The prediction of intelligence in preschool children using alternative models to regression

W. Holmes Finch · Mei Chang · Andrew S. Davis ·
Jocelyn E. Holden · Barbara A. Rothlisberg ·
David E. McIntosh

Published online: 7 May 2011
© Psychonomic Society, Inc. 2011

Abstract Statistical prediction of an outcome variable using multiple independent variables is a common practice in the social and behavioral sciences. For example, neuropsychologists are sometimes called upon to provide predictions of preinjury cognitive functioning for individuals who have suffered a traumatic brain injury. Typically, these predictions are made using standard multiple linear regression models with several demographic variables (e.g., gender, ethnicity, education level) as predictors. Prior research has shown conflicting evidence regarding the ability of such models to provide accurate predictions of outcome variables such as full-scale intelligence (FSIQ) test scores. The present study had two goals: (1) to demonstrate the utility of a set of alternative prediction methods that have been applied extensively in the natural sciences and business but have not been frequently explored in the social sciences and (2) to develop models that can be used to predict premorbid cognitive functioning in preschool children. Predictions of Stanford–Binet 5 FSIQ scores for preschool-aged children is used to compare the performance of a multiple regression model with several of these alternative methods. Results demonstrate that classification and regression trees provided more accurate predictions of FSIQ scores than does the more traditional regression approach. Implications of these results are discussed.

Keywords Prediction · Regression · *CART* · *MARS* · *Neural networks* · *GAM* · *Boosting*

Clinical neuropsychologists are frequently required to determine whether an individual has experienced changes in intellectual functioning resulting from a neurological insult, such as a traumatic brain injury (TBI) or stroke. Accurate diagnosis and the determination of functional decline rely largely on a clinician's ability to compare current test performance with an estimate of premorbid (i.e., prior to injury) performance. It is common in clinical practice for neuropsychologists to use a discrepancy between a predicted and an obtained test score to assist in the determination of whether organic impairment or a progressive disease is present. Thus, an accurate estimation of premorbid intelligence is necessary to prevent errors such as under- or overestimation of a patient's level of cognitive decline (Griffin, Mindt, Rankin, Ritchie, & Scott, 2002), and the availability of techniques demonstrating good validity and reliability for predicting premorbid intellectual functioning is a central concern of clinicians. When premorbid ability levels can be reasonably estimated, a diagnosis can be made with confidence, and cognitive rehabilitation programs can be properly designed, monitored, and modified (Reynolds, 1997).

The demand for accurate, reliable, and objective estimation procedures is even greater in forensic neuropsychology and neuropsychiatry, where claims of loss of intellectual function often imply involvement of potentially compensable matters (Lezak, Howieson, Loring, Hannay, & Fischer, 2004). For example, it is common for neuropsychologists to be employed as expert witnesses in lawsuits involving automobile accidents to determine whether functional decline has occurred as the result of the accident. Moreover, neurodegenerative disorders (e.g., dementia) have prompted the development of proper estimation procedures for the early detection and accurate diagnosis of various degenerative stages, as well as for describing proper treatment

W. H. Finch (✉) · M. Chang · A. S. Davis · J. E. Holden ·
B. A. Rothlisberg · D. E. McIntosh
Department of Educational Psychology, Ball State University,
Muncie, IN 47306, USA
e-mail: whfinch@bsu.edu

methods (Pavlik, Doody, Massman, & Chan, 2006). For example, the early detection of Alzheimer's disease is facilitated by comparing current norm-referenced memory performance against a measure of estimated premorbid functioning; this has increased in importance, given the new medications that have been shown to slow disease for this condition (Almkvist & Tallberg, 2009).

Traditional methods of prediction

A variety of approaches have been proposed and developed for the estimation of premorbid ability estimation, including (1) historical achievement-based and standardized group assessment data (e.g., Baade & Schoenberg, 2004; Schinka & Vanderploeg, 2000); (2) "hold/don't-hold test" estimates (Blair & Spreen, 1989; Lezak et al., 2004); (3) best current performance estimates (Lezak, 1995; 2004); (4) demographic-based regression formulas (e.g., Barona, Reynolds, & Chastain, 1984); (5) combinations of demographic and actual performance data (e.g., Schoenberg, Lange, Brickell, & Saklofske, 2007b; Schoenberg, Lange, & Saklofske, 2007a; Schoenberg, Lange, Saklofske, Suarez, & Brickell, 2008a; Schoenberg, Lange, Saklofske, Suarez, & Brickell 2008b; Schoenberg, Scott, Duff, & Adams, 2002; Vanderploeg, Schinka, & Axelrod, 1996); and (6) current word reading ability tests (e.g., Blair & Spreen, 1989; Wechsler, 2003). However, each approach has been shown to have some limitations in application. Limitations and detail of these approaches will be discussed further in the following paragraphs.

Historical data, such as grades, group administered assessment tests, SATs, and GREs, would provide good estimates of premorbid functioning, but in most clinical situations, these data usually are not readily available (Schoenberg, Lange, Brickell, & Saklofske, 2007a). When direct measurements of ability made prior to the injury are not available, one approach for estimating premorbid level of intellectual status involves a cognitive ability that is thought to remain unchanged ("holds") following brain injury (Lezak et al., 2004). This approach is commonly used with the Wechsler scales (e.g., WISC-IV, Wechsler, 2003; WAIS-III, Wechsler, 1997), in that performance on its subtests of Vocabulary, Information, Picture Completion, and/or Matrix Reasoning are generally resistant to the effects of brain insult, while scores on other subtests—primarily, fluid reasoning, attention, or memory-based decline—"don't hold" (Lezak et al., 2004; Schoenberg, Lange, Saklofske, Suarez, & Brickell, 2008b). Although this method is typically adopted by clinicians, it is not recommended for use in clinical practice, because this approach is problematic for inconsistent research findings and compromising variables such as development pace in

children, aging effects, lesion location, and type of neurologic insult (Reynolds, 1997). In addition, other methods utilizing the hold/don't-hold tests were criticized, since it is unclear to what the extent these "hold" cognitive abilities are resistant to the effects of brain dysfunctions (Lezak, 1995; Lezak, et al, 2004).

The best performance estimate (BPE) method uses the best or highest scores within an instrument battery to estimate prior functional level on the basis of the assumption that the highest levels of performance demonstrated following brain insult are the ones most likely to remain intact (Schoenberg, Duff, Scott, & Adams, 2003). Studies have shown that premorbid ability estimates using the BPE method tend to underestimate overall IQ for certain age groups (e.g., Schoenberg, Duff, Scott, Patton, & Adams, 2006). In addition, current word-reading tests are typically used to estimate premorbid cognitive functioning (e.g., Green et al., 2008; Mathias, Bowden, Bigler & Rosenfeld 2007), such as the widely accepted National Adult Reading Test (Nelson & O'Connell, 1978) and the North American Adult Reading Test (Blair & Spreen, 1989), as well as the more recently developed Wechsler Test of Adult Reading (Wechsler, 2003). The BPE approach is based on the assumption that reading pronunciation skill is highly correlated with intelligence in the general population and is relatively resistant to neurological damage (Willshire, Kinsella & Prior 1991). Despite the popular use of the word-reading tests, research has shown inconsistent findings in its use for accurately predicting premorbid IQ (Crawford, Allan, Cochrane & Parker 1990; Green et al., 2008; Mathias et al., 2007).

The demographic-based regression method for estimating premorbid ability involves developing a multiple linear regression (MLR) model by employing only demographic-based variables (e.g., age, gender, education, ethnicity, and occupation) that are easily obtainable and potentially avoid the effects of current ability levels or neurological injury (Griffin et al., 2002). The approach that combines the use of demographic variables and current actual performance also utilizes MLR to predict scores on an IQ measure. Such an approach is characterized by adding a measure of current ability to demographic information designed to tailor the regression equation to account for more variance in IQ than would be accounted for by equations based on demographic information alone (Yeates & Taylor, 1997). Estimation models employing demographic variables that are objective require no subjective judgments by the clinician and, thus, should provide a more reliable premorbid estimate than several of the other methods previously discussed. For this reason, demographic-based regression equations have become very popular for the prediction of premorbid intelligence. Nonetheless, clinical judgment is still required with respect to which current best performance outcome to use to

create a formula suitable for individuals sustaining different types of neurological injuries (Vanderploeg et al., 1996).

Using MLR to predict premorbid IQ

As was previously discussed, in applications using MLR to estimate IQ, some researchers have developed models based on demographic variables in conjunction with performance on a task such as word reading or some comparable measure (Sellers, Burns, & Guyrke, 2002; Vanderploeg, Schinka, Baum, Tremont, & Mittenberg, 1998; Yeates & Taylor, 1997), while in other cases, only demographic variables have been used. Crawford, Millar, and Milne (2001) found that for adults, the correlation between actual and predicted IQ, based on the demographic variables of education, socioeconomic status, and age, was .76, which was higher than that obtained through clinical judgment. A study focusing on predicting IQ for adolescents included variables such as gender, ethnicity, region of the U.S. in which the subject lived, age, and parental education level (Schoenberg, Lange, Brickell, & Saklofske, 2007a) and was found to provide predictions of full-scale IQ (FSIQ). Powell, Brossart, and Reynolds (2003) compared the performance of two regression models of the demographic information estimation formula index (DI) (Barona et al., 1984) and the Oklahoma Premorbid Intelligence Estimate (OPIE) (Krull, Sherer, & Adams, 1995) for estimating premorbid cognitive functioning in adults. Both models are based on linear equations that predict cognitive functioning using demographic variables (age, gender, race, education, occupation, urban/rural residence, and current performance) on the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1983) Vocabulary and Picture Completion subtests. Their results demonstrated that the DI approach provided more accurate estimates of cognitive decline but was not as accurate when predicting FSIQ for individuals who had not suffered any brain injury. One issue with either of these approaches is that researchers must have access to all of the variables that serve as inputs to the standardized equations. Another issue is that regression-based estimates of premorbid IQ have been shown to be susceptible to error, particularly in outer ranges of intellectual function (Veiel & Koopman, 2001).

Much of the focus in the prediction of premorbid IQ has been on adults, with relatively little research devoted to predicting cognitive functioning in school-aged and younger children (Schoenberg, Lange, & Saklofske, 2007b). However, some research has been conducted on the use of prediction equations based only on demographic variables with school-aged children. For example, Sellers, Burns, and Guyrke (1996) examined the Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI–R; Wechsler, 1989) standardization sample for children and found that parental level

of education, occupation, geographic region, and ethnicity were predictive of FSIQ, with an R^2 of .4, and that parental education was the most powerful predictor. Other studies also have shown correlations between a child's intelligence, family literacy environment, ethnicity, maternal education, and achievement outcomes (Pungello, Iruka, Dotterer, Koonce-Mills, & Reznick, 2009; Roberts, Bornstein, Slater, & Barrett, 1999). Other studies have focused on predicting IQ for adolescents and children, including variables such as age, gender, ethnicity, parental education level, and/or region of the U.S. in which the subject lived (Schoenberg, Lange, Brickell, & Saklofske, 2007a, 2008a), and have produced mixed results. For example, Schoenberg et al. (2008b), using a paired samples *t*-test, found that predicted FSIQ scores were, on average, significantly greater than obtained FSIQ scores for subjects who had sustained a TBI, whereas estimated and obtained FSIQ did not differ when applied to healthy peers.

Despite the strong empirical evidence for positive correlations between demographic variables and children's intellectual functioning and outcomes of performance, limited research has been conducted using demographic variables in conjunction with MLR to estimate premorbid IQ in school-aged and younger children, and what has been published is somewhat inconclusive (Sellers et al., 1996). Indeed, the problem of estimating premorbid IQ in school-aged children has received far less attention than have similar models for adults (Schoenberg, Lange, & Saklofske, 2007b). Furthermore, prior work with adult populations has not definitively demonstrated linear models to be the universally the most effective tools for predicting premorbid IQ scores, as was discussed previously. In addition, some of the variables used in the adult models (e.g., level of education) are not applicable to a very young population. For these reasons, alternative methods for prediction should be investigated in order to find the optimal tool(s) for the important task of obtaining reliable estimates of premorbid intellectual functioning for young children who have undergone a neurological insult and suffered from cognitive impairment.

It should be noted that while the focus of this research was on predicting IQ scores for young children using demographic variables, other recent examples using prediction in the social science literature include predicting school counselor evaluations of student performance (Granello, 2010), college student dropout (Nistor & Neubauer, 2010), impact of character education on social competence (Cheung, & Lee, 2010), and parent training effectiveness (Lavigne, LeBailly, & Gouze, 2010), to name but a few. In the vast majority of this research, some variant of linear regression was used to obtain predictions. However, because it is limited to linear or relatively simple nonlinear forms, regression may not always be the optimal choice for this type of research (Berk, 2008).

The goals of this study were to describe some alternative methods of prediction that could be employed in the context of obtaining estimates of premorbid IQ in preschool-aged children and to compare their performance with the more commonly used MLR approach based on a set of commonly used demographic variables. These predictor variables included age, gender, ethnicity, maternal education, and paternal education, while the FSIQ, as measured by the Stanford–Binet Intelligence Scales–Fifth Edition (Roid, 2003), was the outcome of interest. Prior to discussing the study methodology and results, we will describe several alternative methods of prediction that can be used in this context.

Alternative methods of prediction

Following is a discussion of a number of alternative approaches for prediction that may prove more useful than the standard MLR approach when relationships between variables are not linear and when standard assumptions regarding the data are not met. Given the psychometric problems, discussed earlier, with the traditional approaches of using regression models for predicting premorbid IQ, and because very little work has been done examining the prediction of IQ in very young children, these alternative methods may prove to be interesting tools for this task. After a description of the various methods, the results of a study comparing these techniques for the task of predicting IQ will be discussed.

Classification and regression trees (CART) CART (Breiman, Friedman, Olshen, & Stone, 1984) arrives at predicted values for an outcome variable, Y , given a set of predictors, by iteratively dividing individual members of the sample into ever more homogeneous groups, or nodes, based on values of the predictor variables. It can be thought of as a nonparametric approach, in that there are no assumptions regarding the underlying population from which the sample is drawn or the form of the model linking the outcome and predictor variables. CART begins by placing all subjects into one node, or group, and then searches the set of predictors to find the value of one of those by which it can divide the observations into two new nodes, whose values on Y are as homogeneous as possible. For each of these new nodes, the predictors are once again searched for the optimal split by which the subjects can be further divided into ever more homogeneous nodes, where homogeneity is always based on the similarity of values of Y . This division of the data continues until a predetermined stopping point is reached, when further splits do not appreciably reduce the heterogeneity of the resulting nodes. At this point, the tree is complete, and values of Y for new individuals can be

obtained using the decision tree developed with this original training sample. The data for the new subject are fed into the tree, following the branches from node to node on the basis of the values of the predictor variables, until the individual is placed in one of the final, or terminal, nodes. The predicted value of Y for each individual is then the mean for the training sample in this terminal node.

Neural networks (NNET) Another prediction method examined in this study is neural networks (NNETs; e.g., Marshall & English, 2000; see Garson, 1998, for a more technical description of the method). NNETs identify relationships between Y and a set of predictor variables by using a search algorithm that examines a large number of subsets of the predictors. Interactions between and powers of the predictors (referred to as *hidden layers*) are computed, in conjunction with weights, akin to regression slopes, for each term. These hidden layers are then selected by the algorithm so as to minimize the common least squares criterion used in standard linear regression (i.e., minimizing the sum of squared differences between the observed and predicted values). The hidden layers are generally much more complex than the two- and three-way interactions common in regression, involving several predictors and higher order versions of the predictors in a single interaction (Schumacher, Robner, & Vach, 1996). In order to reduce the likelihood of finding locally optimal results that will not generalize beyond the original (training) sample, random changes to the subset of predictors and interactions, not based on model fit, are also made. This method of obtaining optimal model fit is known as *back-propagation*, where the difference between actual and predicted outputs is used to find optimal weight values. It is one of the most commonly used approaches in NNET applications (Garson, 1998).

A primary strength of NNET models is that they can identify complex interactions among the predictor variables in the hidden layer that other approaches may ignore (Marshall & English, 2000). For example, whereas, in regression, it is common to express the interaction of two predictors as their product, or to square or cube a single variable if the relationship with the response is believed not to be linear, a NNET will create hidden layers as weighted products of, perhaps, several variables, thus allowing the model to be influenced by the predictors in varying degrees. The result is that fairly obscure relationships among the variables will be automatically identified and used in the prediction of the outcome variable, without the researcher having to explicitly request their presence in the model. Conversely, this ability to identify extremely specific models to fit the data presents a potential problem, in that NNETs can substantially overfit the training data used to estimate the model (Schumacher et al., 1996). In other

words, the weights selected for each variable and each hidden layer may be so closely linked to the training sample that the results are not generalizable to the wider population. Thus, predictions for new cases will be much less accurate than those for the original sample. In order to combat this problem, most NNET models apply what is called *weight decay*, which penalizes (i.e., reduces) the largest weights found in the original NNET analysis—in effect, assuming that very large weights are at least partially driven by random variation unique to the observed data.

Multivariate adaptive regression splines (MARS) The MARS model is an extension of linear regression in which nonlinear relationships and interactions are modeled automatically, similar in spirit, if not method, to NNETs. The individual terms in the model are known as *basis functions* (or *hinge functions*) and are piecewise linear, changing direction at a point t on the line, which is known as the *knot* (Hastie, Tibshirani, & Friedman, 2001). For a given basis function, the model coefficients are estimated through ordinary least squares, minimizing the residual sum of squares in much the same way as standard regression.

MARS begins building a predictive model by using a forward stepwise methodology, in which the first step involves the inclusion only of β_0 . The algorithm then proceeds to add a basis function to the model in each step, selecting the one that provides the greatest reduction in the sum of squared residuals. The newly added basis function will include a term already in the model, multiplied by a new hinge function. When deciding which new basis function to add to the existing model, MARS searches across each possible value for all terms (main effects and interactions) currently in the model, as well as all of the independent variables not yet included in the analysis, in order to select the knot for the new basis function. This model building continues until the change in the least squares criterion becomes very small when a new term is entered or until a maximum model size (set by the user) is reached.

As with NNET, a given MARS model may overfit the training dataset. Therefore, a stepwise backward deletion procedure is included, in which the least important (in a statistical sense) term (main effect or interaction) is removed at each step. The optimal model is selected by minimizing a function of the sum of squared residuals penalized by the number of parameters in the model and the sample size. Thus, the MARS model attempts to identify the simplest model (the one with the fewest terms) that maximizes prediction accuracy. The primary advantage of the MARS model building strategy is its ability to work well locally in the function space and, thus, identify precisely where the relationship between a given predictor and Y changes direction (Hastie et al., 2001). Specifically, the use of the basis functions described above allows for the

modeling of interactions only in the range of data for which two such functions have a nonzero value. Thus, unlike other models described here, the entire data space is not required to take a common nonlinear functional form. This means that rather than assessing whether a particular nonlinear term is significant for the entire dataset, MARS is able to identify the nonlinearity within a small span of the data.

Generalized additive models (GAMs) GAMs are a class of very flexible models that allow for the linking of Y with one or more predictor variables, using a wide variety of smoothing functions common in statistics. Each function is fit using a smoothing technique, such as a cubic spline or a Kernel smoother, with the goal of minimizing the penalized sum-of-squares criterion. The penalized sum of squares (PSS) is based on the standard sum of squared residuals (i.e., the difference between the actual and predicted values of the response variable squared), with a penalty applied for model complexity (i.e., the number of main effects and interactions included).

The GAM algorithm works in an iterative fashion, beginning with the setting of β_0 to the mean of Y . Subsequently, a smoothing function is applied to each of the independent variables in turn, selecting the smoothed predictor that minimizes the PSS. This iterative process continues until the smoothing functions stabilize (i.e., the PSS cannot be appreciably reduced further), at which point final model parameter estimates are obtained. The most common smoothing function used with GAMs (and the one used in this study) is the cubic spline. As is the case for several of the methods described in this article, overfitting of the data can be a problem with GAMs. Therefore, it is recommended that the number of smoothing parameters be kept relatively small (Wood, 2006).

Boosting (BOOST) Boosting refers to a prediction algorithm that attempts to improve prediction accuracy of Y by combining the predictions from a set of weak predictor variables in order to obtain a single strong prediction (Freund & Schapire, 1997). Boosting was originally developed for use in predicting a dichotomous outcome variable but has since been generalized to the regression context with continuous responses (see Buhlmann & Hothorn, 2007, for a discussion of this history). Of these extensions, one of the more popular is L_2 boosting. As described by Buhlmann and Hothorn, this algorithm consists of five steps:

1. A regression equation based on the set of predictors is fit to the original response variable.
2. The residuals (observed Y predicted Y) for this model are calculated.

3. The original set of predictors is used to predict the residuals obtained in step 2.
 4. The fitted residual values obtained in step 3 are used to update the fitted value of the outcome variable, Y , obtained in step 1; that is, they are included as independent variables.
5. Steps 1 through 4 are repeated until the change in the fitted value of Y is below some threshold value, in which case convergence to a solution has been reached.

A variety of approaches for determining when to stop the boosting algorithm have been recommended, with perhaps the most recent (Buhlmann & Hothorn, 2007) being the use of the minimum value for Akaike's information criterion (AIC). Therefore, a researcher may elect to use a large number of m iterations and then review the resultant AIC values, selecting the model that corresponds with the smallest of these, which was the approach used in the present study.

L_2 boosting will typically lead to very complex models as the number of iterations (and thus, residual functions) increases. For this reason, it is recommended that the original regression model used to predict Y be fairly simple, consisting of a relatively small number of predictors (Buhlmann & Yu, 2003). Finally, it should be noted that while linear regression is quite often the model to which the boosting algorithm is applied, it is entirely possible to use smoothing splines or other functions to relate the response variable to the predictors and then apply the boosting algorithm (Buhlmann & Hothorn, 2007).

Correlation weights (CORR) In contrast to the complex methods described heretofore, an alternative approach to deriving a prediction equation for some outcome variable that has been suggested is the use of weights based on the zero-order correlation coefficients between the individual standardized predictors and the standardized Y (Dana & Dawes, 2004). Using this technique, the model for predicting Y with three predictors, x_1 , x_2 , and x_3 , would be expressed as (1).

Dana and Dawes note that both Marks (1966) and Goldberg (1972) reported positive results when using these correlations, as opposed to the standard regression coefficients. The potential advantage of the CORR prediction method is its simplicity, particularly in comparison with the other, more complex prediction models described previously.

The present study

This study has two purposes. First of all, it serves as an introduction and description of prediction tools that are

increasingly in use in fields such as medicine and business but have not made their way into the broader social science literature. They represent potential alternatives to the familiar MLR approach to prediction and may offer some advantages either in the form of simplicity of use (CORR) or in their ability to incorporate complex relationships between the predictor and outcome variables (CART, MARS, GAM, NNET, BOOST). More specifically, a number of these methods, including CART, MARS, GAM, and NNET, allow for the automatic investigation of nonlinear model forms in the development of prediction models. In contrast, the MLR approach to prediction includes only nonlinear terms in the model if the user explicitly specifies them. Therefore, these alternative methodologies may prove particularly useful when researchers want to model a complex phenomenon such as cognitive functioning, using a variety of variables when the likely model form (linear or nonlinear) is not well understood. Second, this study represents an initial step in the development of tools to predict premorbid IQ in preschool children, using easily obtained demographic variables such as age, gender, ethnicity, maternal education, and paternal education. We hypothesize that the alternative prediction methods, particularly those accommodating nonlinear terms such as CART, NNET, MARS, and GAM, will provide more accurate predictions of cognitive functioning than will the traditional MLR approach.

Methodology

Subjects The subjects for this study included 200 (103 females, 97 males) preschool children. The sample was obtained from preschool facilities near a mid-sized city in the Midwest during the 2008–2009 school year. They were solicited through the preschoolers' teachers, who sent subject requests/consent forms home with each child. Only children whose parental permission/signed consent was obtained were included in this study. Data on parents declining to participate were not recorded, since their consent forms were not returned; therefore, data on differences among those who participated and those who declined were not available. Demographic information for the total sample appears in Table 1. Only children who were not receiving special education or related services and whose parental consent was obtained were included as subjects.

Procedure Once a signed parental permission form was obtained, the children were administered the Stanford–Binet Intelligence Scales–Fifth Edition (SB5; Roid, 2003) under standardized conditions by trained examiners. Test administration for each child occurred over the course of three sessions, and all data collection was completed within

Table 1 Descriptive statistics for the total, training and cross-validation samples

Variable	Total Sample	Training	Cross-Validation
Gender			
Male	97 (48.5%)	73 (48.7%)	24 (48%)
Female	103 (51.5%)	77 (51.3%)	26 (52%)
Ethnicity			
Caucasian	124 (62%)	93 (62%)	31 (62%)
African-American	49 (24.5%)	38 (25.3%)	11 (22%)
Hispanic/Latino	2 (1%)	2 (1.3%)	0 (0%)
Biracial	20 (10%)	15 (10%)	5 (10%)
Other	3 (1.5%)	0 (0%)	3 (6%)
No report	2 (1%)	2 (1.3%)	0 (0%)
Father's education			
Less than high school	30 (15%)	24 (16%)	6 (12%)
High school/GED	78 (39%)	56 (37.3%)	22 (44%)
1–3 years of college	44 (22%)	35 (23.3%)	9 (18%)
4+ years of college	29 (14.5%)	21 (14%)	8 (16%)
No report	19 (9.5%)	14 (9.4%)	5 (10%)
Mother's education			
Less than high school	16 (8%)	13 (8.7%)	3 (6%)
High school/GED	48 (24%)	35 (23.3%)	13 (26%)
1–3 years of college	89 (49.5%)	65 (43.3%)	24 (48%)
4+ years of college	39 (19.5%)	31 (20.7%)	8 (16%)
No report	8 (4%)	6 (4%)	2 (4%)
Age in months: Mean (<i>SD</i>)	58.86 (5.38)	59.73 (5.50)	60.28 (5.04)
FSIQ	98.10 (11.81)	98.29 (11.17)	97.54 (13.67)

a 2-week period in order to minimize changes related to development. After administration, protocols were scored by the individual examiners, using the computer-scoring program for the SB5. All testing was conducted by six advanced graduate students in a school psychology program, all of whom had received graduate-level training in administering standardized cognitive assessment. Each child was tested one-to-one in a private room or area according to the standardized testing procedures outlined in the examiner's manual (Roid, 2003). Each protocol was subsequently double-checked for errors by an advanced school psychology doctoral student. The data were entered into an SPSS statistical program on a regular basis as it was gathered. "Examiner drift" was not specifically assessed, given that the examiners were trained to emphasize the importance and necessity of following standardized testing procedures in administering the tests to all the children assessed. However, the supervising researchers and the advanced doctoral student did regularly examine the protocols and observed testing by each examiner at regular intervals.

Instrumentation The SB5(Roid, 2003) is an individually administered assessment of IQ appropriate for people between the ages of 2 and 85 years. It is theoretically grounded

in the Cattell–Horn–Cattell (CHC) theory and is meant to represent five CHC factors, including fluid intelligence (*Gf*), crystallized knowledge (*Gc*), quantitative knowledge (*Gq*), visual processing (*Gv*), and short-term memory (*Gsm*). The entire SB5 (five verbal and five nonverbal subtests) was administered to the subjects and generated an FSIQ. In relation to this study, the SB5 FSIQ score was used to indicate the children's comprehensive cognitive abilities. The SB5 was selected for use in this study because it is strongly grounded in CHC theory, has been normed for children as young as those used in this study, and has been shown to be a valid and reliable tool for such assessments.

Prediction models The outcome variable of interest was the FSIQ from the SB5, while the predictors included gender, years of education each for mother and father, ethnicity (Caucasian or minority), and age. These predictors were selected because they are typically available for any subject for whom predicted IQ is required and will not be impacted by a CNS injury. They have also been used in prior IQ prediction studies (Sellers, Burns, & Guyrke, 1996). The models used to predict FSIQ with these demographic variables included MLR, as well as CART, NNET (with 2, 5, and 15 hidden layers), MARS, GAM, BOOST, and CORR. All analyses were carried out using the R software

package (R Development Core Team, 2007). Where applicable the values of tuning parameters were based on recommendations in the literature.

In order to assess the predictive accuracy of the models, the original sample of 200 subjects was randomly divided into training ($n = 150$) and cross-validation ($n = 50$) samples. For each method, the training sample was used to estimate a predictive model, which was, in turn, applied to the cross-validation sample to obtain predicted values for FSIQ. Prediction accuracy for the cross-validated sample was assessed through the bias of the predicted IQ, $\theta_{\text{Predicted}} - \theta_{\text{Actual}}$, the root mean square error (RMSE) of the predictions for the cross-validation

sample, $\sqrt{\frac{\sum (\theta_{\text{Predicted}} - \theta_{\text{Actual}})^2}{n}}$, and R^2 between the predicted

and actual FSIQ values. Bias serves as a measure of the estimation accuracy, RMSE reflects both accuracy and precision of the predicted values, and R^2 is the proportion of variance shared by the predicted and actual FSIQ scores. In general, results with lower bias, lower RMSE, and higher R^2 can be viewed as better.

Results

Table 1 contains descriptive statistics for the training and cross-validation samples, respectively. A t -test was used to compare the mean ages and FSIQ scores between the training and cross-validation samples, and no significant differences were found. In addition, Exact tests were used to compare the frequency distributions for gender, father's education, mother's education, and ethnicity between the two samples, and no significant differences were found. Please note that Bonferroni's correction was used to control the family-wise Type I error rate.

Table 2 includes the mean bias, RMSE, and R^2 results for each prediction method studied here. Note that for NNET, only results for the best-fitting model (15 hidden layers) are presented here. These results suggest that MLR, GAM, and CART produced the least biased predictions of FSIQ for the cross-validation sample, all with values extremely close to 0. This low level of bias suggests that the models developed using the training sample will produce predicted values that are, on average, very close to the actual FSIQ for the cross-validation sample. In contrast, NNET and MARS produced the most biased results of the methods studied here, with the predicted values being higher, on average, than the actual FSIQ scores. BOOST and CORR had bias results in between those of the best and worst performers, and like MARS and NNET, their predictions tended to overestimate the actual FSIQ.

Table 2 Mean prediction bias RMSE and R^2 values for the cross-validation sample

Model	Mean Bias*	RMSE	R^2
MLR	0.0000004	12.54	0.22
BOOST	0.29	14.03	0.09
CORR	0.28	10.09	0.08
NNET	0.61	15.06	0.14
MARS	0.62	15.68	0.01
GAM	0.0000004	12.52	0.25
CART	0.0000003	11.93	0.31

*Mean bias = mean difference between actual and predicted FSIQ scores. Larger values indicate less accurate predictions

RMSE = root mean square error. Larger values indicate less accurate predictions

R^2 = proportion of variance in actual FSIQ scores accounted for by predicted FSIQ scores. Higher values indicate more accurate predictions

As was mentioned above, the RMSE reflects both the degree of bias in the predicted values and their variation. Thus, higher values indicate less stable estimates. In this case, CORR displayed the lowest value of RMSE, followed by CART, while the highest RMSE values belonged to MARS, NNET, and BOOST. These results suggest that of the methods included here, CART and CORR had the lowest combination of bias and variation in the estimates. With respect to the degree of variation shared between the predicted and actual values, CART had the highest value, followed by GAM and MLR, both of which performed very similarly. In contrast, CORR, BOOST, and MARS shared the least amount of variability with the actual FSIQ values, indicating that the quality of their predictions was somewhat less than that of the other approaches.

Discussion

Prediction is an important aspect of statistical practice in psychology and the other social sciences. Prior studies in the area of premorbid IQ prediction have generally focused on using MLR-based models with adolescent and older populations. However, it has been argued that the regression-based approach may not always be optimal (Veiel & Koopman, 2001), nor has it been shown that such predictions can be accurately made for preschool age children. The goal of this study was to demonstrate how modern methods of prediction that are currently in use in fields as diverse as business, medicine, and national defense can also be applied to problems in the social sciences. These methods are potentially quite useful in behavioral research, either because they are very simple to carry out (CORR) or because they allow for the modeling of complex

relationships between the predictor and outcome variables (CART, GAM, MARS, NNET, BOOST).

To briefly summarize the results of this study, it appears that in terms of the outcome variables included here, including bias, RMSE and R^2 , CART provided the most accurate predictions of FSIQ. It demonstrated the least bias, the lowest RMSE value, and the highest proportion of explained FSIQ variance in the cross-validation sample. Both MLR and GAM produced the second most accurate sets of results, performing very similarly to one another and better than the other approaches studied here, in terms of the outcome variables. Among the other methods, no one could be said to perform better than the rest, given mixed findings with regard to bias, RMSE, and R^2 . Thus, a researcher interested in predicting IQ scores in young children may find that CART provides the most accurate results of the methods studied here. One primary reason for this improved accuracy could be the flexibility with which CART deals with a mixture of categorical and continuous predictor variables, something that has proven more difficult for some of the other methods in previous research—particularly, MARS and NNET (Berk, 2008). In addition, given that prior Monte Carlo simulation work has shown that linear models such as MLR perform poorly for prediction when there are a number of interactions among predictor variables in the population (Garson, 1998), we may be able to infer from the relative success of MLR in this case that the relationships among these predictors and FSIQ are largely linear.

The results of this research have several implications for practice among practitioners and researchers in the social sciences. Perhaps foremost of these is that the ubiquitous MLR model may not be the optimal tool for predicting FSIQ using demographic variables. While it generally produced predicted values that were close to the actual FSIQ scores for the cross-validation sample, it did not have the lowest RMSE values, nor was its R^2 the largest. In contrast, CART displayed lower RMSE and higher R^2 results than did MLR, as well as slightly less bias. This combination of results would suggest that for this problem, CART may have provided, overall, the highest quality predicted values of the methods studied here. On the other hand, on the basis of these outcome metrics, neither MARS nor NNET did a relatively good job of predicting SB5 FSIQ scores. Their predicted values demonstrated the greatest bias and the highest RMSE, and MARS had the single lowest R^2 . Researchers interested in making predictions should therefore consider the merits of sophisticated methods, such as CART.

The simplest prediction method included in this study was CORR, which requires the user only to obtain the zero-order Pearson correlation coefficient values between each predictor and the outcome and then apply them as weights

to the standardized predictor variables. In this instance, the predictions obtained using CORR were not competitive with those from other methods, including CART, GAM, and MLR, each of which displayed much lower bias and higher R^2 values.

In summary, the results of this study suggest that MLR may not always be the best approach for obtaining predicted values in the social sciences and that some of the alternative methods studied here, particularly CART and GAM, may be viable options, although they have not been widely used in the social sciences. It is hoped that social scientists will find this example useful and will find application of one or more of these modern prediction techniques to their own research problems useful.

Directions for future research

While it is hoped that this study serves as a good introduction to some modern methods of prediction, it is by no means the final word on this matter. Future research needs to be conducted to expand the knowledge base regarding these methods and to demonstrate their utility beyond the FSIQ data that were the focus of this study. Indeed, application of these approaches to problems in which nonlinear relationships between predictors and the outcome variable are known or strongly suspected would be particularly important in future studies. In addition, the sample for this study was relatively small, as was the pool of variables used for prediction. The limited set of demographic variables was selected consciously because it contains those measures that should be easily accessible to neuropsychologists and others interested in predicting premorbid intellectual ability for young children. At the same time, this relatively small number of predictors does not allow for the full use of some of these models. Thus, future studies might extend this literature by including more and a broader range of variables for prediction, such as region of residence, parental occupation, and achievement test data. Furthermore, studies including less explored predictors to improve the accuracy of regression formulas may be fruitful. For instance, adding current intellectual functioning level of unimpaired, healthy family members (parents and/or siblings) may contribute to increasing accuracy of premorbid ability estimation in clinical samples of young children. In addition, although a research literature containing studies of various methods for assessing premorbid level of intellectual functioning is readily available, it is unclear how and to what extent these methods are actually used by clinicians. It would be encouraging to see more research investigating what actuarial approach clinical professionals are adopting in determining whether a decline in intellectual function has

occurred and to what extent they utilize these research findings presently available. Finally, it must be noted that the present study was based on data from a healthy, nonclinical sample. However, as was noted in the introduction, the need to obtain predictions for premorbid cognitive functioning most often occurs in the context of a brain injury of some type. Therefore, in future research, these methods should be examined using a clinical sample that has suffered some type of brain trauma and whose preinjury functioning is known.

References

- Almkvist, O., & Tallberg, I. M. (2009). Cognitive decline from estimated premorbid status predicts neurodegeneration in Alzheimer's disease. *Neuropsychology*, 23, 117–124.
- Baade, L. E., & Schoenberg, M. R. (2004). A proposed method to estimate premorbid intelligence utilizing group achievement measures from school records. *Archives of Clinical Neuropsychology*, 19, 227–243.
- Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology*, 52, 885–887.
- Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.
- Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *Clinical Neuropsychologist*, 3, 129–136.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole.
- Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.
- Buhlmann, P., & Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *American Statistical Association*, 98, 324–339.
- Cheung, C.-K., & Lee, T.-Y. (2010). Improving social competence through character education. *Evaluation and Program Planning*, 33, 255–263.
- Crawford, J. R., Allan, K. M., Cochrane, R. H. B., & Parker, D. M. (1990). Assessing the validity of NART estimated premorbid IQ in the individual case. *British Journal of Clinical Psychology*, 29, 435–436.
- Crawford, J. R., Millar, J., & Milne, A. B. (2001). Estimating premorbid IQ from demographic variables: A comparison of a regression equation vs. clinical judgment. *British Journal of Clinical Psychology*, 40, 97–105.
- Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29, 317–331.
- Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Science*, 55, 119–139.
- Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London: Sage.
- Goldberg, L.R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs*, 7(2).
- Granello, D. H. (2010). Cognitive complexity among practicing counselors: How thinking changes with experience. *Journal of Counseling & Development*, 88, 92–100.
- Green, R. E., Melo, B., Christensen, B., Ngo, L. A., Monette, G., & Bradbury, C. (2008). Measuring premorbid IQ in traumatic brain injury: An examination of the validity of the Wechsler Test of Adult Reading (WTAR). *Journal of Clinical and Experimental Neuropsychology*, 30, 163–172.
- Griffin, S. L., Mindt, M. R., Rankin, E. J., Ritchie, A. J., & Scott, J. G. (2002). Estimating premorbid intelligence: Comparison of traditional and contemporary methods across the intelligence continuum. *Archives of Clinical Neuropsychology*, 17, 497–507.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.
- Krull, K. R., Sherer, M., & Adams, R. L. (1995). A comparison of indices of premorbid intelligence in clinical populations. *Applied Neuropsychology*, 2, 35–38.
- Lavigne, J. V., LeBailly, S. A., & Gouze, K. R. (2010). Predictors and correlates of completing behavioral parent training for the treatment of oppositional defiant disorder in pediatric primary care. *Behavior Therapy*, 41, 198–211.
- Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
- Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.
- Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services. *Psychological Methods*, 5, 102–124.
- Marks, M. R. (1966). *Two kinds of regression weights which are better than betas in crossed samples*. New York: Paper presented at the meeting of the American Psychological Association.
- Mathias, J. L., Bowden, S. C., Bigler, E. D., & Rosenfeld, J. V. (2007). Is performance on the Wechsler test of adult reading affected by traumatic brain injury? *British Journal of Clinical Psychology*, 46, 457–466.
- Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex*, 14, 234–244.
- Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers & Education*, 55, 663–672.
- Pavlik, V. N., Doody, R. S., Massman, P. J., & Chan, W. (2006). Influence of premorbid IQ and education on progression of Alzheimer's disease. *Dementia & Geriatric Cognitive Disorders*, 22, 367–377.
- Powell, B. D., Brossart, D. F., & Reynolds, C. R. (2003). Evaluation of the accuracy of two regression-based methods for estimating premorbid IQ. *Archives of Clinical Neuropsychology*, 18, 277–292.
- Pungello, E. P., Iruka, I. U., Dotterer, A. M., Koonce-Mills, R., & Reznick, J. S. (2009). The effects of socioeconomic status, race, and parenting on language development in early childhood. *Developmental Psychology*, 45, 544–557.
- Reynolds, C. R. (1997). Postscripts on premorbid ability estimation: Conceptual addenda and a few words on alternative and conditional approaches. *Archives of Clinical Neuropsychology*, 12, 769–778.
- Roberts, E., Bornstein, M. H., Slater, A. M., & Barrett, J. (1999). Early cognitive development and parental education. *Infant and Child Development*, 8, 49–62.
- Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.): *Technical manual*. Itasca, IL: Riverside.
- Schinka, J. A., & Vanderploeg, R. D. (2000). Estimating premorbid level of functioning. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 43–68). Mahwah, NJ: Erlbaum.
- Schoenberg, M. R., Duff, K., Scott, J. G., & Adams, R. L. (2003). An evaluation of the clinical utility of the OPIE-3 as an estimate of

- premorbid WAIS–III FSIQ. *Clinical Neuropsychologist*, 17, 308–321.
- Schoenberg, M. R., Duff, K., Scott, J. G., Patton, D., & Adams, R. L. (2006). Prediction errors of the Oklahoma Premorbid Intelligence Estimate-3 (OPIE-3) stratified by 13 age groups. *Archives of Clinical Neuropsychology*, 21, 469–475.
- Schoenberg, M. R., Lange, R. T., Brickell, T. A., & Saklofske, D. H. (2007a). Estimating premorbid general cognitive functioning for children and adolescents using the American Wechsler Intelligence Scale for Children–Fourth Edition: Demographic and current performance approaches. *Journal of Child Neurology*, 22, 379–388.
- Schoenberg, M. R., Lange, R. T., Brickell, T. A., & Saklofske, D. H. (2007b). Estimating premorbid general cognitive functioning for the American WISC–IV: Demographic and current performance approaches. *Journal of Child Neurology*, 22, 379–388.
- Schoenberg, M. R., Lange, R. T., & Saklofske, D. H. (2007c). Estimating premorbid FSIQ scores for the Canadian WISC–IV: Demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology*, 29, 867–878.
- Schoenberg, M. R., Lange, R. T., & Saklofske, D. H. (2007d). A proposed method to estimate fullscale intelligence quotient (FSIQ) for the Canadian Wechsler Intelligence Scale for Children–Fourth Edition (WISC–IV) using demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology*, 29, 867–878.
- Schoenberg, M. R., Lange, R. T., Saklofske, D. H., & Suarez, M. (2008a). Validation of the child premorbid intelligence estimate method to predict premorbid Wechsler Intelligence Scale for Children–Fourth Edition Fall Scale IQ among children with brain injury. *Psychological Assessment*, 20, 377–384.
- Schoenberg, M. R., Lange, R. T., Saklofske, D. H., Suarez, M., & Brickell, T. A. (2008b). Validation of the child premorbid intelligence estimate method to predict premorbid Wechsler Intelligence Scale for Children–4th Edition Full Scale IQ among children with brain injury. *Psychological Assessment*, 20, 377–384.
- Schoenberg, M. R., Scott, J. G., Duff, K., & Adams, R. L. (2002). Estimation of WAIS–III intelligence from combined performance and demographic variables: Development of the OPIE-3. *Clinical Neuropsychologist*, 16, 426–438.
- Schumacher, M., Robner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis*, 21, 661–682.
- Sellers, A. H., Burns, W. J., & Guyrke, L. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence–Revised as a function of stratification variables. *Applied Neuropsychology*, 9, 65–73.
- Sellers, A. H., Burns, W. J., & Guyrke, L. (1996). Prediction of premorbid intellectual functioning of young children using demographic information. *Applied Neuropsychology*, 9, 65–73.
- Vanderploeg, R. D., Schinka, J. A., & Axelrod, B. N. (1996). Estimation of WAIS–R premorbid intelligence: Current ability and demographic data used in a best-performance fashion. *Psychological Assessment*, 8, 404–411.
- Vanderploeg, R. D., Schinka, J. A., Baum, K. M., Tremont, G., & Mittenberg, W. (1998). WISC–III premorbid prediction strategies: Demographic and best performance approaches. *Psychological Assessment*, 10, 277–284.
- Veiel, H. O. F., & Koopman, R. F. (2001). The bias in regression-based indices of premorbid IQ. *Psychological Assessment*, 13, 356–368.
- Wechsler, D. (1983). *Wechsler Adult Intelligence Scale–Revised (WAIS–R): Technical Manual*. New York: Psychological Corporation.
- Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence–Revised*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, 3rd Edition (WAIS–III)*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.
- Willshire, D., Kinsella, G., & Prior, M. (1991). Estimating WAIS–R IQ from the National Adult Reading Test: A cross-validation. *Journal of Clinical and Experimental Neuropsychology*, 13, 204–216.
- Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall.
- Yeates, K. O., & Taylor, H. G. (1997). Predicting premorbid neuropsychological functioning following pediatric traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, 19, 825–837.