HHS Public Access

Author manuscript

Int J Obes (Lond). Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

Int J Obes (Lond). 2016 June ; 40(6): 887–894. doi:10.1038/ijo.2015.214.

The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research

Andrada E. Ivanescu[1], **Peng Li**[2], **Brandon George**[2], **Andrew W. Brown**[2], **Scott W. Keith**[3], **Dheeraj Raju**[4], and **David B. Allison**[2,5]

[1]Department of Mathematical Sciences, Montclair State University

[2]Office of Energetics and Nutrition Obesity Research Center, University of Alabama at Birmingham

[3]Department of Pharmacology and Experimental Therapeutics, Division of Biostatistics, Thomas Jefferson University

[4]School of Nursing, University of Alabama at Birmingham

[5]Department of Biostatistics, University of Alabama at Birmingham

Abstract

Deriving statistical models to predict one variable from one or more other variables, or predictive modeling, is an important activity in obesity and nutrition research. To determine the quality of the model, it is necessary to quantify and report the predictive validity of the derived models. Conducting validation of the predictive measures provides essential information to the research community about the model. Unfortunately, many articles fail to account for the nearly inevitable reduction in predictive ability that occurs when a model derived on one dataset is applied to a new dataset. Under some circumstances, the predictive validity can be reduced to nearly zero. In this overview, we explain why reductions in predictive validity occur, define the metrics commonly used to estimate the predictive validity of a model (e.g., R^2, mean squared error, sensitivity, specificity, receiver operating characteristic, concordance index), and describe methods to estimate the predictive validity (e.g., cross-validation, bootstrap, adjusted and shrunken R^2). We emphasize that methods for estimating the expected reduction in predictive ability of a model in new samples are available and this expected reduction should always be reported when new predictive models are introduced.

Keywords

Cross-validation; Model Validation; Nutrition; Obesity; Overfitting; Prediction Modeling; Regression; Research Reporting; Shrinkage; Statistics

*Correspondence to: David B. Allison, Ph.D., Office of Energetics and Nutrition Obesity Research Center, University of Alabama at Birmingham, Ryals Building, Room 140J, 1665 University Boulevard, Birmingham, Alabama 35294. Phone: (205) 975-9169. dallison@uab.edu.

INTRODUCTION

Deriving statistical functions from data to predict another variable is a widely used and important activity in science in general. Classic examples include prediction equations for resting metabolic rate[1] and the prediction of health outcomes, such as the risk of having a heart attack[2]. In this article, we use the word *prediction* to refer to either the true future prediction of an unobserved event or variable that has not yet occurred (e.g., time of death of a currently living individual) or the estimation of an event or variable that has already occurred but has not yet been observed (e.g., body fat of an individual on whom only anthropometric variables have been measured); we collectively refer to models that predict as *predictive modeling*. Any such model aims to predict the value of a response variable by using as input values the measured predictor variables. Studies that rely on predictive modeling typically apply regression and its extensions. To determine the quality of the model, it is necessary to quantify and report the predictive validity of the derived models. Some examples of studies that have used predictive models in obesity and nutrition research and that have measured the quality of the predictions (rightmost column) are summarized in Table 1.

Many techniques exist for deriving predictive modeling functions, and we will not provide a thorough review here. For in-depth overviews, there are many books on this topic, e.g., Kuhn & Johnson[16]. Instead, we will focus on an aspect of predictive modeling that is frequently neglected or apparently not fully understood in nutrition and obesity research, namely, the estimation of validity shrinkage. In brief, *validity shrinkage* refers to the fact that a predictive model derived from a finite sample in a manner that maximizes its predictive validity within that sample will almost assuredly not predict as well on the overall population from which the sample was drawn, on a fresh sample from the same population, or on a new sample from a different population[17]. For example, a model that predicts energy expenditure from accelerometry data may have phenomenal accuracy in a sample data set, yet in another data set, the accuracy may decline radically (see, for example, Schmid et al.[18]). This phenomenon can occur regardless of how well the model predicts the values in the sample from which it was derived. Validity shrinkage must be evaluated when a predictive model is developed, because it is not only the in-sample data (data values used to fit the model) but also the out-of-sample data (data values not used for the model fit) and their respective ability to predict that are of interest to the research and applied communities.

In other words, as will be described more fully below, the true test of a model's predictive capacity is when the model is tested on an independent dataset that was not used to develop the model. That is, there are two datasets: a training set (on which the model is derived) and a validation set (an independent testing set on which the model is evaluated). It is the validation in this testing set that is often overlooked; most studies end by maximizing predictive accuracy in the training set.

The purpose of this article is to introduce the concept of validity shrinkage in predictive modeling, to emphasize the importance of estimating the validity shrinkage when offering a new predictive model, and to describe methods for doing so. For the reader's convenience, a glossary of statistical terms is provided in Table 2.

METHODS

The Concept of Validity Shrinkage

In the next section, we will discuss the metrics that can be used to quantify the predictive validity of a model, such that a larger (e.g., R^2) or smaller (e.g., the MSE) value for a specific metric indicates greater model validity. First we address one limitation to these metrics: that they are susceptible to validity shrinkage, in which the predictive validity of a model shrinks when the model is applied to a new dataset. The cause of shrinkage is straightforward. The metrics computed from data are themselves random variables, meaning that they change from sample to sample. When we fit a model to data to maximize some function (e.g., R^2) or minimize some function (e.g., the MSE), the algorithm adjusts the model parameters to optimize that function. In doing so, the algorithm adjusts the model to fit the observed data, which will be a function of the true signal in the data and also any idiosyncrasies ('noise') due to measurement error, random sampling variance, or biased sample selection. Because these idiosyncrasies are properties of the dataset, to the extent that the model is tuned to evaluate this dataset, the model will not as effectively predict new observations in new datasets that do not share those idiosyncrasies. This phenomenon is referred to by Hitchcock and Sober[19] as 'accommodation,' but may also be called 'over-fitting,' 'capitalizing on chance,' or 'fitting to noise.'

Consider a situation in which an additional predictor, eye color, is included in a model and found to significantly improve the model validity. The contribution of the new predictor to the validity metric of the model could be considered a random variable. If eye color has no relation to the response (e.g., resting energy expenditure, or REE), we may expect its contribution to be zero. In our example, by random chance, eye color increased the model validity for that particular sample. If another sample is taken from the same population, eye color will likely not contribute to the model's validity in the new sample because of regression to the mean (variables tend to be closer to the average on subsequent measurements)[20]. This drop in validity can be described as *stochastic shrinkage* when the samples come from the same population. The methods used to estimate stochastic validity shrinkage will be discussed in a later section.

Another type of shrinkage to be concerned about is *generalizability shrinkage*, where the validity metric shrinks when the predictive model generated on a sample from one population is applied to data from an entirely different population. This may occur when something that is a useful predictor in one group is not as useful in another group. For example, if one develops a formula to predict remaining lifespan from anatomic and metabolic variables in a sample of young men, the formula may not be as strong a predictor if applied to a sample of elderly women. In response to generalizability shrinkage, external validation[21, 22] is recommended when researchers need to expand the scope and use of the predictive model. We will not discuss generalizability shrinkage further here.

Metrics of Predictive Validity

The performance of a statistical prediction model can be assessed by measuring how close the predicted values are to the observed values or by measuring the correspondence,

similarity, or agreement between the predicted and the observed values. For example, in a model that predicts REE from anthropometric and demographic variables (see Heymsfield et al.[23]), including age, adipose tissue, adipose tissue-free mass, brain mass, liver mass, and sex, one would expect the predicted REE values to closely approximate the true REE. When predicting a binary outcome (e.g., has type 2 diabetes vs. does not have type 2 diabetes) using a statistical model, individuals who experience some characteristic, such as excess adiposity as measured by a higher body mass index (BMI) value, are identified to have a higher predicted risk of type 2 diabetes than do individuals who have normal BMI (see Ley et al.[24]). According to the type of outcome (e.g., continuous, binary, ordinal, time-to-event) and the statistical model to be used, predictive validity can be evaluated by different methods and metrics, which are implemented in many available statistical software packages. Several of the most common metrics are described below, although this list is by no means exhaustive. It should be noted that all the metrics from a specific sample are biased estimates and their accuracy must be tested in other independent samples or in the overall population.

The coefficient of determination, or R^2—R^2, also known as the coefficient of determination, estimates the amount of the total variance in the outcome measurements that is explained by the predictive model. Expressed as a ratio, R^2 ranges from 0 to 1. R^2 is commonly used for continuous outcomes, but some variations can also be used for other data types[25]. For a predictive regression model, a value of R^2 closer to 1 indicates better prediction. One drawback to the use of R^2 for estimating predictive validity is that its value in any sample can only increase or stay the same when more predictor variables are added into the model. Thus, the use of a large number of predictors may lead to an over-fitted predictive model.[20] In addition, the sample statistic value of R^2 is an upwardly biased estimator of the true population parameter R^2. The adjusted R^2 or shrunken R^2, presented in a later section, should be computed to assess the predictive performance.

The mean squared error, or MSE—Another broadly used method to evaluate predictive validity for continuous outcomes is the mean squared error (MSE). The MSE is calculated as the average of the squared differences between the observed and predicted values. A smaller value for MSE indicates that the predicted values are closer to the observed data and therefore a better prediction. Similar to the R^2, the MSE of a predictive model in a new sample is generally larger than the MSE in the training sample because the new sample is not used for fitting the predictive model.

Sensitivity, specificity, and the receiver operating characteristic—In certain cases, we may wish to predict whether a patient will develop a binary outcome such as obesity from a set of patient characteristics at baseline. Since the output of the predictive model will be a predicted risk, and the patient either will or will not have the disease after a certain amount of time, the above metrics are unsuitable for this scenario. Hence, for binary predictions with a statistical method (e.g., logistic regression), researchers commonly consider specificity and sensitivity as metrics of model validity and visualize them with the receiver operating characteristic (ROC) curve[26]. The possible results of binary predictions are given in Table 3. We define *sensitivity* as the ratio of correctly predicted positives (sometimes called events or denoted outcome = 1, denoted as *d* in Table 3) divided by the

total number of actual positives ($c+d$); in our example, sensitivity would be the fraction of patients we had predicted would become obese over all the patients who actually became obese. We then define *specificity* as the ratio of the correctly predicted negatives (sometimes called *nonevents* or denoted outcome = 0, denoted as a in Table 3) to the total of actual negatives ($a+b$); here, specificity would be the fraction of patients we had predicted would not become obese over all the patients who actually did not become obese. In this framework, perfect prediction (meaning no false positives or false negatives) would have sensitivity and specificity both equal to 1.

If a predictive model provides predicted risks for subjects, then the actual prediction of the outcome will be based on defining a cutoff risk for the disease. For example, we could predict that all patients with a risk over 60% will become obese, while those with a risk under 60% will not. In practice, the choice of cutoff will affect both sensitivity and specificity such that increased sensitivity may result in lower specificity, or vice versa. As an extreme example, predicting that every patient will become obese (i.e., the cutoff is at 0% predicted risk) would have a sensitivity of 1 but a specificity of 0. The ROC curve visualizes this potential trade-off by plotting sensitivity vs (1 - specificity) across a series of cutoff points that change the prediction of positive and negative outcomes and therefore change observed sensitivity and specificity. The area under an ROC curve (AUC) is used as a quantitative measure of the ROC and prediction assessment of the statistical method used. AUCs are bounded by 0 and 1, where in general an AUC close to 1 indicates a nearly perfect prediction, an AUC of 0.5 indicates that the prediction is no better than chance, and an AUC of less than 0.5 indicates worse prediction than by chance. For situations in which the prevalence of an event is to be accounted for, the metrics of positive predicted value (PPV), negative predicted value (NPV), see e.g., Kuhn & Johnson[16], and related measures (e.g., Ozenne et al.[27]) are available beyond sensitivity, specificity, and the ROC.

The concordance index, or *c*—The concordance index, denoted by c, is a statistic that measures the agreement between the rank order of the observed and the predicted values (see Harrell et al.[20]). For example, the c index quantifies whether the subject with the highest observed REE also has the highest predicted REE, whether the subject with the second-highest observed REE has the second-highest predicted REE, and so on. Sometimes referred to as the c-statistic, the c index can be interpreted as the probability of concordance between the predicted and the observed outcomes and can be interpreted much like the ROC AUC: a value of 1 indicates perfect prediction, whereas a value of 0.5 indicates a prediction no better than chance. Because the c index does not involve the underlying distributions of the outcomes, it can be used to evaluate predictive validity for continuous, binary, ordinal, and censored time-to-event outcomes. Hence, the c index and its variants are especially useful for survival (time-to-event) prediction. If the outcome to be predicted is a binary response, c is equivalent to the ROC AUC.

How to Quantify Shrinkage and Assess Validity Under Shrinkage

As previously mentioned, an integral part of model validation is the estimation of how much the model's validity will hold or shrink when the model is used on an independent dataset. When a predictive model is being developed, shrinkage is best quantified by cross-

validation. If one is looking to estimate the validity of a model reported in the literature, back-of-the-envelope calculations such as the adjusted R^2 and shrunken R^2 are available.

Cross-validation—The main function of cross-validation is to provide qualitative and quantitative insight[20] on the model's potential to predict outcomes in new independent samples without collecting additional data. In cross-validation, the available data are split into two sets of data that are used for two distinct purposes. One set of data is used to construct the predictive model using the measured values of the outcome and predictor variables; this part is called the training set. The second set of data is used to test the predictive model; this set is called the validation set.

In practice, this process of cross-validation can be done by using data splitting of varying sizes to give a more robust estimate of the true validity of the predictive model. A common approach is *k*-fold cross-validation, the steps of which are listed below.

1. Randomly partition the original sample into *k* sets of equal size (*k* is often chosen as 5 or 10).
2. Choose one of the *k* sets to be the validation set, and combine the other *k*-1 sets to create the training set.
3. Fit the predictive model on the training set.
4. Calculate the metric of predictive validity using the validation set.
5. Repeat steps 2–4 *k* times, so that each of the *k* sets is used as the validation set once.
6. Average the *k* metrics of predictive validity to get the estimated predictive measure of the true validity of the model. Shrinkage can be calculated as the difference between this metric and the metric from using all the data together.

Note that *k*-fold cross-validation is preferable to single data-splitting (also called the validation set approach[16, 20, 28–30]), where the data are split into two sets and only one is used as a validation set, because each of the two resulting sets may not be representative of the entire dataset. A special case of *k*-fold cross-validation is leave-one-out cross-validation, where *k* is equal to the sample size and we consider the validity of the model on each subject in turn when removed from the training set. Leave-one-out cross-validation is useful for examining whether any one observation has an undue effect on the properties of the predictive model and in cases where the sample may be undersized for cross-validation. For datasets where the outcome variable is binary and the prevalence in the sample is greatly different from 50%, one may wish to partition the data (step 1) by using stratified sampling[16] so that the training set more closely resembles the entire original dataset.

Bootstrap—Another method of utilizing already-collected data, the bootstrap, can be used to estimate the shrinkage[20] of predictive performance for a metric. A bootstrap sample is generated from the entire original data by drawing a random sample with replacement, such that a particular observation may be included in the bootstrap sample once, or more than once, or possibly not at all. The predictive performances across bootstrap samples are

combined[31,32,33]. Bootstrap-derived measures of predictive validity are obtained separately (see, for example, Ambroise and McLachlan[32], Molinaro et al.[33]) from cross-validation-derived metrics.

Variable selection when assessing validity—An important source of validity shrinkage is when estimates of predictive validity are based on data used to select a subset of the most promising variables drawn from a set of possible predictors for the 'final' selected model. In Chapter 7 of their book, Hastie et al.[29] point out the gravity of applying variable selection (for example, subset selection and stepwise methods) once on the entire sample before obtaining a metric of predictive performance. If the data from the training set are used for both the selection and the validation of predictors selected for the 'final' predictive model, it is a logical consequence that those predictors will have exaggerated validity. The predictive performance of this model in those data should not be expected to represent well the performance of the model in another independent sample from the population. This issue illustrates why a separate validation set is important for evaluating the predictive performance of a model. Ambroise and McLachlan[32] explain that cross-validation or bootstrap approaches can be applied appropriately in the context of data-driven model variable selection, but care must be taken to appropriately account for the downwardly biased prediction error estimation caused by assessing model performance within the selection process. The key to doing this is to keep the model selection independent of the prediction error evaluation. So, rather than fitting the same 'final' model consisting of previously selected variables in each of the training subsets generated by the cross-validation or bootstrapping routine, the same variable selection procedure used to establish the 'final' model must be performed in each of the training subsets to arrive at, possibly, newly selected models to be evaluated for prediction error in the corresponding validation subsets.[32]

Adjusted and shrunken R^2—As mentioned above, the R^2 tends to overestimate a model's validity because it does not account for over-fitting of the model. Modifications to the R^2 have been proposed that penalize the metric for an increased number of predictors relative to the total sample size.

The adjusted R^2 is constructed to estimate the true R^2 in the population. The adjusted R^2 takes into account the sample size (n) and the number of predictors (p), based on an adjustment that involves the ratio p/n. The larger the number of predictors p relative to the sample size n, the larger the necessary reduction because a larger amount of shrinkage is expected. The adjusted R^2 given by Wherry's formula[34] provides an estimate ($\hat{\tau^2}$) for the population parameter R^2 (τ^2) and is as follows:

$$\hat{\tau}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2).$$

Furthermore, the shrunken R^2 (Lord-Nicholson type shrinkage[34] given by a derived formula of Cattin[35]) is applied to estimate how well the regression obtained in the given sample would predict future samples from the same population from which the derivation sample was drawn. The shrunken R^2 formula[35] is given by

$$\hat{\tau}_s^2 = \frac{(n-1)\tau^4 + \tau^2}{(n-p)\tau^2 + p}$$

where τ^2 can be estimated by using the Wherry's estimator defined above. Note that the number of predictors (p) in the denominator accounts for how a model with a larger number of predictors relative to sample size (n) is more likely to experience greater shrinkage in validity when applied to a new sample.

Adjusted and shrunken R^2 are relatively easy to implement and require no added computational tools for resampling the data. They offer well-recognized and established measures of validity shrinkage for goodness of prediction with linear regression. The adjusted R^2 is usually reported by most statistical analysis programs, even MS Excel. Given the ease of implementation and inclusion in the standard reporting of statistical software such as MS Excel, SAS, Stata, R, and SPSS, the adjusted or shrunken R^2 should always be reported by researchers proposing predictive models involving the R^2 metric. These statistics should also be considered when selecting the variables to go into the predictive model.

***Case study for the* R^2 *metric*—**We applied validity shrinkage estimation by shrunken R^2 to a sample of published results of predictive performance. Shrunken R^2 was obtained by using the formulas presented in the previous section. Table 4 presents the collection of studies that used linear regression and R^2 to measure predictive performance for study outcomes. We can see that in studies such as Yamanaka et al.[37] that used a large number of predictors (p=17) relative to sample size (n=36), the amount of validity shrinkage was substantial (R^2=0.672 reduced to shrunken R^2 of 0.207). Note that adding a greater number of predictors is not inherently deleterious if one also uses a larger sample size, as in Blalock et al.[36] However, we see that in general an overreliance on validity metrics susceptible to shrinkage could lead to overstated conclusions of model validity that can mislead readers concerning the actual capacity of the model to predict new samples.

Examples of Well Done Validation

Examples in the literature in which validation was conducted and the resulting predictive performance was reported include the following: (1) a single split-sample, or validation set approach, dealing with a single training sample and a single validation sample[11–14, 45, 46]; (2) leave-one-out cross-validation[8,10]; (3) k-fold cross-validation[16]; and (4) bootstrap[32]. As mentioned above, the gold standard for assessing predictive ability is validation on an external dataset. Although this approach may be difficult to do, it can be done and should always be considered. Examples of external validation include Rush et al.[14], Jackson et al.[47], and Steyerberg et al.[48]

Common Misconceptions on Acquiring Model Validity

It is possible to avoid many mistakes when trying to establish the validity of a proposed predictive model. Two particularly common misconceptions that are easily avoidable are a lack of validation, which we discuss next, and the variable selection when assessing validity, already discussed in a previous section.

No validation—A common misconception in the development and reporting of predictive models is that model validation beyond performance on the training set is not necessary. As we have shown above, foregoing the assessment of potential validity shrinkage may give a false sense of model accuracy to other investigators interested in applying that model to their datasets. Conducting validation of the predictive measures provides essential information to the research community about the model.

In the case where an independent sample is not readily available for external validation, the study of predictive modeling[38, 49] should be accompanied by a cross-validation-derived predictive validity measure[20] or a metric that is a formula-derived (e.g., shrunken R^2) adjustment.

Case examples—The literature on obesity and related exposures and outcomes is rich in articles detailing statistical models that estimate associations that do not take into account the predictive accuracy of those models during optimization of the model or when applying the model to external samples. This is a shortcoming in many of these cases, such as when the focus is on the association between an outcome and a particular exposure of interest[50]. It becomes more clearly problematic when a model or a component of a model is presented as a tool for predicting outcomes in external samples. For instance, Xi et al.[51] devised an obesity genetic risk score for predicting insulin resistance. They derived the risk score model in a training sample of Chinese children without conducting the proper evaluation of the model's prediction error. As such, because prediction error was neither minimized nor subsequently evaluated by, for example, cross-validation, it remains unclear whether this obesity genetic risk score is useful in practice.

DISCUSSION

How to Minimize Shrinkage

In this section we point out several strategies for constructing predictive models that may result in minimizing validity shrinkage. We discuss increasing the sample size of the study, reducing the number of predictors, informing predictive models by theory or prior knowledge, and using robust methods. In any given study of predictive modeling, these strategies can plausibly be applied jointly, and necessarily in conjunction with validation, such as cross-validation.

Increase sample size—Collecting data on more individuals from the population is the obvious way to increase the sample size of a study. Appropriately handling of observations with missing data can also increase the sample size available for analysis. As such, the methods used for model fitting should be carefully selected in the presence of missing data (Little and Rubin[52], Graham[53]).

Reduce the number of predictors—Model parameters provide model flexibility, leading to the quotation attributed to John von Neumann: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." It is common for researchers to want to put all possible predictors in the model (estimating at least one model parameter for each), but, as we have discussed, this can create issues for validity shrinkage, such as the

possibility of reporting un-validated metrics of predictive performance. When a larger number of predictors are considered, too many predictors are sometimes included given the sample size. As a result, the model may have a large reduction in predictive validity when applied to new data. In addition, including redundant variables, such as highly correlated variables, in a predictive model does not add more information but leads to multicollinearity that may result in unstable regression models. The use of redundant variables should therefore be avoided. When a large number of variables are available, some model selection strategies may be applied such as LASSO, which is explained under robust methods.

Inform model by theory or prior knowledge—Experts, discipline-specific research, and prior studies may provide knowledge about the relationship between study variables and the response. Pertinent scientific reason may inform the consideration of predictors. Knowledge independent of the current data can be helpful for prediction in other data. For example, prior knowledge indicating a nonlinear relationship between a predictor and the response can be accounted for directly when constructing the predictive model. An example is that an individual's height should be accounted for when predicting total body mass owing to the square-cube law.

Robust methods—Given a library of predictor variables, each model component may not be equally important for the outcome. Each predictor introduces at least one additional model component that needs to be estimated. Least absolute shrinkage and selection operator (LASSO) and ridge regression are "shrinkage" methods[29, 54] for obtaining estimated coefficients by applying a calculated penalty to reduce the magnitude of the model coefficients. The resulting parameter coefficients are "shrunk" (i.e., penalized toward a null value of zero) depending on their perceived importance for predicting the response. Ridge regression cannot zero out the parameter coefficients, whereas the LASSO method can select predictors by shrinking the parameter coefficients to zero for those predictors with small effects. Thus, LASSO is preferable when some of the predictors in the model may not be important. The methods are robust to estimate model coefficients in the presence of many predictors with large or small magnitude[55]. LASSO-related methods include elastic net, which applies a LASSO and ridge-combined penalty type to select and shrink the coefficients. For example, to predict type 2 diabetes remission after bypass surgery, a binary variable (yes/no), Cottilard et al.[56] studied a predictive model using LASSO with a set of 10 preoperative measures as predictor variables: *sex*, *age*, *BMI*, *fasting glycemia*, *HbA1c*, *hypertension*, *T2D duration*, *insulin therapy*, *number of anti-diabetic drugs*, and *C-peptide value*.

Conclusions

Predictive modeling can have great value in many applications and is frequently used in obesity and nutrition research. Readers and potential future users of a published predictive model will want to know how good the predictive validity of that model is, that is, how well the model will predict future, yet-to-be-observed data. Investigators typically provide metrics that give information about the predictive validity of the model in the sample on which it was derived, but commonly neglect to provide how well the model can be expected to perform in future samples. As we have explained herein, the model will almost assuredly

not predict as well in future samples as it did in the sample from which it was derived, and this degree of validity shrinkage can be quite large. Methods for estimating and reporting validity shrinkage and the expected predictive validity in new samples are available and should be used.

Acknowledgments

Supported in part by NIH grants R25DK099080, R25HL124208 and P30DK056336. The opinions expressed are those of the authors and do not necessarily represent those of the NIH or any other organization. We gratefully acknowledge the anonymous reviewers for their helpful suggestions, which substantially improved this article.

References

1. Heshka S, Feld K, Yang MU, Allison DB, Heymsfield SB. Resting energy expenditure in the obese: a cross-validation and comparison of prediction equations. Journal of the American Dietetic Association. 1993; 93(9):1031–6. [PubMed: 8360408]
2. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA : the journal of the American Medical Association. 2009; 302(21): 2345–52. [PubMed: 19952321]
3. Baan CA, Ruige JB, Stolk RP, Witteman JC, Dekker JM, Heine RJ, et al. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. Diabetes care. 1999; 22(2):213–9. [PubMed: 10333936]
4. Dixon JB, Chuang LM, Chong K, Chen SC, Lambert GW, Straznicky NE, et al. Predicting the glycemic response to gastric bypass surgery in patients with type 2 diabetes. Diabetes care. 2013; 36 (1):20–6. [PubMed: 23033249]
5. Forns X, Ampurdanes S, Llovet JM, Aponte J, Quinto L, Martinez-Bauer E, et al. Identification of chronic hepatitis C patients without hepatic fibrosis by a simple predictive model. Hepatology. 2002; 36(4 Pt 1):986–92. [PubMed: 12297848]
6. Hayes MT, Hunt LA, Foo J, Tychinskaya Y, Stubbs RS. A model for predicting the resolution of type 2 diabetes in severely obese subjects following Roux-en Y gastric bypass surgery. Obes Surg. 2011; 21(7):910–6. [PubMed: 21336560]
7. Li S, Zhao JH, Luan J, Luben RN, Rodwell SA, Khaw KT, et al. Cumulative effects and predictive value of common obesity-susceptibility variants identified by genome-wide association studies. Am J Clin Nutr. 2010; 91(1):184–90. [PubMed: 19812171]
8. Thomas DM, Ivanescu AE, Martin CK, Heymsfield SB, Marshall K, Bodrato VE, et al. Predicting successful long-term weight loss from short-term weight-loss outcomes: new insights from a dynamic energy balance model (the POUNDS Lost study). Am J Clin Nutr. 2015; 101(3):449–54. [PubMed: 25733628]
9. Cancello R, Tordjman J, Poitou C, Guilhem G, Bouillot JL, Hugol D, et al. Increased infiltration of macrophages in omental adipose tissue is associated with marked hepatic lesions in morbid human obesity. Diabetes. 2006; 55(6):1554–61. [PubMed: 16731817]
10. Chen H, Sullivan G, Quon MJ. Assessing the predictive accuracy of QUICKI as a surrogate index for insulin sensitivity using a calibration model. Diabetes. 2005; 54(7):1914–25. [PubMed: 15983190]
11. Clasey JL, Bradley KD, Bradley JW, Long DE, Griffith JR. A new BIA equation estimating the body composition of young children. Obesity. 2011; 19(9):1813–7. [PubMed: 21681223]
12. Garcia AL, Wagner K, Hothorn T, Koebnick C, Zunft HJ, Trippo U. Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. Obes Res. 2005; 13(3): 626–34. [PubMed: 15833949]
13. Huang TT, Watkins MP, Goran MI. Predicting total body fat from anthropometry in Latino children. Obes Res. 2003; 11(10):1192–9. [PubMed: 14569044]
14. Rush EC, Chandu V, Plank LD. Prediction of fat-free mass by bioimpedance analysis in migrant Asian Indian men and women: a cross validation study. International journal of obesity (2005). 2006; 30(7):1125–31. [PubMed: 16432545]

15. Russell M, Mendes N, Miller KK, Rosen CJ, Lee H, Klibanski A, et al. Visceral fat is a negative predictor of bone density measures in obese adolescent girls. The Journal of clinical endocrinology and metabolism. 2010; 95(3):1247–55. [PubMed: 20080853]
16. Kuhn, M.; Johnson, K. Applied Predictive Modeling. Springer; 2013.
17. Copas JB. Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society Series B (Methodological). 1983; 45(3):311–354.
18. Schmid M, Riganti-Fulginei F, Bernabucci I, Laudani A, Bibbo D, Muscillo R, et al. SVM versus MAP on accelerometer data to distinguish among locomotor activities executed at different speeds. Computational and mathematical methods in medicine. 2013; 2013:343084. [PubMed: 24376469]
19. Hitchcock C, Sober E. Prediction versus accommodation and the risk of overfitting. Brit J Philos Sci. 2004; 55(1):1–34.
20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996; 15(4):361–87. [PubMed: 8668867]
21. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012; 98(9):691–8. [PubMed: 22397946]
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015; 13:1. [PubMed: 25563062]
23. Heymsfield SB, Thomas D, Bosy-Westphal A, Shen W, Peterson CM, Muller MJ. Evolving concepts on adjusting human resting energy expenditure measurements for body size. Obes Rev. 2012; 13(11):1001–14. [PubMed: 22863371]
24. Ley SH, Hamdy O, Mohan V, Hu FB. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. Lancet. 2014; 383(9933):1999–2007. [PubMed: 24910231]
25. Mittlbock M, Schemper M. Explained variation for logistic regression. Statistics in medicine. 1996; 15(19):1987–97. [PubMed: 8896134]
26. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007; 115(5):654–7. [PubMed: 17283280]
27. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015; 68(8):855–9. [PubMed: 25881487]
28. Harrell, FFJ. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; New York: 2001.
29. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. Springer; 2009.
30. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning. Springer; 2013.
31. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Vol. 57. Chapman & Hall/CRC Monographs on Statistics & Applied Probability; NY: 1993.
32. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(10):6562–6. [PubMed: 11983868]
33. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005; 21(15):3301–7. [PubMed: 15905277]
34. Bobko, P. Correlation and Regression: Applications for Industrial Organizational Psychology and Management. Sage Publications; 2001.
35. Cattin P. Estimation of the predictive power of a regression model. Journal of Applied Psychology. 1980; 65(4):407–414.
36. Blalock SJ, Currey SS, DeVellis RF, Anderson JJ, Gold DT, Dooley MA. Using a short food frequency questionnaire to estimate dietary calcium consumption: a tool for patient education. Arthritis care and research : the official journal of the Arthritis Health Professions Association. 1998; 11(6):479–84. [PubMed: 10030180]

37. Yamanaka N, Okamoto E, Kuwata K, Tanaka N. A multiple regression equation for prediction of posthepatectomy liver failure. Annals of surgery. 1984; 200(5):658–63. [PubMed: 6486915]
38. Scalfi L, Marra M, De Filippo E, Caso G, Pasanisi F, Contaldo F. The prediction of basal metabolic rate in female patients with anorexia nervosa. International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity. 2001; 25(3):359–64.
39. Mahon AD, Marjerrison AD, Lee JD, Woodruff ME, Hanna LE. Evaluating the prediction of maximal heart rate in children and adolescents. Research quarterly for exercise and sport. 2010; 81(4):466–71. [PubMed: 21268470]
40. Roediger HL 3rd, Watson JM, McDermott KB, Gallo DA. Factors that determine false recall: a multiple regression analysis. Psychonomic bulletin & review. 2001; 8(3):385–407. [PubMed: 11700893]
41. Nomani MZ, Khan AH, Shahda MM, Nomani AK, Sattar SA. Predicting serum gastrin levels among men during Ramadan fasting. Eastern Mediterranean health journal = La revue de sante de la Mediterranee orientale = al-Majallah al-sihhiyah li-sharq al-mutawassit. 2005; 11(1–2):119–25.
42. McKeon JL, Murree-Allen K, Saunders NA. Prediction of oxygenation during sleep in patients with chronic obstructive lung disease. Thorax. 1988; 43(4):312–7. [PubMed: 3406918]
43. Puyau MR, Adolph AL, Vohra FA, Zakeri I, Butte NF. Prediction of activity energy expenditure using accelerometers in children. Medicine and science in sports and exercise. 2004; 36(9):1625–31. [PubMed: 15354047]
44. Siervo M, Prado C, Hooper L, Munro A, Collerton J, Davies K, et al. Serum osmolarity and haematocrit do not modify the association between the impedance index (Ht(2)/Z) and total body water in the very old: the Newcastle 85+ study. Archives of gerontology and geriatrics. 2015; 60(1):227–32. [PubMed: 25288578]
45. Hoffman DJ, Toro-Ramos T, Sawaya AL, Roberts SB, Rondo P. Estimating total body fat using a skinfold prediction equation in Brazilian children. Ann Hum Biol. 2012; 39(2):156–60. [PubMed: 22324842]
46. Lee JJ, Freeland-Graves JH, Pepper MR, Yao M, Xu B. Predictive equations for central obesity via anthropometrics, stereovision imaging and MRI in adults. Obesity. 2014; 22(3):852–62. [PubMed: 23613161]
47. Jackson AS, Stanforth PR, Gagnon J, Rankinen T, Leon AS, Rao DC, et al. The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity. 2002; 26(6):789–96.
48. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21(1):128–38. [PubMed: 20010215]
49. Apfelbacher CJ, Loerbroks A, Cairns J, Behrendt H, Ring J, Kramer U. Predictors of overweight and obesity in five to seven-year-old children in Germany: results from cross-sectional studies. Bmc Public Health. 2008; 8:171. [PubMed: 18495021]
50. Gerhard GS, Benotti P, Wood GC, Chu X, Argyropoulos G, Petrick A, et al. Identification of novel clinical factors associated with hepatic fat accumulation in extreme obesity. Journal of obesity. 2014; 2014:368210. [PubMed: 25610640]
51. Xi B, Zhao X, Shen Y, Wu L, Hou D, Cheng H, et al. An obesity genetic risk score predicts risk of insulin resistance among Chinese children. Endocrine. 2014; 47(3):825–32. [PubMed: 24619288]
52. Little, RJA.; Rubin, DB. Statistical analysis with missing data. Wiley; New York: 1987.
53. Graham JW. Missing data analysis: making it work in the real world. Annual review of psychology. 2009; 60:549–76.
54. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met. 1996; 58(1): 267–288.
55. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11(12):880–6. [PubMed: 21045869]

56. Cotillard A, Poitou C, Duchateau-Nguyen G, Aron-Wisnewsky J, Bouillot J-L, Schindler T, et al. Type 2 diabetes remission after gastric bypass: what is the best prediction tool for clinicians? Obesity Surgery. 2015; 25:1128–1132. [PubMed: 25387683]

Table 1

Example studies of predictive modeling in obesity and nutrition

Study	Predictive Model Topic	Assessment measure of predictive performance*
Categorical response variable		
Baan et al.[3]	Identify risk for undiagnosed diabetes in a healthcare setting	ROC/AUC Sensitivity/Specificity PPV/NPV
Dixon et al.[4]	Explore predictors of diabetes remission and conversely inadequate glycemic control after gastric bypass surgery	ROC/AUC Sensitivity/Specificity PPV/NPV
Forns et al.[5]	Identify liver fibrosis in patients with chronic hepatitis C	ROC/AUC Sensitivity/Specificity PPV/NPV
Hayes et al.[6]	Predict the resolution of type 2 diabetes in severely obese patients after gastric bypass surgery	Correctly/Incorrectly Classified (%)
Li et al.[7]	Predict obesity risk as a function of genetic loci and different anthropometric traits	ROC/AUC
Thomas et al.[8]	Predict long-term successful weight management from outcomes measured in the first 3 months	ROC/AUC
Continuous response variable		
Cancello et al.[9]	Compare associations of macrophages in white adipose tissue with hepatic damage in obese patients	R^2
Chen et al.[10]	Develop a calibration model to compare the ability of quantitative insulin-sensitivity check index and supplementary surrogates to predict clamp index of insulin sensitivity	MSE
Clasey et al.[11]	Develop body composition prediction equations for pediatric use	R^2
Garcia et al.[12]	Predict body fat by using measurements of skinfold and body part circumferences	R^2
Huang et al.[13]	Predict total body fat by using weight, height, age, gender, Tanner stage, and abdominal skinfold in children	R^2
Rush et al.[14]	Predict fat-free mass by using body composition measures	R^2
Russell et al.[15]	Determine associations of regional fat mass and adipokines with body mineral density	R^2

*ROC, receiver operating characteristic; AUC, area under the ROC curve; PPV, positive predictive value; NPV, negative predictive value; R^2, coefficient of determination; MSE, mean squared error.

Table 2

Glossary of statistical terms

Term	Definition
Adjusted R^2	A metric of predictive ability, such that values closer to 1 suggest higher predictive ability. Modifies the R^2 to account for the number of predictor variables in the model relative to the sample size. Less susceptible to validity shrinkage than the basic R^2. It is intended to be an unbiased estimator of the squared correlation between the predicted and actual values in the population from which the derivation sample was drawn.
AUC, area under the curve	A metric of predictive validity, such that values closer to 1 suggest higher predictive ability. Defined as the area under the ROC curve.
Bootstrap	A resampling procedure that can be used for estimating the possible validity shrinkage of a predictive model. Works by drawing random sets of data from the observed data.
c, the concordance index	A metric of predictive validity, such that values closer to 1 suggest higher predictive ability. Measures the concordance between observed and predicted outcomes.
Cross-validation	A procedure for estimating the possible validity shrinkage of a predictive model. Works by separating the observed data into separate training and validation sets.
Generalizability shrinkage	Validity shrinkage occurring due to the model being applied to data from a different population than the one it was built in.
Mean squared error (MSE)	A metric of predictive validity, such that values closer to zero suggest better predictive ability. The average of the square of the differences between the observed and predicted values.
Population parameter	The true, unobservable value defining the distribution of a variable or associations between variables in the population.
Prediction	The true future prediction of an unobserved event or variable that has not yet occurred OR the estimation of an event or variable that has already occurred but has not yet been observed.
Predictive ability	How ‘good’ a predictive model is at predicting an unobserved event or variable from provided values of predictor variables.
Predictive modeling	Statistical models used to make predictions based on values of predictor variables.
R^2, the coefficient of determination	A metric of predictive ability, such that values closer to 1 suggest higher predictive ability. The proportion of variance in a continuous outcome that is explained by the predictive model.
Response variable	The study variable needing to be predicted. Also known as dependent, outcome, or target variable.
ROC, receiver operator characteristic curve	A curve of sensitivity versus (1-specificity), defining how the predictive ability of a model changes with different cutoffs for predicted risk.
Sample statistic	An estimate of the true population parameter obtained from a finite sample from the population.
Sensitivity	The proportion of ‘positive’ events successfully predicted by the model.
Shrunken R^2	A metric of predictive ability, such that values closer to 1 suggest higher predictive ability. Modifies the R^2 to account for the number of predictor variables in the model relative to the sample size. Less susceptible to validity shrinkage than the basic R^2. It is intended to be an unbiased estimator of the squared correlation between the predicted and actual values in a new sample from the same population from which the derivation sample was drawn.
Specificity	The proportion of ‘negative’ events successfully predicted by the model.
Stochastic shrinkage	Validity shrinkage occurring due to variations from one finite sample to another.
Validity shrinkage	The reduction in predictive ability of a predictive model when moving from the data used to construct the model to and a new, independent dataset.

Table 3

Cross-tabulation of actual values versus possible predictions for a binary outcome

		Prediction	
		No Outcome (Y=0)	**Has Outcome (Y=1)**
Reality	No Outcome (Y=0)	a (True negative)	b (False positive)
	Has Outcome (Y=1)	c (False negative)	d (True positive)

Table 4

R^2 and the corresponding shrunken R^2 as a function of sample size (n) and number of variables (p) for prediction using linear regression

Study	R^2	*Shrunken* R^2	*Amount* R^2 *shrinkage*	*Percent* R^2 *shrinkage*	*n*	*p*
Yamanaka et al.[37]	0.672	0.207	0.465	69.13%	36	17
Scalfi et al.[38]	0.160	0.131	0.029	18.12%	86	2
Mahon et al.[39]	0.290	0.248	0.041	14.41%	52	2
Roedringer et al.[40]	0.680	0.594	0.085	12.50%	55	7
Nomani et al.[41]	0.753	0.664	0.088	11.76%	30	5
McKeon et al. [42]	0.810	0.719	0.090	11.19%	24	5
Puyau et al.[43]	0.789	0.718	0.070	8.92%	32	5
Blalock et al.[36]	0.972	0.940	0.031	3.23%	268	97
Siervo et al.[44]	0.970	0.968	0.002	0.14%	252	6