Boosting predictions of treatment success

Michael LeBlanc[1] and Charles Kooperberg

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

For a given treatment and a patient with a specific set of attributes, what is the probability of a successful outcome after receiving treatment? This is a pervasive applied question in medicine. In PNAS, a report by Banerjee et al. (1) considers such a problem with respect to personalized prediction of success of in vitro fertilization (IVF) treatments. The majority of IVF procedures do not achieve a live birth; therefore, providing predictions of success for subsequent IVF treatments should assist a patient with decisions, given the financial, physical, and emotional costs of undergoing IVF therapy. The authors note that, at this time, estimates of the success of IVF treatment are primarily based on age-based stratification (1).

Their statistical analysis uses available information on clinical diagnoses, IVF treatment, and outcomes for IVF cycles performed at Stanford Hospital and Clinics that has been recorded in their database; cases are limited to those for which complete embryo data were available. Statistical inferences and any use of predictions based on their models will be influenced by the nature of patient selection and reasons for going on to subsequent IVF cycles. Potential unmeasured variables will likely lead to some bias in the estimates of success probabilities. The authors address this issue by conducting a sensitivity analysis with respect to availability of cycle 1 and cycle 2 IVF data, and their results appear to support their conclusions (1). We think that fully addressing the selection process is complex, and therefore, some concerns remain, especially if using the estimated model for predictions at other institutions where there may be unmeasured differences in the characteristics of patient prognostic factors and treatment delivery. Ultimately, the nature and collection of the data are paramount in the types of inferences that can be drawn in any setting. In addition, while they report 1,000× improvement in the fit of their model compared to the age based control based on the likelihood ratio, we think alternative measures of comparative predictive performance would be preferred and give less dramatic results. Our motivation for this commentary relates primarily to the statistical technologies used for the predictions, which are applicable to much more than IVF. For instance, our applied interests relate to chronic diseases and, primarily, cancer.

For this analysis, Banerjee et al. (1) use a powerful statistical methodology called boosting, which has wide applicability for prediction and other statistical modeling applications. Boosting algorithms were developed by Freund and Schapiro (2) and put into a statistical context by Friedman et al. (3). Boosting was shown to be related to a technique based on penalized regression (using an L_1 norm on the regression-model coefficients to control model complexity) called the Lasso (4). Taken as a whole, these methods and their many extensions likely constitute the most important development in statistical learning and prediction in the last two decades, and they have generated many important statistical publications. A nice review is given in ref. 5, and ref. 6 puts these methods in the broader context of statistical learning algorithms.

Although the IVF outcome is binary, live birth or no live birth, it is important that the statistical models developed actually give the estimated probability of success and not just a success or no success classification. It is clear that estimates of success probability are needed to make decisions regarding the tradeoffs to be made for undergoing therapy. Boosting and penalized logistic regression algorithms provide probabilities and therefore are more appropriate than some machine learning algorithms that classify a subject only as a case or a control. In addition, for other prediction problems, patient outcome is more complex than a simple binary event. For instance, in many applications, the goal of a new therapy is to prolong life or to prevent or lengthen time until disease recurrence, and therefore, time until event or survival-analysis methods are required.

Banerjee et al. use a powerful statistical methodology called boosting.

Boosting Algorithms

Because the manuscript does not describe the form of the boosted model and algorithm used for predicting the success for IVF treatment, we briefly outline the components here. An important variant used by the authors, gradient boosting, is an iterative algorithm in which, at each step, a predictor (or simple combination of predictors) that improves the model most is added to the model; it is also multiplied by some small weight so as not to move the solution too quickly in that direction:

$$f_m(x) = f_{m-1}(x) + \nu T_m(x).$$

Here, the term $T_m(x)$ indicates the simple single predictor or combination of predictors. The boosted model can also be represented as a simple weighted sum (an additive combination) of these simple model terms:

$$f(x) = \sum_m w_m T_m(x).$$

The individual component terms can either be individual predictor variables or other simple combinations of predictors. Banerjee et al. (1) use a regression tree (7) for each component $T_m(x)$. A regression tree is a binary decision model that yields a prediction that is constant for a set of predictor values defined by boxes in the predictor space. Therefore, a tree model for a simple term can be represented as:

$$T_m(x) = \gamma_j \text{ for } x \in R_j$$

where Rj represents a box shaped region in the predictor space.

It is a good choice for a simple model, because it is invariant to monotone transformations for variables, and it works well with categorical and continuous predictors. In addition, the sequence of binary decisions is flexible enough to find interactions. Boosting typically does not yield a simple model that one can easily record but rather, is a construction of many simple models or an ensemble of such.

Interpretation of Models

Appropriate boosted regression models have been shown to outperform simple linear logistic regression models or single tree-based models in terms of prediction-error performance. However, what is typically lost is a simple representation (or a simple formula) for the prediction model. For instance, there is no easy way for another IVF researcher to apply the prediction model described in the report to a new dataset, although that researcher had collected the appropriate predictor variables. This, of course, could be easily rectified by supplying the estimated prediction-model computer code or providing

Author contributions: M.L. and C.K. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 13570.

[1]To whom correspondence should be addressed. E-mail: mleblanc@fhcrc.org.

a web-based calculator that others could use. In addition, there are statistical tools to help interpret the variables used in the complex model; the authors (1) display variable importance calculations (8), which show the relative contribution of variables used to make predictions of IVF success.

Although boosting is a powerful prediction method, an alternative that these authors choose to explore in the SI section of their manuscript is models that yield simple interpretations. For instance, one can use regression-tree models that divide the space of predictors into regions to yield simple Boolean decision rules. One rule they gave in the SI was of the form {*Blastocyst development* $\geq 11\%$} and {*Age* ≥ 36}, which yields an estimated live-birth success probability of 38%. It is a component rule from the same class of models, regression trees, used in their boosting procedure, except that, for boosting, many regression trees are used. Because of the discreteness of the solution of a single tree, the improved interpretation of these rules will typically come at the cost of some reduced prediction performance relative to boosted regression models. Trees are not the only way to achieve Boolean rules. For example, there are other tools available. Logic regression (9) is a technique that builds up Boolean rules based primarily on categorical variables and has seen considerable usage in genetic applications; other methods, such as the patient rule induction method (PRIM) (10) and extreme regression (11), are better suited for ordered variables but also give decision-rule predictions. These latter two methods allow one to construct rules to predict average outcome for a specified fraction of patients. For instance, one can specify Boolean rules to describe the 25% of patients with the best probability of success of treatment.

Model Selection and Computation

Overfitting data is of critical concern in flexible statistical modeling. In the context of a prediction problem, it could manifest as an overly large boosted regression model that seems to give wonderful predictions on an original training sample but gives relatively poor predictions when tested on a similar but new set of data. The importance of model selection is widely appreciated, and these authors use cross-validation (where subsets of the data are repeatedly left out) and the IVF models are refitted to try to approximate what the prediction error would be on a new dataset; the model size that gives a good prediction result is chosen. That final model is used on the new dataset to evaluate the performance of the prediction model. These are critical steps in prediction-model building across many applications.

The authors also note that this boosted tree algorithm "allows many variables to be analyzed simultaneously, without need to select variables a priori" (1). In this case, with at most ~50 potential variables and a reasonable number of patient cases, this is likely true. However, does this statement scale to other applications one may want to consider? Our experience in higher-dimensional settings such as those resulting from genome-wide association studies (GWASs) with up to 1 million SNPs and high-dimension gene expression array data are that additional filtering or pre-selection can yield improved predictions. In those settings, which we believe to be of interest to some PNAS readers, even with an L_1 penalized regression (as a replacement for boosting), one likely needs to be more aggressive with respect to controlling the variance of the modeling strategy to avoid getting tricked by spurious associations. For instance, for a GWAS, there were improvements in risk prediction if the number of SNPs used was filtered based on univariate or marginal test statistics to select a small fraction of the total number of variables (SNPs) to <1,000, as reported in ref. 12. A flexible off-the-shelf algorithm such as boosting still needs to be used with some application-specific considerations.

Extensions: Models for Treatment Selection

The authors note that the prediction model could be helpful in making choices about whether to undergo subsequent IVF cycles. However, for many medical issues, there can be alternative competitive treatments available either at baseline or after a failed previous treatment. We need improved datasets (probably best from one or more randomized studies) and statistical methods that compare predictions of success between two different therapies based on patient characteristics, potentially including genetic factors, to predict for a given patient if one treatment (A) should be preferred to another treatment or (B) the reverse and by how much. This can be recognized as a generally unachieved goal of personalized medicine. These proposed statistical models would go beyond simple prediction to be prescriptive (13), because such models would aid in patient-level treatment selection. For many applications, these statistical methods would also need to allow for survival data or other longitudinal data. It is interesting to note the strong connection to the active research area of identification of gene × environment interactions and modeling, especially in the context of genomics and GWASs (14). There are commonalities in the strategies for these areas, and both problems address statistical interactions or varying effects (treatment or environment) in subgroups of subjects. There is room for improved statistical methods and algorithms, potentially, for example, by developing specialized extensions of the regression tree and boosting algorithms that Banerjee et al. (1) applied to the IVF prediction problem.

ACKNOWLEDGMENTS. We thank Mark Blitzer for his assistance in manuscript preparation. Work related to statistical methodology and prediction methods for M.L. and C.K. were supported through National Institutes of Health Grants R01 CA90998 and P01 CA53996.

1. Banerjee P, et al. (2010) Deep phenotyping to predict live birth outcomes in in vitro fertilization. *Proc Natl Acad Sci USA* 107:13570–13575.
2. Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning* (Morgan Kaufman, San Francisco), pp 148–156.
3. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. *Ann Stat* 28:337–407.
4. Hastie T, Tibshirani R, Friedman J (2009) *Elements of Statistical Learning* (Springer, New York), 2nd Ed.
5. Buhlmann P, Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Stat Sci* 98:477–515.
6. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B* 58:267–288.
7. Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees* (Wadsworth Advanced Books and Software, Belmont, CA).
8. Friedman JH (2001) Greed function approximation: A gradient boosting machine. *Ann Stat* 29:1189–1232.
9. Ruczinski I, Kooperberg C, LeBlanc M (2003) Logic regression. *J Comput Graph Statist* 12:475–511.
10. Friedman J, Fisher N (1999) Bump hunting in high dimensional data. *Stat Comput* 9:123–143.
11. LeBlanc M, Moon J, Kooperberg C (2006) Extreme regression. *Biostatistics* 7:71–84.
12. Kooperberg C, LeBlanc M, Obenchain V (2010) Risk prediction models for genome-wide association studies. *Genet Epidemiol*, in press.
13. Gunter L, Zhua J, Murphy SA (2010) Variable selection for qualitative interactions. *Stat Methodol*, in press.
14. Kooperberg C et al. (2009) Structures and Assumptions: Strategies to Harness Gene × Gene and Gene × Environment Interactions in GWAS. *Stat Sci* 24:472–488.