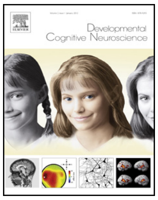
Developmental Cognitive Neuroscience 5 (2013) 172–184

Contents lists available at SciVerse ScienceDirect

Developmental Cognitive Neuroscience

journal homepage: http://www.elsevier.com/locate/dcn



Predicting IQ change from brain structure: A cross-validation study

C.J. Price*, S. Ramsden, T.M.H. Hope, K.J. Friston, M.L. Seghier

Wellcome Trust Centre for Neuroimaging, UCL, London, UK

ARTICLE INFO

Article history:
Received 13 November 2012
Received in revised form 10 February 2013
Accepted 9 March 2013

Keywords:
IQ change, Neuroimaging, Circular inference
Biased sampling
Non-independent errors
Cross-validation

ABSTRACT

Procedures that can predict cognitive abilities from brain imaging data are potentially relevant to educational assessments and studies of functional anatomy in the developing brain. Our aim in this work was to quantify the degree to which IQ change in the teenage years could be predicted from structural brain changes. Two well-known *k*-fold cross-validation analyses were applied to data acquired from 33 healthy teenagers – each tested at Time 1 and Time 2 with a 3.5 year interval. One approach, a Leave-One-Out procedure, predicted IQ change for each subject on the basis of structural change in a brain region that was identified from all other subjects (i.e., independent data). This approach predicted 53% of verbal IQ change and 14% of performance IQ change. The other approach used half the sample, to identify regions for predicting IQ change in the other half (i.e., a Split half approach); however – unlike the Leave-One-Out procedure – regions identified using half the sample were not significant. We discuss how these out-of-sample estimates compare to in-sample estimates; and draw some recommendations for *k*-fold cross-validation procedures when dealing with small datasets that are typical in the neuroimaging literature.

1. Introduction

Neuroimaging data are most commonly used to find brain areas where the functional response, or structural measurement, can be predicted by experimental, behavioural or demographic variables. In this case, the mapping of interest is from behavioural measurements (the independent variable) to brain measurements (the dependent variable) and enable one to infer that certain brain areas are associated with the experimental manipulation. These inferences can then be empirically tested with new data. For example, after demonstrating that the right cerebellum was activated during verbal fluency tasks (Petersen et al., 1988, 1989), the same authors reported that damage to the right cerebellum impaired verbal fluency (Fiez et al., 1992). In other words, a functional imaging study of healthy participants predicted functional specialisation that was confirmed with a neuropsychological (structural imaging) study of patients, which led to a clinically relevant conclusion.

Inferences about cognitive abilities from brain imaging data have also been made in the developmental context. For example, Hoeft et al. (2007) predicted children's reading skills from a combination of behavioural and neuroimaging measures. Cross-validation procedures of the kind reported in Hoeft et al. are essential in this context, because predictions about behaviour will only generalise if they apply to subjects that were not used to select the brain region used to make those predictions. In cross validation procedures, one sample is used to identify brain regions mediating the behavioural phenotype, and another

---

* Corresponding author at: Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London WC1N 3BG, UK. Tel.: +44 203 448 4345.
E-mail addresses: c.j.price@ucl.ac.uk, c.price@fil.ion.ucl.ac.uk (C.J. Price).

1878-9293
http://dx.doi.org/10.1016/j.dcn.2013.03.001

sample is used to predict the behavioural phenotype using those regions. If the same subjects are used in both steps, the predictive validation is circular (non-independent) and there is no replication of the structure-function relationships. This is referred to as “circularity”, “double dipping”, “the non-independence problem” or “biased estimates” (Kriegeskorte et al., 2009; Poldrack and Mumford, 2009; Vul et al., 2009).

In the current study, we illustrate the use of two different cross-validation procedures, with the aim of quantifying how much of the variance in IQ change, measured over the teenage years (Ramsden et al., 2011, 2012), can be predicted from structural brain changes – when the predictions for IQ change are made for subjects that did not contribute to the selection of predictive brain areas (i.e., the region selection and prediction process used independent data). The validation approaches we report, “Leave-One-Out” and “Split-half”, are commonly used to test whether the results of a statistical analysis generalise to an independent sample, and they are especially useful when new samples are costly or difficult to collect.

There are many *re-sampling* techniques that we could have adopted in this context, including bootstrap, jackknife, permutation tests and cross-validation. The “Split half” and “Leave-One-Out” procedures we assess here are both variants of k-fold cross validation. They involve partitioning the full set of data into k non-overlapping samples or sets: $k-1$ samples are used as training sets (e.g., to generate a hypothesis/model) and the remaining sample is used as a validation set (e.g., to test a hypothesis/model). In the present context, this translates to $k-1$ samples being used to identify brain areas associated with a behavioural phenotype (behaviour-to-brain) and the remaining sample being used to predict behaviour from brain measurements (brain-to-behaviour). This procedure is then repeated with different training and validation sets (derived from the same overall sample) – and the results from each iteration/fold are averaged to produce a single estimate. The advantage of repeating the procedure (with k folds) is that all observations are used for both training and validation, without replacement. More specifically, k-fold cross validation only assumes that the original sample is chosen at random from the population and the samples (partitions or subsets) are in turn chosen at random from that original sample. Although the validation and training sets are drawn from the same population, cross-validation is not considered to produce biased results (Efron and Tibshirani, 1997; Hastie et al., 2009).

Variations in the k-fold procedure differ according to how the full sample is partitioned and with the number of iterations used (i.e., the value of k); see review in Arlot and Celisse, 2010. The robustness and appropriateness of each k-fold procedure can be assessed against several criteria; including bias, variance, sensitivity, completeness and computational cost. Different k-fold procedures are expected to perform comparably when the sample is relatively large; however, differences may emerge when the sample size is too small (e.g., in the case of a biased or skewed distribution). Ideally, different procedures can be tested with increasing values of k (varying between 2 to the number of subjects); however, the computational cost can become unmanageable; particularly when many iterations must be performed for a given k value. In the current paper, we compared k-fold cross-validation when the value of k was set to either its lower limit ($k=2=$ the Split-half analysis) or to its upper limit (for our sample size $k=33=$ the Leave-One-Out analysis). These k values reflect the two extremes for the given number of subjects (i.e., half of the subjects for $k=2$ to all but one subject for $k=33$). The effectiveness of the two procedures can then be compared on the basis of: (i) type-II errors during region selection, (ii) the proportion of variance in measured IQ change that could be accounted for by structural change, when tested on the remaining (independent) subsets, and (iii) the computational cost of both procedures.

In more detail, to implement the Leave-One-Out approach, all but one of the available observations are used in the training set and the remaining observation (that is left out) is used to validate the results (Hastie et al., 2009). The procedure is then repeated k times, with k being equal to the number of observations in the full sample, and with each observation occurring once in the test set and $k-1$ times in the training set. The advantages of this approach are that (i) power in the training set is maximised (by including all but one observation) and (ii) there are a (usually comparatively small) finite number of splits that is equal to the number of observations, see Efron and Tibshirani (1997), Hastie et al. (2009), Strother et al. (2002) for further discussion. The Leave-One-Out procedure should therefore be efficient (statistically speaking) for small sample sizes.

To implement the Split-half analysis, the full sample is split in half by randomly assigning data to two sets (*A* and *B*), so that both sets are of (approximately) equal size. In this 2-fold or Split-half cross-validation, training starts on *Set A*, with testing on *Set B*, followed by training on *Set B* and testing on *Set A*. The main disadvantage of the Split-half approach is that the ‘training sets’ (*Set A* in the first iteration, and *Set B* in the second) are smaller than they could be. Put simply, if the power per sample is low, then small training set sizes could reduce the sensitivity of detecting effects for subsequent validation in the test set. As discussed in Kohavi (1995), when the sample is small – in the context of a small k value (here $k=2$) – there is variance due to the random effects of the training sets themselves (Kohavi, 1995). One solution is to average the results after repeating the procedure with multiple two-way splits. However, if the overall sample size is too small, none of the training analyses will have sufficient power to detect effects of interest (Poldrack and Mumford, 2009). Moreover, for a reasonably sized sample, there will be an almost infinite number of possible partitions of the same data. If only a few random partitions are tested, some observations may never be selected in the validation subsample, whereas others may be selected more than once. These considerations suggest that – for a maximally sensitive analysis – the Leave-One-Out procedures may be preferable over split half procedures. In what follows, we test this conjecture quantitatively, using a reanalysis of previously reported data.