When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging

Robert Whelan and Hugh Garavan

The ability to predict outcomes from neuroimaging data has the potential to answer important clinical questions such as which depressed patients will respond to treatment, which abstinent drug users will relapse, or which patients will convert to dementia. However, many prediction analyses require methods and techniques, not typically required in neuroimaging, to accurately assess a model's predictive ability. Regression models will tend to fit to the idiosyncratic characteristics of a particular sample and consequently will perform worse on unseen data. Failure to account for this inherent optimism is especially pernicious when the number of possible predictors is high relative to the number of participants, a common scenario in psychiatric neuroimaging. We show via simulated data that models can appear predictive even when data and outcomes are random, and we note examples of optimistic prediction in the literature. We provide some recommendations for assessment of model performance.

Key Words: Addiction, imaging, machine learning, methods, prediction, simulation

"Prediction is very difficult, especially if it's about the future."
Niels Bohr

Identifying neurobiological predictors of clinically important outcomes (e.g., which young adults will transition to psychosis; which abstinent drug users will relapse) is important because they could inform mechanistic models of disease and have clinical, diagnostic utility. However, developing a regression model to predict a particular outcome for a previously unseen individual (as opposed to inferring a significant difference in between-group means) is subject to some methodologic and statistical considerations necessary to accurately assess model performance. Such considerations, although almost axiomatic in other fields (e.g., machine learning), are typically not required for neuroimaging analyses, and therefore imaging researchers may be unaware of them. Our goal is to describe how regression models can appear—incorrectly—to be predictive, and to describe methods for quantifying, and improving, model reliability and validity.

Measures of neural activity such as magnetic resonance imaging, positron emission tomography, and electroencephalography yield a potentially large number of putative predictor variables (voxels, electrodes, or regions of interest) that may also be combined with other variables such as age, sex, IQ, and so on. Thus, neuroimagers usually have many more data points relative to the number of subjects (note that the issues we describe are not restricted to neuroimaging, but apply to other domains, such as genetics (1–4). In these cases, statistical methods predicting outcomes such as group membership (e.g., logistic regression), survival models such as time to relapse (e.g., Cox regression) or regression with variable selection (e.g., stepwise regression) will result in overfitting and optimism unless particular precautions are taken. Overfitting occurs because a model derived from a particular sample will partly reflect the unique data structure of that particular sample—including the noise in the data (Figure 1). Thus, given some training data, the observed ("apparent") error will be less than the ("actual") error that is found when we then apply the model to novel test data, a difference that reflects our (unwarranted) optimism about the model (this reduction is also known as shrinkage). A challenge in generating predictive models is to minimize, and quantify, this inherent optimism.

Quantifying model performance can be achieved in a number of ways (e.g., percent correct per outcome category). However, the receiver operating characteristic (ROC) curve, which compares sensitivity versus specificity at various discrimination thresholds, is a particularly useful metric of model performance. Importantly, the ROC is not influenced by base rates, the prevalence of the disease in the population, which influences a biomarker's diagnosticity. The area under the curve (AUC; Figure 2) of the ROC quantifies the model's ability to correctly assign a patient to the disease group. A value of .5 denotes no prediction accuracy, 1 denotes perfect accuracy and heuristically, .6 to .7 can be regarded as weak, .7 to .85 as moderate, and more than .85 as good, although the convention varies considerably by discipline and analysis goal. Other measures include d′, the distance between the signal and the noise means in units of standard deviations [see Stanislaw and Todorov for more examples (5) and Bayes' rule (6)].

Crucially, and perhaps counterintuitively to those who deal primarily with the general linear model, optimism increases as a function of the decreasing number of participants and the increasing number of predictor variables in the model. (i.e., models appear better as sample size decreases). To illustrate the ease with which predictive models can apparently be created, we generated simulated data across varying numbers of observation and predictors (Figure 3). Assume we designate 25 data sets as responders (or relapsers), 25 data- sets as nonresponders, and generate 13 predictors—each randomly related to the outcome. Given these data, one observes an AUC of .80 in a logistic regression (i.e., a moderate to good performance). Similarly, assigning a random time to relapse to each member of the relapse group produces a significant Cox regression model (overall model significance of $p \leq .012$ and 5 of 13 betas significant at $p = .05$). A stepwise regression with entry value set to $p < .05$ and removal set to .1 also produces a significant model ($p = .014$, $r^2 = .166$). Of course, purely random data are unlikely in practice. Adding even a modest effect size to each predictor (e.g., a mean Cohen's d of .33) will increase the apparent AUC to .996, whereas the actual ROC is .84. Optimism in real data was described recently in a study predicting relapse in a sample of cocaine users (7). Here, the training data yielded an apparent ROC of .85, dropping to approximately .60 on test data. However,

From the Departments of Psychiatry and Psychology, University of Vermont, Burlington, Vermont.

Address correspondence to Hugh Garavan, Ph.D., Department of Psychiatry, UHC, 1 South Prospect Street, Burlington, VT 05401-1419; E-mail: Hugh.Garavan@uvm.edu.

Received Feb 26, 2013; revised May 1, 2013; accepted May 15, 2013.

0006-3223/$36.00
http://dx.doi.org/10.1016/j.biopsych.2013.05.014

many studies do not try to quantify inherent optimism (8–17), which makes it difficult for the reader to evaluate the true predictive accuracy of a particular model.

We briefly provide some recommendations for the development and assessment of regression models. An obvious solution to attenuating optimism, albeit expensive in the context of neuroimaging, is to collect more data. A minimum ratio of 10 cases per predictor is a common (18), although not a universal (19), recommendation. Optimism can be lowered by introducing a regularization term—a penalty for model complexity—to constrain the size of the parameter values. Variable selection can also be performed in combination with optimism attenuation [e.g., (20–22)], and such approaches are generally preferable to automated variable selection (e.g., stepwise regression). J-pruning (23) can be used to prune decision trees and Bayesian approaches, using previous information to constrain model complexity, are also useful (many regularization approaches can be interpreted from a Bayesian perspective).

Estimating the optimism can be achieved in a number of ways. Bootstrapping (24), or variants thereof (25), involves selecting—with replacement—the same number of data points as the original sample. This resampling is repeated many times (i.e., >1000), and the model performance for the bootstrapped samples is compared with performance for the full sample. Permutation (26) involves the random reassignment of labels (e.g., relapse or nonrelapse) to participants, and again compares the performance on the permuted data, in which the structure of the data are preserved but the outcome is random, to performance on the original data. Cross-validation tests the ability of the model to generalize and involves separating the data into subsets. A model is developed with a subset of the data (the "training" set), and then the model's predictive prowess is tested in the fully independent remainder of the data (the "test" set). At the extreme, data can be split in half, but this is wasteful. Tenfold validation (27) is efficient: a model is developed on 90% of the sample and the model's prediction accuracy is tested on the remaining 10%. This process is repeated 10 times (i.e., each fold serves as the test set once). Nested cross-validation (cross-validation within the training data) is useful to optimize parameters for some regularization techniques (e.g., the Elastic Net). If multiple models are being assessed, then unadjusted metrics of optimism become unreliable as the probability of overfitting to the test data increases with multiple comparisons (28). Recent versions of the MATLAB (The MathWorks, Natick, Massachusetts) Statistics Toolbox contain lassoglm, used to implement the methods described in (20,21,29), bootstat for bootstrap sampling, many functions for Bayesian analysis, and the bioinformatics toolbox contains crossvalind for generating training and testing sets for cross-validation. Recent versions of SPSS (30) have bootstrapping options. Future research could investigate the costs and benefits of bootstrapping, which is computationally expensive but efficient in that all the data are used, versus cross-validation for imaging data.

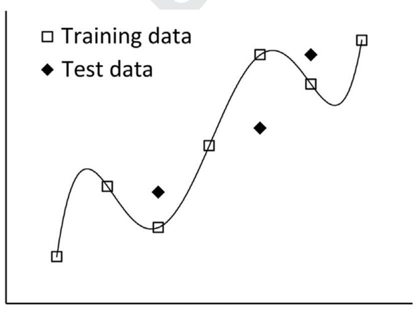

Figure 1. An example of an overfit model. The (approximately linear) relationship was modeled with a sixth-order polynomial function, which fit the training data perfectly. However, the model generalizes poorly to the test data.

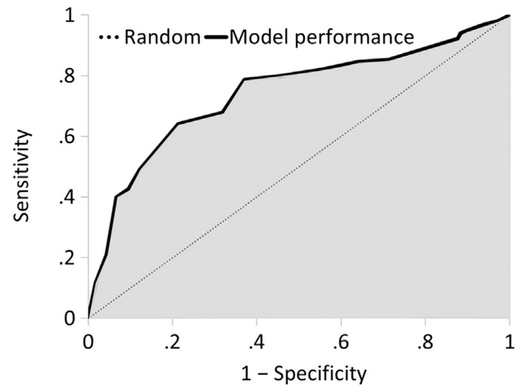

Figure 2. An example of a receiver operating characteristic curve, displaying sensitivity versus 1–specificity at various thresholds. The dashed 45° line represents random classification accuracy. The area under the solid line (shaded in gray) represents the area under the curve, a summary metric for classification performance.

One important precaution when testing the generalizability of a model is that the training and testing subsets must always be kept completely separate; any cross-contamination will result in optimism. For example, restricting analyses to regions of interest that were determined in an initial analysis that included all participants will render invalid the subsequent cross-validation. Again, simulated data can help make this point (25 participants in each group, 13 random predictors). First, we conducted a between-groups *t* test and only retained significant predictors, maintaining a strict Bonferroni cutoff (.05/13 = .0038), repeating this procedure 10,000 times to ensure an adequate sampling of false positives. Next, a 10-fold validation was conducted on any predictors that, by chance, were significant: the AUC on the "test" data was .755 (the AUC derived from the whole group was .756). Separating the training and testing subjects before the *t* test, then cross-validating, returns the expected AUC of approximately .5. We then repeated this simulation but added an effect size of .33 to each predictor. The AUC for the cross-validated data was .756 (.776 for the whole group) when, as above, the predictors were identified before separating the data into training and test sets. In contrast, doing the separation first then identifying the predictors on the training set yielded an AUC of just .601 on the test data. In essence, preselecting variables provides inaccurate information about the generalizability of a model, although it is possible to find examples of incomplete separation of data in the literature (31–34).

The use of neurobiological features to predict outcome provides us with a different perspective on neural functioning [cf. Poline and Brett (35)]. Our goal here was to highlight the need to account for the optimism that is inherent in regression models. We particularly hope that, in future, findings will be discussed with respect to the optimism-corrected results rather than the apparent error, conveying more accurately the ability of imaging data to predict and diagnose disorders.

We thank the Complex Systems group at the University of Vermont for helpful discussion during preparation of this article. The authors acknowledge the Vermont Advanced Computing Core, which is supported by the National Aeronautics and Space

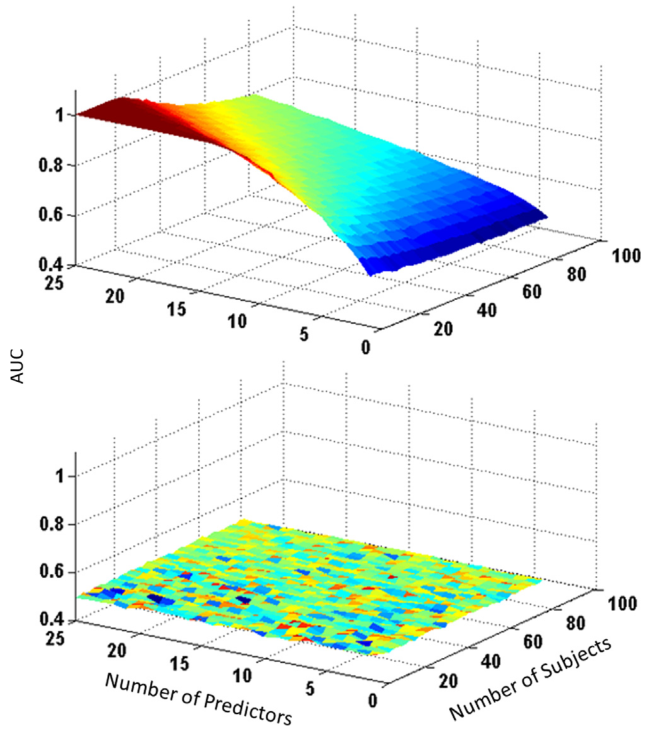

Figure 3. Normally distributed random data, half designated as treatment responders and half as nonresponders with varying numbers of predictor variables (e.g., regions of interest) and numbers of participants. A logistic regression was used to classify participants into groups (results averaged over 280 regressions). The upper panel shows apparent predictive ability increasing rapidly as the number of predictors increases and the number of participants decreases, whereas the generalization to new data, as expected, remains at chance (lower panel). AUC, area under the curve.

Administration (NNX 06AC88G), at the University of Vermont for providing high-performance computing resources that have contributed to the research results reported within this article.

The authors report no biomedical financial interests or potential conflicts of interest.

1. Ambroise C, McLachlan G (2002): Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 99:6562–6566.
2. Goddard ME, Wray NR, Verbyla K, Visscher PM (2009): Estimating effects and making predictions from genome-wide marker data. *Stat Sci* 24:517–529.
3. Evans D, Visscher P, Wray N (2009): Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 18:3525–3531.
4. Powell J, Zietsch B (2011): Predicting sensation seeking from dopamine genes: Use and misuse of genetic prediction. *Psychol Sci* 22:413–415.
5. Stanislaw H, Todorov N (1999): Calculation of signal detection theory measures. *Behav Res Methods Instruments Comput* 31:137–149.
6. Lee PM (2012): *Bayesian Statistics: An Introduction*. London: Wiley.
7. Luo X, Zhang S, Hu S, Bednarski S, Erdman E, Farr O, *et al.* (2013): Error processing and gender shared and specific neural predictors of relapse in cocaine dependence. *Brain* 1356:1231–1244.
8. Lavretsky H, Zheng L, Weiner M, Mungas D, Reed B, Kramer J, *et al.* (2010): Association of depressed mood and mortality in older adults with and without cognitive impairment in a prospective naturalistic study. *Am J Psychiatry* 167:589–597.
9. Garner B, Pariante C, Wood S, Velakoulis D, Phillips L, Soulsby B, *et al.* (2005): Pituitary volume predicts future transition to psychosis in individuals at ultra-high risk of developing psychosis. *Biol Psychiatry* 58:417–423.
10. Rando K, Hong KI, Bhagwagar Z, Li CS, Bergquist K, Guarnaccia J, *et al.* (2011): Association of frontal and posterior cortical gray matter volume with time to alcohol relapse: A prospective study. *Am J Psychiatry* 168:183–192.
11. Walterfang M, Yung A, Wood A, Reutens D, Phillips L, Wood S, *et al.* (2008): Corpus callosum shape alterations in individuals prior to the onset of psychosis. *Schizophr Res* 103:1–10.
12. Devanand D, Bansal R, Liu J, Hao X, Pradhaban G, Peterson B (2012): MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease. *Neuroimage* 60:1622–1629.
13. Tupler L, Krishnan KR, Greenberg D, Marcovina S, Payne M, MacFall J, *et al.* (2007): Predicting memory decline in normal elderly: Genetics, MRI, and cognitive reserve. *Neurobiol Aging* 28:1644–1656.
14. Zipoli V, Goretti B, Hakiki B, Siracusa G, Sorbi S, Portaccio E, *et al.* (2009): Cognitive impairment predicts conversion to multiple sclerosis in clinically isolated syndromes. *Mult Scler* 16:62–67.
15. Braverman ER, Blum K, Damle UJ, Kerner M, Dushaj K, Oscar-Berman M, *et al.* (2013): Evoked potentials and neuropsychological tests validate positron emission topography (PET) brain metabolism in cognitively impaired patients. *PloS One* 8:e55398.
16. Lin Y-T, Liu C-M, Chiu M-J, Liu C-C, Chien Y-L, Hwang T-J, *et al.* (2012): Differentiation of schizophrenia patients from healthy subjects by mismatch negativity and neuropsychological tests. *PloS One* 7:e34454.
17. Prichep L, John E, Ferris S, Rausch L, Fang Z, Cancro R, *et al.* (2006): Prediction of longitudinal cognitive decline in normal elderly with subjective complaints using electrophysiological imaging. *Neurobiol Aging* 27:471–481.
18. Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A (1996): A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 49:1373–1379.
19. Vittinghoff E, McCulloch C (2007): Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 165:710–718.
20. Zou H, Hastie T (2005): Regularization and variable selection via the elastic net. *J R Stat Soc B Stat Methodol* 67:301–320.
21. Tibshirani R (1996): Regression shrinkage and selection via the lasso. *J R Stat Soc B Stat Methodol* 58:267–288.
22. Moons K, Donders A, Steyerberg E, Harrell F (2004): Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: A clinical example. *J Clin Epidemiol* 57:1262–1270.
23. Bramer M (2002): Using J-pruning to reduce overfitting in classification trees. *Knowledge Based Syst* 15:301–308.
24. Efron B, Tibshirani RJ (1993): *An Introduction to the Bootstrap. (Vol. 57)*. New York: Chapman & Hall.
25. Efron B, Tibshirani R (1997): Improvements on cross-validation: The 632+ bootstrap method. *J Am Stat Assoc* 92:548–560.
26. Magdon-Ismail M, Mertsalov K (2010): A permutation approach to validation. *Stat Analysis Data Mining* 3:361–380.
27. Kohavi R (1995): A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2:1137–1143. Q4
28. Ng AY. Preventing overfitting of cross-validation data. Presented at the 14th International Conference on Machine Learning (ICML), 1997. Available at: http://robotics.stanford.edu/~ang/papers/cv-final.pdf. Accessed May 28, 2013.
29. Hoerl AE, Kennard RW (1970): Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.
30. IBM Corp. IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp.
31. Gomar J, Bobes-Bascaran M, Conejero-Goldberg C, Davies P, Goldberg T (2011): Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Arch Gen Psychiatry* 68:961–969.
32. Clark V, Beatty G, Anderson R, Kodituwakku P, Phillips J, Lane T, *et al.* (2012): Reduced fMRI activity predicts relapse in patients recovering from stimulant dependence [published online ahead of print September 27]. *Hum Brain Mapp.* http://dx.doi.org/10.1002/hbm.22184.
33. Devanand D, Liu X, Tabert M, Pradhaban G, Cuasay K, Bell K, *et al.* (2008): Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biol Psychiatry* 64:871–879.
34. Mechelli A, Riecher-Rössler A, Meisenzahl E, Tognin S, Wood S, Borgwardt S, *et al.* (2011): Neuroanatomical abnormalities that predate the onset of psychosis: A multicenter study. *Arch Gen Psychiatry* 68:489–495.
35. Poline J-B, Brett M (2012): The general linear model and fMRI: Does love last forever? *Neuroimage* 62:871–880.