**Appendix: Cross-validation and Machine Learning**

As mentioned in our main article an interesting alternative to Null-hypothesis testing is assessing prediction accuracy of unseen cases as a measure of evidence is based on work by statisticians in the 1970s (Allen, 1974, Browne, 1975, Geisser, 1975, Harrell et al., 1996, Stone, 1974).

In the simplest approach the data set is randomly is randomly split into two (Figure 1a). In the larger set a model will be trained, such as estimating the parameters for a regression. In the smaller test data set we then assess the model e.g. by estimating the prediction accuracy. If sample size is large the prediction accuracy of our model of unseen cases provides a nearly unbiased estimate of the prediction accuracy of new cases. This model validation method is called internal validity. However, we only get valid estimates if we do not perform any model selection. Model selection and model assessment are two separate goals and cannot be assessed on the same unseen data set (Hastie et al., 2009, Varma and Simon, 2006) ). This is often ignored in the machine learning community (Cawley and Talbot, 2010).

If we want to perform model selection we would need to randomly divide the dataset into three parts: a training set, a validation set, and a test set (Figure 1b). The training set is used to fit the models. The validation data set is used to estimate prediction error for model selection. The model with the smallest prediction error is selected as the best model. The test set is used to assess the generalization error of the final chosen model (Hastie et al., 2009). The nomenclature may be confusing because internal validty is assessed using the test data set! This three-way split-sample approach may be feasible for big-data but is usually not possible in medical research and is anyway inefficient and potentially unreliable (Harrell, 2015, Steyerberg, 2009). Two other methods are therefore recommended, cross-validation and bootstrap validation (Harrell, 2015, Hastie *et al.*, 2009).

In n-fold cross-validation, illustrated in Figure 1c for n=5, the single available dataset is randomly divided into n-folds (equal subsets). In turn each fold is used as the unseen data (test set) with the remaining n-1 folds pooled together as the training set. Prediction performance is the average over the n-folds. Though increasingly used, few realise that this approach still gives an over-optimistic assessment. To allow model building and selection (including missing data imputation, transformations or inclusion/exclusion of predictor variables for the training dataset) to be separate from model assessment requires the use of nested cross-validation (Stone, 1974). Illustrated in Figure 1d, the data set is divided into a training set, a test set and a validation set. Different models will be assessed using the validation data set and the best model re-assessed on the independent test fold. This validation and test procedure will again be done on each fold and the final average. Nested cross-validation allows us to get a nearly unbiased estimate of the prediction accuracy of a final model even after performing extensive model selection procedures.

In the alternative bootstrapping approach (Harrell, 2015, Hastie et al., 2009, Steyerberg, 2009) a random sample with replacement is drawn from the data. Models are developed in the bootstrap sample and prediction accuracy tested in the original sample or on cases not selected by the bootstrap sampling. Frank Harrell (Harrell et al 1996 and Harrell 2015) recommends an extension of the bootstrap to estimate the optimism of a performance estimates and to correct own estimates adequately.

****

**Figure 1: Model selection and model assessment using hold-out samples and cross-validation.**

The simplest approach is to randomly split the data set into two (a). In the larger set a model will be trained, such as estimating the parameters for a regression. In the smaller test data set we then assess the model e.g. by estimating the prediction accuracy. If selection of a best model is needed, we split the data into training, test and validation sets (b). The training data set is used to fit different models and based on their performance in the validation data set we select one as our final model. The performance of this model will be then assessed in the independent test data set. If sample size does not allow data to be simply split, we use cross-validation. 5-fold cross validation shown in (c) retains one fifth as an independent test set for model assessment and the remaining for as training. This is repeated for 5 data splits, hence each case is used once as part of a validation data set. The results of the 5 tests folds are averaged and the final model fitted using the total sample. If model selection is undertaken we have to perform nested cross-validation (d). Here, for each split of the data we nest a further 5-fold splitting of the training data. The inner loop is used for model selection and the outer loop for model testing.

**Machine learning**

Cross-validation and bootstrapping to select a model and to estimate the predictive power of unseen cases from the same population (or internal validity) can be used for all statistical models. What modern machine learning methods allow is the searching over hundreds of different models and, if properly implemented, a realistic assessment of the performance of the model in another sample drawn from the same population. Unlike the frequentist and Bayesian approaches, it allows the inclusion of data-preprocessing steps, such as variable transformation, variable selection or imputation of missing data within the model selection process (Kuhn and Johnson, 2013). However, most machine learning methods are “black boxes” and models can lack clinical or scientific interpretability (Shmueli, 2010). Statistical learning methods based on a probability model can provide an integration of good prediction and model interpretability (Iniesta et al., 2016). Regularized and penalized generalized linear models (Lasso, elastic net) are among the most popular (Hastie et al., 2009) and because they are modifications of GLMS, are relatively easy to apply using standard software.

References

**Allen, D. M.** (1974). Relationship between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics* **16**, 125-127.

**Browne, M. W.** (1975). Comparison of Single Sample and Cross-Validation Methods for Estimating Mean Squared Error of Prediction in Multiple Linear-Regression. *British Journal of Mathematical & Statistical Psychology* **28**, 112-120.

**Cawley, G. C. & Talbot, N. L. C.** (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11**, 2079-2107.

**Geisser, S.** (1975). Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association* **70**, 320-328.

**Harrell, F.** (2015). *Regression Modeling Strategies*. Springer: New York, USA.

**Harrell, F. E., Lee, K. L. & Mark, D. B.** (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

**Hastie, T., Tibshirani, R. & Friedman, J.** (2009). *The Elements of Statistical Learning*. Springer-Verlag: New York.

**Iniesta, R., Stahl, D. & McGuffin, P.** (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine* **46**, 2455-2465.

**Kuhn, M. & Johnson, K.** (2013). *Applied Predcitive Modelling*. Springer-Verlag: New York.

**Shmueli, G.** (2010). To Explain or to Predict? *Statistical Science* **25**, 289-310.

**Steyerberg, E. W.** (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York.

**Stone, M.** (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **36**, 111-147.

**Varma, S. & Simon, R.** (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91-91.