitting, validation and application. *Journal of Physiotherapy* 2019;65(4):243-245.
- [3] Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
- [4] Thao LTP, Geskus R. A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal* 2019;61(2):343-356.
- [5] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996;58(1):267-288.