

Statistical modeling methods: challenges and strategies

Steven S. Henley, Richard M. Golden & T. Michael Kashner

To cite this article: Steven S. Henley, Richard M. Golden & T. Michael Kashner (2019): Statistical modeling methods: challenges and strategies, *Biostatistics & Epidemiology*, DOI: [10.1080/24709360.2019.1618653](https://doi.org/10.1080/24709360.2019.1618653)

To link to this article: <https://doi.org/10.1080/24709360.2019.1618653>



Published online: 22 Jul 2019.



Submit your article to this journal [↗](#)



Article views: 614



View related articles [↗](#)



View Crossmark data [↗](#)



Statistical modeling methods: challenges and strategies

Steven S. Henley^{a,b,c}, Richard M. Golden^d and T. Michael Kashner^{a,b,e}

^aDepartment of Medicine, Loma Linda University School of Medicine, Loma Linda, CA, USA; ^bCenter for Advanced Statistics in Education, VA Loma Linda Healthcare System, Loma Linda, CA, USA; ^cMartingale Research Corporation, Plano, TX, USA; ^dSchool of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX, USA; ^eDepartment of Veterans Affairs, Office of Academic Affiliations (10A2D), Washington, DC, USA

ABSTRACT

Statistical modeling methods are widely used in clinical science, epidemiology, and health services research to analyze data that has been collected in clinical trials as well as observational studies of existing data sources, such as claims files and electronic health records. Diagnostic and prognostic inferences from statistical models are critical to researchers advancing science, clinical practitioners making patient care decisions, and administrators and policy makers impacting the health care system to improve quality and reduce costs. The veracity of such inferences relies not only on the quality and completeness of the collected data, but also statistical model validity. A key component of establishing model validity is determining when a model is not correctly specified and therefore incapable of adequately representing the Data Generating Process (DGP). In this article, model validity is first described and methods designed for assessing model fit, specification, and selection are reviewed. Second, data transformations that improve the model's ability to represent the DGP are addressed. Third, model search and validation methods are discussed. Finally, methods for evaluating predictive and classification performance are presented. Together, these methods provide a practical framework with recommendations to guide the development and evaluation of statistical models that provide valid statistical inferences.

ARTICLE HISTORY



Received 13 June 2018
Accepted 24 April 2019

KEYWORDS

Goodness-of-fit; Information Matrix Test; model misspecification; model selection; specification analysis

1. Introduction

Statistical modeling methods [1–17] are widely used in clinical science, epidemiology, and health services research to analyze and interpret data obtained from clinical trials as well as observational studies of existing data sources, such as claims files and electronic health records. Diagnostic and prognostic inferences from statistical models are critical if researchers are to advance science, clinical practitioners along with their patients are to make informed care decisions, and administrators and policy makers are to positively

CONTACT Steven S. Henley  stevenh@martingale-research.com  Department of Medicine, Loma Linda University School of Medicine, Loma Linda 92357, CA, USA; Center for Advanced Statistics in Education, VA Loma Linda Healthcare System, Loma Linda 92357, CA, USA; Martingale Research Corporation, 101 E. Park Blvd., Suite 600, Plano 75074, TX, USA

impact the health care system to improve quality, enhance access, and reduce costs. The veracity of such inferences relies not only on the quality and completeness of the collected data, but also on the validity of the researcher's model.

Models that can adequately represent the true underlying process by which the data were created, known as the Data Generating Process (DGP) [18,19], reveal data structure, provide performance of parameter estimators, evidence robustness of statistical inferences, and allow evaluation of the underlying assumptions that support interpreting analytic findings [20–23]. Such models also serve an important communicative function by facilitating 'model transparency' [20,24], which supports future scientific inquiry and more effective distribution of research findings.

This article describes challenges and provides recommended strategies for developing valid statistical models that more accurately approximate the underlying DGP. In particular, we focus on methods for regression analysis that pertain to generalized linear models (GLM) [11], generalized additive models (GAM) [10], and the exponential family nonlinear models [25], which also includes methods using supervised learning that are routinely used in machine learning (ML) [7,26–32]. Discussions of current analytical methods are provided with recommendations and more advanced strategies are identified to provide a practical framework for developing improved statistical models.

As there is vast literature on these subjects, this is not a comprehensive discussion. Rather, our article provides an overview of methods with guidance and references underscoring the critical nature of considering model specification as part of the development process that: (i) describes model fit, model misspecification tests, and model selection tests, (ii) addresses data representation strategies, (iii) discusses automated model-building approaches and validation methods, and (iv) reviews predictive and classification measures. We generally focus our discussion on regression methods that are typically applied in practice [1,3,5–7,12,13,16,33–39] on complete data. For an overview of methods related to analyzing data sets containing missing values [40–44] the reader is referred to Zhou [45]. Further, while graphical methods [5–7,10,12,33,39,46–50] are important tools that are recommended as part of the model development process, they are not the focus of this article. Additionally, nonparametric statistical methods are discussed briefly as they pertain to univariate analysis in support of multivariate modeling. See, for example, Wasserman, Corder and Foreman, Hollander et al. for further information on nonparametric statistics [51–53]. Finally, issues related to causality and its relationship to model validity are not explicitly explored. However, see Kashner et al. [54], which addresses making causal inferences [55–57] using regression analysis methods.

This article is organized around strategies that are necessary to the development of correctly specified models that can evidence causal relationships and provide predictive performance. In general, we address the frequentist paradigm where emphasis is placed on statistical testing for model fit and misspecification including model development, classification and prediction. Bayesian methods for model averaging [58] and classification are also presented. Each section begins by stating the analytic challenges and explaining their significance, followed by recommended strategies to address those challenges. Section 2 discusses model fit measures, specification tests, model selection tests, and nonparametric tests. In particular, statistical methods used to establish external and internal validity [59,60] are addressed as they pertain to model development when dealing with the challenge of model fit and specification. Section 3 describes transformations and interactions

that may be applied for improving the representation of the DGP for subsequent multivariate analysis. Section 4 discusses automated model-building methods that include both single and multimodel [58,61–63] paradigms. Section 5 contains computation validation methods for establishing both internal and external model validity. Finally, sections 6 and 7 review prediction and classification measures for statistical modeling.

2. Model fit / model specification

2.1. Statement of challenge – model fit / model specification

Establishing model validity is challenging even when the researcher has a well-formed conceptual model of the outcome of interest and its relevant variables. There are different types of validity. Face validity addresses the issue of whether the model appears plausible. For example, within a regression modeling context, a model that indicates that mortality rate depends upon eye color might be considered to have low face validity even if the model makes useful predictions. Criterion-based validity is applied by the researcher to determine if a model behaves as expected. Typically, this is accomplished by comparing predicted outcomes of a model with established findings. External validity [64] addresses the robustness of the model's results [20,65,66] to determine if its inferences will generalize to related populations [12,67–70]. Internal validity addresses the issue of whether a model makes consistent predictions with respect to a particular data set which is presumed to be representative of the general population. Internal validity may be evaluated in predictive models by checking if the model is correctly specified, makes internally consistent predictions, and exhibits good performance using cross-validation simulation methods [20,71].

2.2. Explanation

Effect sizes are interpretable with respect to a specific statistical (probability) model of the data generating process. In some cases, these models may encompass large classes of models, while in other cases the set of models may be more restrictive. Valid modeling results and their interpretation depends on both the researcher's underlying assumptions for posited probability model with the application of the appropriate statistical methods. For example, using a Best Approximating Model (BAM) [72–76] approach to account for moderating variables and model misspecification, Westover et al. [74,75,77] reanalyzed data from a randomized, double blind, placebo controlled, intent-to-treat study of adults with Attention Deficit Hyperactivity Disorder (ADHD). This model development approach [75,76], which utilizes robust model search and specification analysis methods discussed in this article, allowed Westover et al. [74,77] to evidence significant changes in smoking cessation following the use of a stimulant medication osmotic-release oral system of methylphenidate, even though the original study indicated no treatment effect [78]. Further, simulation studies [79,80] have shown that there exist situations where models that indicate fit to the data using typical summary measures (§ 2.2.1), but do not otherwise represent the DGP may yield wrong conclusions.

Thus, to adequately assess the validity of the statistical (probability) model, researchers should examine both model fit and model misspecification [79–81]. *Model fit* measures the similarity of the 'fitted model' to actual outcome values generated by the DGP (§ 2.2.1).

Model misspecification indicates whether the choice of variables and how those variables are recoded or transformed are fit in a model that best approximates the DGP. A model that represents the DGP is called a ‘correctly specified model’. A model that can not represent the DGP is said to be misspecified. We test for failure to represent the DGP using model misspecification tests that test the null hypothesis the model is correctly specified (§ 2.2.4). In contrast to typically used summary level goodness-of-fit tests (GOF) [6,12], model misspecification tests are designed to examine various aspects of model misspecification [18,79–81].

Both the assessment of model fit and the possible presence of model misspecification are critically important and, while related, often capture quite different aspects of model validity. For example, suppose the researcher modeled risk factors that predict whether a sample of patients with depressive symptoms will attempt self harm within 30 days following a visit to a hospital’s psychiatric emergency department. Here, the limited dependent variable is the probability that a patient will attempt self-harm and enters into the model as a logit. In this situation the researcher may estimate parameter coefficients using a logistic regression that properly accounts for the binary outcome variable (attempt v. no attempt). The use of a simple linear regression to predict a binary outcome would correspond to a situation where the model was misspecified because linear regression assumes the outcome variable is continuous.

2.2.1. *Model fit measures*

Essential to data interpretation [79,80], model fit is a measure of the discrepancy between the observed empirical distribution of the observations in the data set and the ‘best-fitting’ probability distribution computed from the estimated probability model. Given the specification of a parameterized model and the data, model parameters may be estimated to fit the model. For example, maximum likelihood estimation methods seek the parameters that make the observed data most likely and may be interpreted as minimizing a cross-entropy discrepancy measure between predicted probabilities and observed relative frequencies.

The problem of assessing model fit is challenging when researchers want to measure fit that accounts for variability in model complexity, model misspecification, and small sample size. Examples of global or summary model fit measures that often used to assess overall model fit include sum of squared errors (SSE), log-likelihood (LL), as well model selection criteria. Table 1 contains examples of widely used model fit measures including information theoretic model selection criteria.

2.2.2. *Model specification analysis*

Model specification analysis [18,27,81,100–104] addresses the question of whether the researcher’s probability model represents the theoretically correct model for the DGP. Specifically, a correctly specified probability model has the property that it contains the true probability distribution that generated the observed data [18,79–81]. Statistical methods are available to assess model fit and misspecification. For example, graphical residual diagnostics [5–7,10,12,33,38,39,46–50,105] are useful for identifying the presence of model misspecification for the class of GLMs [11,106] and the larger class of exponential family nonlinear models [25]. Such diagnostics are also useful for identifying outliers that correspond to situations where the probability model is not applicable or the data sampling

Table 1. Examples of summary level model fit measures.

No.	Model fit measures	Description
1	Sum of Squared Errors (SSE)	Sum of squared differences (residuals) between predicted and observed values. Measures deviation from actual values [5,6,39].
2	R^2 , adjusted R^2 , Pseudo- R^2 Statistics	Coefficient of determination (R^2) compares the predictive performance of the model to a constrained version of the model [5,6,10,12,33,36,82,83].
3	Log-Likelihood (LL)	Kullback-Leibler based measure of model fit to observed data. Selects the model that makes the in-sample data (training data) most likely [12,38,84].
4	Akaike Information Criterion (AIC)	AIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the model that makes the out-of-sample data most likely. Assumes all models are correctly specified [18,61,80,81,85–87].
5	Akaike Information Criterion with finite sample correction (AICc)	AICc allows comparison between nested, overlapping, or nonnested models having different numbers of parameters, with small sample size correction. Selects the model that makes the out-of-sample data most likely. Assumes all models are correctly specified [61,87–89].
6	Bayesian Information Criterion (BIC), also know as the Schwarz Criterion (SIC)	BIC/SIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the most probable model given the data. Applicable for both correctly specified and misspecified models [61,86,87,90].
7	Bayesian predictive information criterion (BPIC)	Hierarchical modeling generalization of the AIC and BIC [91,92].
8	Generalized Akaike Information Criterion (GAIC)	GAIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the model that makes the out-of-sample data most likely. Applicable for both correctly specified and misspecified models [61,86,87,93,94].
9	Generalized Bayesian Information Criterion (GBIC)	GBIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters, with small sample correction superior to BIC. Selects the most probable model given the data. Applicable for both correctly specified and misspecified models [95–97].
10	Kullback Information Criterion (KIC)	KIC is an asymptotically unbiased estimator of Kullback's symmetric divergence that allows comparison between nested, overlapping, or nonnested models having different numbers of parameters [98,99]. Assumes all models are correctly specified.

process assumptions are severely violated. However, the determination of whether an outlier corresponds to a flaw in the probability model or an exception in the data sampling process may not always be resolved through inspection of graphical residual diagnostics. In such cases, the use of hypothesis testing approaches to outlier detection [107–109] and, more generally, specification analysis [18,81,110] may be more helpful than measures of fit that support model comparisons, but are not designed to detect misspecification in a particular model. Several specification tests for statistical models with continuous and categorical outcomes including newer Information Matrix Tests (IMT) [18,79–81] are further discussed with examples provided in Table 2.

Link specification tests [34,38] are applicable for testing the assumption of linearity in the link function (e.g. logit), however they are not designed to detect other types of model misspecification. Another widely used method is the encompassing model selection test, which is applicable when the researcher's model is fully nested within a highly flexible model capable of representing both the researcher's model as well as virtually any arbitrary probability distribution. An example is the Likelihood Ratio Test (LRT) [111] that compares an encompassing model with the researcher's model. The detection of a difference in model fit between the encompassing model and the researcher's model

Table 2. Examples of specification tests for assessing a model's representation of the DGP.

No.	Specification tests	Description
1	Box-Tidwell Test	Detects misspecification of linearity assumptions in logistic regression models [38].
2	Breush-Pagan Test	Detects presence of heteroskedastic residual errors in the linear regression model [138].
3	Durbin-Watson Test	Detects presence of autocorrelation (i.e. correlated residual errors) for regression models that assume uncorrelated residual errors [139,140].
4	Encompassing Model Selection Test	Detects model misspecification by using a correctly specified probability model which is sufficiently flexible for representing both the DGP and the researcher's model [18,141].
6	Generalized Information Matrix Test(s) (GIMT)	Detects model misspecification by comparing nonlinear combinations of two different covariance matrix estimators (Hessian, OPG) to test if the Fisher Information Equality holds for large class of smooth probability models [18,79–81].
7	Hosmer-Lemeshow Test	Detects deviation of observed frequency distribution from a theoretical (expected) probability distribution for logistic regression modeling [12,38,114,142].
8	Information Matrix Test(s) (IMT)	Detects model misspecification by comparing linear combinations of two different inverse covariance matrix estimators (Hessian, OPG) to test if the Fisher Information Equality holds for a large class of smooth probability models [18,79–81].
9	Ljung-Box Test	Detects presence of autocorrelation (i.e. correlated residual errors) for auto-regressive moving average (ARMA) models [143,144].
10	Pearson-chi-square χ^2	Detects deviation of observed frequency distribution from theoretical (expected) probability distribution [12,38].
11	Tukey-Preigibon Link Test	Detects misspecification of linearity assumptions in logistic regression models [38].
12	White Test (IMT)	Detects presence of heteroskedastic residual errors or residual error dependence on predictors or other types of misspecification in the linear regression model [100].

indicates the possible presence of model misspecification stemming from potentially omitted predictors. However, this strategy is limited because the resulting chi-squared statistic often has high degrees of freedom and therefore poor statistical power. For linear regression modeling, several specialized tests have been developed including the White Test for Heteroskedasticity [100], which is a special case of the IMT [81]. Additionally, specialized model misspecification tests have been developed to detect one or more aspects of model misspecification. For example, specification tests for time-series analyses have been developed [112].

Chi-square GOF tests [113,114] are utilized for probability models with categorical response variables. For example, the classic Pearson chi-squared GOF test may be used, which is asymptotically equivalent to an Encompassing MST using the LRT. In the case of logistic regression, a variety of model misspecification tests have been developed based upon the chi-squared GOF test [115,116]. A chi-squared GOF test [12,114,117] compares the predicted probabilities of a model to the observed frequencies. In regression modeling when the outcome variable is categorical, the conditional probability mass function of the outcome variable conditioned upon the predictors needs to be checked for model misspecification for unique patterns of predictor variable values. However, in practice such tests require that observations (exemplars) with similar predictor patterns be grouped together in order to avoid the problem of excessive degrees of freedom [12,118]. A drawback of this approach is that grouping methods are actually testing a 'grouped version' of the researcher's model rather than the researcher's original logistic regression model [114].

Generalized Information Matrix Tests (GIMT) [79,80,119–125] are nonlinear extensions of White’s IMT [18,81,126] that test for model misspecification and are applicable to a wide range of statistical models including generalized linear models (GLM) [11,37,127] and the family of nonlinear exponential models [25]. GIMTs are based on classical asymptotic statistical theory that assumes the probability model can adequately represent the DGP, implying that the Fisher Information Matrix Equality holds [18,79–81,128]. So, if the fitted probability model is correctly specified, then the inverse outer-product-gradient (OPG) and inverse Hessian estimates of the variance-covariance matrix (hereafter covariance matrix) of the maximum likelihood estimates (MLE) will be approximately equal corresponding to the case where the Fisher Information Matrix Equality holds. Note that OPG estimates of the covariance matrix are based on first derivatives of the likelihood function, while Hessian estimates use second derivatives. In the situation when these two covariance matrix estimators (CME) are not equivalent, the presence of model misspecification is indicated [18,81]. If the model is misspecified, then both of these formulas (Hessian, OPG) are incorrect for the covariance matrix of the MLEs and the robust sandwich CME [18,81] should be used instead. Different methods for comparing the Hessian and OPG CMEs correspond to different types of tests for detecting model misspecification in the GIMT framework [18,79–81].

An important advantage of GIMT methods over Encompassing MST methods is that the number of degrees of freedom is often manageable, which provides statistical tests with good level and power performance [79,80]. Another advantage is that the detection of model misspecification also involves identifying situations where the use of the sandwich CME is required. GIMTs have been applied to detect the presence of misspecification in regression models in both randomized controlled trials and observational studies. GIMTs and other GOF tests [114,117,129–136] are provided in Table 2. Validation methods [6,7,33,137] that also provide an opportunity for assessing both model fit and specification are discussed in section 5.1.

2.2.3. Model selection tests

Researchers often fit two or more models to the same dataset, and then test which of the estimated models has the best ‘fit’ to the dataset. A statistical test to determine the better fitting model is called a ‘Model Selection Test’ (MST) [86,111,145–148]. To support model development, such tests can be used for comparing the fit of competing models. These tests, however, have limitations. The Likelihood Ratio Test (LRT) [111] is widely used to determine whether a correctly specified ‘full model’ fits the DGP more effectively than a ‘reduced’ or ‘nested’ model as a special case of the full model. It is commonly used to compare the full model to the intercept only model as a test of overall model fit. A more robust approach to the LRT is the Bootstrap Likelihood Ratio Test (BLRT) [149], which uses an empirical estimate of the null hypothesis distribution. Generalizations of the LRT such as the Vuong’s Model Selection Test and Golden’s Discrepancy Risk Model Selection Test (DRMST) [145,147] may be used to compare models that are not nested or possibly not correctly specified [146,148]. Further, MSTs may also be used to determine optimal recoding or transformation strategies for a predictor that has been recoded or transformed before being fit in a model. This is done by comparing two regression models comprising the same variables, but where the model variables are recoded or transformed differently. Table 3 contains examples of model selection tests.

Table 3. Examples of model selection tests (MST) that compare competing model fits.

No.	Specification test	Description
1	Bootstrap Likelihood Ratio Test (BLRT)	Robust statistical test for comparing fit of possibly misspecified or nonnested statistical models [149–152].
2	Discrepancy Risk Model Selection Test (DRMST)	Robust statistical test for comparing possibly misspecified or nonnested time-series models. Model selection criteria such as AIC, BIC, or GAIC may be used to adjust models being compared. This is a generalization of Vuong's Model Selection Test [80,145,147].
3	Likelihood Ratio Test (LRT)	Statistical test for comparing fit of statistical models where the full model is correctly specified and the reduced model is fully nested within the full model [12,38,111].
4	Lo-Mendell-Rubin Likelihood Ratio Test (LMR)	Statistical test that allows comparison between models that have different numbers of components in a normal mixture. This is a special case of Vuong's Model Selection Test [153].
5	Rivers & Vuong Nonlinear Dynamic Model Selection Test	Statistical test for comparing possibly misspecified and essentially nonnested time-series models using a variety of MSC (e.g. AIC, BIC, GAIC). This is a generalization of Vuong's Model Selection Test [148].
6	Vuong's Model Selection Test	Robust statistical test for comparing possibly misspecified or nonnested models. Model selection criteria such as AIC or BIC may be used to adjust models being compared [146,148].

2.2.4. Nonparametric tests

Nonparametric tests may be used to check selected properties of the DGP without postulating an explicit probability model. Nonparametric tests are often preferred to parametric tests because the former relies on weaker assumptions regarding the underlying probability distribution [51,53,154–157]. Nonparametric test statistics are useful when: (i) the data is continuous and the distribution is unknown, (ii) the data is ordinal, ranked or categorical, (iii) sample size is very small, or (iv) reliable models of the occurrence of outliers [108,109,158,159] are not available, and little is known about the relationship among and between predictors and outcome variables. Nonparametric tests can facilitate better understanding of the variables that are used in subsequent multivariate modeling. For a broader discussion of nonparametric statistics see Wasserman, Corder and Foreman, Hollander et al. [51–53]. Table 4 contains examples of some nonparametric tests.

Table 4. Examples of nonparametric tests.

No.	Statistical test	Description
1	Anderson-Darling Test	Detects if empirical distribution function estimated from the data sample differs from null hypothesis distribution function [160].
2	Kolmogorov-Smirnov Test	1. Detects if the empirical distribution function estimated from the data sample differs from null hypothesis distribution function. 2. Detects if two empirical distribution functions estimated from two data samples are different [14,51,161–164].
3	Shapiro-Wilk Test	Detects violation of assumption that observations are from a normally distributed population [165].
4	Spearman's rank correlation coefficient	Detects if two random variables are correlated without assuming a specific joint distribution for the two variables [14].
5	Tau Test based on Kendall rank correlation coefficient	Detects if two random variables are correlated without assuming a specific joint distribution for two variables [14,166,167].
6	Wilcoxon signed-rank Test	Detects if median difference between pairs of observations from two respective data samples differ from zero. Does not assume a specific joint distribution for pairs [14].

2.3. Recommendations

The challenge in performing model fit and model specification analyses is to determine and apply the appropriate statistical methods for the researcher's statistical model. These methods should be identified as part of the researcher's data analysis plans [168] where the type of model summary fit measures and significance of regression tests, as well as the summary GOF tests and model misspecification tests available for data analyses, depend on the particular statistical modeling approach being applied. When regression modeling paradigms (e.g. linear regression, nonlinear regression, categorical regression, etc.) are employed, the following is recommended.

Recommendations: Model Fit and Model Specification Testing

1. *Minimum Acceptable Statistical Modeling* – A summary measure and model fit test (Table 1) as well as a summary model specification (goodness-of-fit) test (Table 2) are performed on the researcher's final model. Results also include graphical diagnostic plot(s) [6,7,12,38,46,169]. *Researcher* specifies the fitness measure, fit and specification (goodness-of-fit) tests with the graphical diagnostic plot(s) for the statistical model [6,7,12,38] in the data analysis plan [168]. The model selection criteria for comparing or testing competing models (Table 3) are included in the plan. The researcher must also justify why the proposed model fit and model comparison plans are appropriate for the planned model comparisons proposed in the research plan.
 2. *Advanced Statistical Modeling* – Summary tests for model fit (Table 1) and a battery of specification tests (see Table 2) [6,7,12,38] are performed to check different aspects of possible model misspecification for the researcher's final model. This should be accompanied with multiple graphical diagnostics for both predictive results and residual plots. Additionally, computational specification analysis methods such as nonparametric bootstrapping or split-sample methods are performed to evaluate model fit and specification and employed. *Researcher* specifies the fitness measure, fit and specification (goodness-of-fit) tests with the graphical diagnostic plots for the statistical model [6,7,12,38] in the data analysis plan [168]. Further, the model selection criteria for comparing or testing competing models (Table 3) and computational analysis methods are included in the plan.
-

3. Data representation strategies

3.1. Statement of challenge – choice of variables, transformations, and interactions

To create a probability model of the response outcome, analysts must identify the relevant variables in the dataset and then transform or recode [6,7,12,31] them into covariates (predictors) to develop a model that can represent the true DGP [18].

3.2. Explanation

In classical asymptotic statistical theory, analysts usually assume that their probability model adequately represents the DGP. As previously discussed, the Fisher Information Matrix Equality holds in this situation [18,81,128]. When this equality is violated, in practice this indicates: (i) a data field, containing a variable, was not considered or is not in the dataset, (ii) the data variable was considered, but not properly recoded (discrete, ordinal)

or transformed (continuous, ordinal) into a covariate, (iii) an important interaction among two or more covariates was not included, and/or (iv) the particular form of the response variable distribution assumed by the statistical analysis has major (or minor) discrepancies.

3.2.1. *Plausible variables*

Analysts often begin with a dataset containing raw fields that must be scored, indexed, and missing data recoded into research-ready variables appropriate for data analyses [170,171]. Plausible variables are those available research-ready variables that the analyst considers relevant for: (i) the specific study purpose, such as estimating effect sizes, testing hypotheses, confirming theory, or estimating risks of an occurrence or outcome event, (ii) study design, (iii) prior experience including literature reviews of prior studies, and most importantly, (iv) theory, including all variables derived from competing theories. Plausible variables may also be assessed empirically using univariate regressions [1,7,12] to determine separately if each variable is associated with responses.

3.2.2. *Covariates: recoding and transformation*

To determine whether a plausible variable from the dataset should be transformed or recoded before entering into the model, the analyst may fit the response variable in separate univariate regressions on each variable with the intercept term, and then test the estimated single variable model for both model fit (Table 3) and model misspecification using tests identified in Table 2. A single predictor model with poor model fit indicates the model's predictor may be less important in generating predictive model responses. Further, the presence of model misspecification in a single predictor model indicates the probabilistic relationship between the response variable and predictor variable is misspecified and suggests alternative representations or transformations of the predictor should be considered. Plausible variables may be transformed or recoded to candidate covariates for subsequent multivariate modeling as follows. Categorical (nominal) and ordinal variables may be recoded by collapsing (pooling) or disaggregating categories using reference cell coding [12]. Continuous variables may be transformed using: (i) power transformations [172], (ii) fractional polynomial methods [9], (iii) spline methods [6,7], (iv) kernel smoother methods [7], or (v) splitting the ordinal / continuous plausible variable into consecutive ranges where each is represented by a binary design variable. Thus, age could be partitioned into ranges: < 18 years, 19–30, 31–45, 46–64, and 65+, with respective binary design variables that assume a value of one if the subject is within the age interval, and zero otherwise [7,12]. As previously discussed, robust model selection tests (e.g. Vuong's MST [146], Golden's DRMST [145]) which allow comparing non-nested models may also be used to determine an optimal recoding or transformation for a variable by comparing models fitted with its alternate recodings or transformations. All variable recodings / transformations are thus performed before being entered into the candidate covariate pool.

3.2.3. *Interactions*

Interaction terms are computed by multiplying two or more covariates together. Interactions are used to address specification issues (i.e. GOF), improve model fit [175], model relationships between covariates during model development [6,176–180], and to assess how the effect of one covariate on the outcome variable varies with another covariate

[6,9,38,175,181]. An interaction may describe either a causal or non-causal relationship. In this situation interactions may be specified or may be driven by model selection. Each interaction is then treated as an additional predictor, which is included in the candidate list of variables for the modeling procedure. The issue goes beyond merely attempting to measure the independent effect of two interacting factors on outcome. For example, in the DGP that explains the onset of a given disorder, an ‘older female’ patient may not necessarily be only the product of an age and gender factors, but rather the characteristic of being an older-female and other predictors such as family history of disorder. Interactions may also be used to investigate hypothesized moderated effects [74,182] and to perform before and after studies [174].

3.3. Recommendations

The challenge is in determining variable transformations and specifying interactions for the researcher’s statistical model. These approaches should be identified as part of the researcher’s data analysis plan, with the type of transformations and interactions of interest for data analyses dependent on the particular statistical modeling being utilized. The following is recommended.

Recommendations: Transformation / Interactions

1. *Minimum Acceptable Statistical Modeling* – Analysis should be performed on a single variable model which predicts the response variable for each variable to determine fit, specification, and predictive performance. Transformations that improve fit (e.g. power transformations [172], fractional polynomial methods [9], spline methods [6,7], kernel smoother methods [7]) and interpretability (e.g. median dichotomization) are applied as required. Interactions are specified as combinations of variables from the pool of potential covariates. *Researcher* specifies the relevant types of variable transformations and interactions for the statistical model in the data analyses plan and explains why they are relevant. Multicollinearity diagnostics [6,12] are identified.
 2. *Advanced Statistical Modeling* – Transformation analyses (when applicable) that include supervised transformations (e.g. fractional polynomials, cubic splines with free knots) and multiple predictive measures (e.g. AIC, GAIC) in conjunction with specification testing and model validation methods (e.g. KCV, bootstrapping) should be performed to identify opportunities to improve variable fit and predictive performance for each variable. Interactions may be specified or data-driven. *Researcher* specifies all relevant model prediction measures, transformations, interactions, and validation methods in data analyses plan and explains why they are relevant. Multicollinearity diagnostics are identified.
-

4. Model search strategies

4.1. Statement of challenge – model search

Researchers often analyze datasets without having clear theoretical or empirical guidance from prior studies as a basis for selecting one model. Compounding this source of uncertainty [183–185] are issues associated with theoretically related variables that are highly collinear [186], thus leaving the researcher to select among competing models. Further,

there are no assurances that a single model exists that [187] adequately represents the DGP [96]. In such situations utilizing automated model-building strategies can be supportive in this regard when a plausible set of covariates consisting of recoded or transformed variables as well as interaction terms that have been identified by the researcher(s) to be within the design framework. However, model(s) resulting from automated model-building methods also require validation and specification testing as well as critical scientific review. The following automated model-building discussion assumes that the number of observations is larger than the number of variables. Thus, statistical methods [32,182–188] for developing models from datasets with small numbers of observations that have thousands to hundreds of thousands of available variables are not addressed [32,188–194].

4.2. Explanation

Typically, a regression model [6] may be specified by a subset of predictors (covariates) from the original sample where each predictor subset corresponds to a particular regression model. Using a data-driven approach, the model building problem may then be reformulated as a search for the model that provides the best possible fit to the observed data. Such model(s) may exhibit improved predictive accuracy and interpretability. There are multiple approaches to model search for a single best model [195]. More advanced approaches that address the uncertainty inherent in the single model (SM) paradigm use multimodel (MM) methods that search for a collection of models [58,61,63]. Table 5 shows a comparison of SM and MM model search methods, which are discussed in the following sections.

4.2.1. Single model (SM) methods

Stepwise Regression. There are a variety of stepwise regression methods [197] that are used to find a best model from among a set of candidate predictor variables. The guiding principle of stepwise regression models is that the search algorithm is constantly comparing one model that is nested within another model. This places a special collection of ‘nesting constraints’ on the search algorithm procedure that severely limits the space of possible models that the analysis will consider as a final model. Such nesting constraints on the model space restrict the stepwise procedure’s capacity to yield a final model that approximates the DGP. Although these methods are widely used even when the number of predictors is small, SM stepwise regression methods are inferior to exhaustive search (all subsets) methods, especially in the presence of multicollinearity among covariates and in higher-dimensional spaces when stepwise methods may become trapped in local maxima [206]. Further, as complex patterns of features increase in the presence of multicollinearity, stepwise methods may not provide adequate representations of the model space.

Forward stepwise regression begins by regressing the outcome variable separately on each variable in the predictor pool, and comparing its fit to that of the intercept-only regression model using the Likelihood Ratio Test (LRT) or the Wald Test (with the Hessian CME). The predictor corresponding to the model with the largest likelihood relative to the intercept-only model is then selected and added into the final model. This process is

Table 5. Comparisons of model search methods.

Model search methods ^a	Seeks optimal model(s)	Robust for comparing misspecified ^b models	Accounts for multicollinearity ^c	Controls for model uncertainty
Single Model (SM) approach ^c				
Stepwise Regression using Likelihood Ratio Test (LRT) [196,197]	<i>No, solution set is algorithm dependent</i>	<i>No, LRT assumes correct specification</i>	<i>No, typically does not check for multicollinearity.</i>	<i>No</i>
Exhaustive Search (all subsets) with MSC [197]	<i>YES, solution set is algorithm independent</i>	<i>YES, provided robust MSC used</i>	<i>Yes, if multicollinearity is checked.</i>	<i>No</i>
Implicit (e.g. B&B) Exhaustive Search [198–201]	<i>YES, solution set is algorithm independent under certain conditions</i>	<i>YES, provided robust MSC used</i>	<i>Yes, if multicollinearity is checked.</i>	<i>No</i>
Stochastic Search with MSC [202–205]	<i>YES</i>	<i>YES, provided robust MSC used</i>	<i>Yes, if multicollinearity is checked.</i>	<i>No</i>
Multimodel (MM) approach				
Frequentist Model Averaging (FMA) [61–63,244–247]	<i>YES, solution set is algorithm independent</i>	<i>YES</i>	<i>Yes, if multicollinearity is checked.</i>	<i>YES</i>
Bayesian Model Averaging (BMA) [58,183,226,230,236,242,243]	<i>YES, solution set is algorithm independent</i>	<i>na</i>	<i>Yes, if multicollinearity of models that are averaged is checked.</i>	<i>YES</i>

^aThe use of stepwise regression methods is not recommended for final model determination. Exhaustive search, branch and bound search of the entire model space, or stochastic search of the model space in conjunction with model validation and specification testing are preferred.

^bModel selection criteria [61,93,94,96] (e.g. GAIC, GBIC) designed to assess out-of-sample performance that are robust to model misspecification [18,81] may be utilized by search algorithm.

^cModels may be filtered for multicollinearity based on a specified threshold for large variance-covariance matrix condition numbers. Ridge regression methods [197] may also be used to control for multicollinearity.

repeated for each of the remaining $p-1$ predictors in the predictor pool until the additional predictor produces no significant increase in likelihood of the final model.

Backward elimination stepwise regression begins with a regression model consisting of all p predictors and an intercept is compared to a nested regression model consisting of $p-1$ predictors using either the LRT or Wald test. The predictor that is least significant is dropped from the model and the process is repeated until dropping a predictor from the model significantly reduces the likelihood. Backward stepwise regression methods are intended to circumvent the problem of omitting crucial predictors from the model yet still suffer from all of the problems of forward stepwise regression as well as the additional problem of multicollinearity (i.e. making statistical inferences from a model containing highly correlated predictors), which may result in unreliable statistical inferences.

Forward–backward stepwise regression is an important variant of stepwise regression [197]. As described in Efronson [196], forward–backward stepwise begins as forward stepwise regression. However, in subsequent steps, after a new predictor is added to the model the statistical significance of all predictors currently in the model is checked and predictors that are not contributing in a statistically significant manner to the predictive performance of the model are dropped.

Stepwise approaches may be used for exploratory purposes, but are not recommended to create final models for subsequent analysis. First, forward stepwise regression methods may omit the inclusion of important predictors since the ordering in which predictors are considered for model inclusion can substantially influence the final form of the model discovered by such methods. Second, in order for the Likelihood Ratio Test (LRT) or the Wald Test (with Hessian CME) to be applicable, it is assumed that the ‘full’ model is correctly specified. This assumption is often violated during the course of stepwise model building. Third, the experiment-wise Type 1 error rate in a forward stepwise regression problem is essentially meaningless, which means that some model comparisons will inevitably be rejected by chance. A fourth problem in stepwise approaches is the limitations of LRT and Wald test methods. Under the null hypothesis, the test statistic for the LRT is assumed to have an asymptotic chi-square distribution, but only if the model is correctly specified. This becomes a problem for the researcher as possibly misspecified models are frequently compared during the stepwise model building process. In fact, this is a pitfall of many automated model building approaches. Further, the LRT compares only nested models and requires the full model to always be correctly specified, thus making the use of LRT inappropriate for ‘exhaustive searches’ where all possible pairs of models are compared since situations where at least one model is misspecified are common.

Exhaustive Search. The exhaustive model search method (i.e. all subsets) is a straightforward model search approach for finding a single best fitting model that is appropriate when the number of models to be searched is computationally tractable [197,198,207,208]. In this method, the likelihood of every possible model (i.e. all subsets of predictors) given the observed data is estimated and the model with the best fit is selected. The exhaustive model search is a powerful approach when employed with a model selection criterion (e.g. AIC, GAIC, BIC) and validation methods, but depending on the number of predictors may be computationally limited. If there are p plausible predictors to explain an outcome variable based on accepted theory (including all covariates obtained after appropriate transformations have been applied), then there are 2^p subsets of predictors (including intercept-only model) that must be considered in searching for the best fitting model. There are computational algorithms that search exhaustively over as many as $p = 30$ predictors (2^{30} models), however specialized algorithms that perform implicit exhaustive searches are required for larger model spaces (e.g. $p = 60$) [196–199,206–210] and the topic continues to be a subject of research [209,210]. Exhaustive search methods can also be applied to rapidly find groups of ‘best’ models that support multimodeling.

Exhaustive Search (Implicit). When the number of predictors is large, then an exhaustive search of all possible subsets of predictors may not be computationally feasible. In this situation, Branch and Bound (B&B) search algorithms may be used to find subsets of predictors that best fit the observed data while avoiding an exhaustive search [197,199,200,211–218]. The approach that B&B methods use is to construct a search tree and then identify branches of the tree that can be ‘pruned’ in order to significantly reduce the size of the search space. Such methods have been shown to find the best model as effectively as an exhaustive search approach [200]. In many cases, dramatic gains in computational efficiency may be realized using these methods [200,212]. Implicit exhaustive search methods such as B&B can also be applied to rapidly find groups of ‘best’ models

that support multimodeling. However, as with other search methods, there is no assurance that the final model found using B&B is correctly specified.

Stochastic Search. Stochastic search methods such as Markov chain Monte Carlo (MCMC) [219–223] can often rapidly find ‘good’ models. However, such algorithms designed to explore model spaces rely on Gibbs sampling [224] or on the Metropolis-Hastings algorithm [204,225] and may be ineffective in very high dimensions with hundreds or thousands of predictors due to slow convergence. New model search methods that draw from MCMC but exploit simultaneous model search using parallel methods [202] have been shown to be effective when searching over larger model spaces. Stochastic search methods such as MCMC can be applied to rapidly find groups of ‘best’ models that support model averaging methods [58].

4.2.2. Multimodel (MM) search and averaging methods

A search for the best model to approximate the DGP assumes that a ‘best’ model exists, is discoverable, and is distinguishable from second best alternative models. While single model estimation and inference is widely used in health-related studies, such approaches neglect model uncertainty that arises when the researcher’s model deviates from other observationally equivalent models that approximate the DGP. Accounting for model uncertainty allows researchers to: (i) detect additional statistical regularities (e.g. treatment effects, risk factors) among groups of observationally equivalent models, (ii) improve the precision of statistical inferences for estimation and prediction / classification, (iii) control for overfitting and model selection biases, and (iv) include a much larger set of highly correlated, multicollinear risk factors in the data analysis plan that would impose estimation obstacles if they were all to appear in a single model.

MM search methods involve several steps. First, a model space is specified based a set of candidate predictors as selected by the researcher based on prior knowledge of what variables should be included in the model. In many applications, this model space can be large. In such cases, combinations of deterministic and stochastic single model search methods may be employed to reduce the number of models in the original model space to a model space with a small number of models which exhibit good predictive performance. And second, MM effect sizes and predictions across all fitted models in the model space are averaged through a weighted averaging scheme. Such model averaging methods have been shown to be both theoretically [58,226,227] and empirically [228–241] superior to SM inference methods applied to health-related data analyses [228–233, 235,239–241].

MM Search Methods. Exhaustive model search is a straightforward model search approach for finding a subset of best approximating models, which can be utilized when the number of models to be searched is computationally tractable [197,198,207,208]. An exhaustive search is appropriate when the researcher has identified approximately 30 or fewer predictors based upon theory-driven considerations. In this situation the likelihood of every possible model given the observed data is estimated and either all the models or the models within a specified neighborhood of the ‘best’ model are used for multimodeling. When used in this manner, exhaustive search can be powerful tool for quantifying model uncertainty. If the number of predictors is large, then an exhaustive search of all possible

subsets of predictors to support a multimodeling approach may not be computationally feasible. In this situation, B&B search algorithms, which implicitly search all models, may be used to find best subsets of predictors that fit the observed data, thus avoiding an exhaustive search [197,199,200,211–215]. As discussed B&B [200] algorithms construct a search tree and then identify branches of the tree that can be ‘pruned’ in order to reduce the size of the search space. The models obtained from B&B are then used for multimodeling. Previously discussed stochastic search methods such as Markov chain Monte Carlo (MCMC) [202,219–223] can also be applied to rapidly find groups of ‘best’ models that support model averaging methods [58].

Bayesian Model Averaging (BMA). BMA is a multimodeling method for estimation and inference that deals with model uncertainty [58,183,226,230,236,242,243]. It is appropriate when the researcher has good prior knowledge of both model and parameter specification. Given a large collection of models which defines a ‘model space’, methods such as exhaustive model search seek a single best model in the model space. In contrast to this approach, BMA approaches generally focus upon using the ‘most probable’ models within a constrained model space by applying Occam’s Window [58] to identify a group of best models, rather than searching through, evaluating each, and finally selecting the best models from among all possible models in what otherwise may be a computationally intractable model space. The essential idea of BMA is that one uses the fit of each model to the data to estimate the expected response given all of the models and the data. This is achieved by defining the predicted response as a weighted sum of the responses of all of the models in the model space where the weight of each model is the estimated probability of that model given the data and specified prior knowledge. Thus, a prediction is generated by a weighted consensus of multiple models rather than attempting to select one out of many models.

Frequentist Model Averaging (FMA). FMA [61,63,244–247] is a relatively new multimodeling approach for dealing with model uncertainty that addresses several issues associated with Bayesian methods [58,183,226,230,236,242,243]. In particular, FMA doesn’t require prior distributions be specified for either predictors or models and permits flexibility in weight choice for FMA estimators [246]. Both BMA and FMA approaches often use a constrained model space by applying Occam’s Window [58] to identify a group of best models. However, in computationally tractable model spaces the FMA method may use all possible models. The essential idea of FMA is that a predicted response, as in BMA, is defined as a weighted sum of the responses of all of the models in the model space, but arbitrary weighting functions can be used to specify the relative importance of each model.

4.3. Recommendations

The challenge in determining model search methods for the researcher’s statistical modeling depends on multiple factors including: (i) posited probability model (e.g. linear regression, logistic regression), (ii) number of covariates, and (iii) sample size. The model search approach should be identified as part of the researcher’s data analysis plan, with the type of transformations and interactions of interest for data analysis dependent on the particular statistical modeling being utilized. The following is recommended.

Recommendations: Model Search

1. *Minimum Acceptable Statistical Modeling* – Single Model search methods (when applicable) that include model validation (ref § 5) are included in the researcher’s data to determine the best model that meets the stated fitness criteria. Note that stepwise regression methods are not recommended for final model development, although they can offer computationally tractable approaches for initial predictor screening and exploratory modeling. Further, many automatic search methods typically do not control for multicollinearity and may also find final models that are misspecified. *Researcher* specifies the candidate variables, transformations, and model search method and validation method in the data analysis plan. Final model(s) are tested for model misspecification.
 2. *Advanced Statistical Modeling* – Multimodel search methods that include model validation are performed on the researcher’s data to determine the best model that meets the stated fitness criteria. Both BMA and FMA methods handle model uncertainty. However, FMA provides the hypothesis testing framework for a final multimodel that is widely used in health-related research. Additionally, as opposed to BMA, the FMA approach does not require prior knowledge (parameter prior, model prior) be specified. *Researcher* specifies the candidate variables and model search method and validation method in the data analysis plan.
-

5. Model validation strategies**5.1. Statement of challenge – model validation**

Model validation methods [7,137,248] include three components to evaluate model performance and assess biases: (i) model fit, (ii) model specification [18,80,81], and (iii) predictive / classification performance [6,7]. The purpose is to determine whether a model, fitted using one dataset, will perform well when applied to new data. Model validation [7,137,248] is thus an essential component of the model development process.

5.2. Explanation

There are a variety of possible model validation methods. *First*, a well-known approach often applied with large datasets is split sample methods [6,7]. Here, the sample is randomly divided into a training dataset that is used to develop the model, and a test data set that is used to validate the fitted model. When a sufficiently large number of observations are present in the dataset then a 3-way split may be used to generate training, validation, and testing samples. This approach develops the statistical model on the training sample, evaluates it on the validation sample allowing possible changes, and then generates the final model once on the test sample [7,33]. The major limitation of split sample methods is that sufficient data may not be available. *Second*, the cross-validation (CV) methodology [250,251] known as k -fold CV [6,7] is an alternative to the classical split sample method when sample size is limited. Often applied when the researcher believes the observations (exemplars) in the dataset are independent and identically distributed (i.i.d.), the k -fold cross validation randomly divides the n total observations (exemplars) into k groups (folds) (e.g. $k = 10$) of observations, and the model is fit using $k-1$ groups

of observations to estimate model parameters and then evaluated on the remaining ‘out-of-sample’ group of observations. This process is repeated an additional $k-1$ times so that all k groups participate as ‘out-of-sample’ groups. The estimated performance statistics averaged across all k groups measures the precision of the predictions in a realistic data sample. *Third*, an approach called leave-one-out cross-validation (LOOCV) [6,7] can also be used to evaluate statistical models. It leaves out one observation from n observations and then fits a model on the remaining $n-1$ observations, which is then used to predict the holdout observation response. The observation is replaced and a subsequent observation is removed to repeat the estimation and prediction. This is performed until all n observations have been processed as holdouts. The predicted results are then averaged to compute performance measures. *Fourth*, bootstrapping or resampling [6,7,252] methods involve sampling observations with replacement m times from the original dataset of n observations, to create m new datasets where each new dataset consists of n observations. These m new datasets are called *bootstrap samples*. For each bootstrap sample, the model is fitted and the mean and standard error (sample standard deviation) across all m bootstrap samples are computed. Ideally, the number of bootstrap samples m should be chosen sufficiently large so that the computed error of the mean standard error converges to zero.

5.3. Recommendations

The challenge in performing model validation is to determine and apply the appropriate validation methods for the researcher’s statistical model. These method(s) should be identified as part of the researcher’s data analysis plan, with the type of validation for data analysis based on the particular statistical modeling being utilized. When regression modeling methods (e.g. linear regression, nonlinear regression, categorical regression, etc.) are employed, the following is recommended.

Recommendations: Model Validation

1. *Minimum Acceptable Statistical Modeling* – Model validation is performed on the researcher’s final model to determine whether it will generalize to new data. Computational methods (e.g. k -fold CV, bootstrapping, LOOCV, split-sample) are generally the preferred methods for model validation. Alternatively, in some situations that may depend on sample size, model complexity, and data quality, a model selection criteria (e.g. AIC, AICc, GAIC) that estimates out-of-sample performance can be utilized to select the best model [61]. *Researcher* specifies the model validation method (e.g. k -fold CV, bootstrapping, LOOCV, split-sample) for the statistical model in the data analysis plan. Model validation measures (e.g. predictive performance, model fit) are included in the plan.
 2. *Advanced Statistical Modeling* – Multiple model validation methods [6,7] are performed to evaluate different aspects of final model fit, predictive performance, and generalization to new data. k -fold or split-sample CV methods are repeated at least 10 times on sample randomizations and results are then averaged. *Researcher* specifies the model validation methods (e.g. k -fold CV, bootstrapping, LOOCV, split-sample) for the statistical model in the data analysis plan. All model validation measures (e.g. predictive performance, model fit) are included in the plan.
-

6. Prediction

6.1. Statement of challenge – prediction

A common problem faced by epidemiological and health care researchers is making effective predictions from their models [6,7]. A widely used approach for making predictions is performed using a fitted statistical model to predict out of sample values [8,12,26,33,253–256]. The use of correctly specified probability models [18,80,81] supports robust predictions.

6.2. Explanation

Predictive performance for statistical models are computed using a variety of measures [6,7,33]. Table 6 shows examples of predictive measures for linear and nonlinear regression models with continuous dependent variables.

6.3. Recommendations

The challenge in reporting predictive performance is to determine and apply the appropriate measure for the researcher’s statistical model. These measures should be identified as part of the researcher’s data analysis plans where the type of predictive measure for data analyses depends on the particular statistical modeling being utilized. When regression modeling analysis methods (e.g. linear regression, nonlinear regression) are employed, the following is recommended.

Recommendations: Predictive Performance

1. *Minimum Acceptable Statistical Modeling* – Predictive analyses specifies error measures and model validation methods that are performed on the researcher’s final model to determine whether the model is likely to accurately predict responses. *Researcher* specifies and justifies the model prediction measures and validation method for the statistical model in the data analysis plan.
 2. *Advanced Statistical Modeling* – Predictive analyses include multiple predictive measures with model validation are performed to evaluate different aspects of final model fit and predictive performance. Multiple models are averaged to obtain better predictive performance. *Researcher* specifies and justifies multiple model prediction measures and validation methods in data analysis plan.
-

7. Classification

7.1. Statement of challenge – classification

Categorical regression models [1,3,6,12] predict probabilities that may be used to make classification decisions. Epidemiological and health care researchers often apply decision or allocation rules [256] to probabilistic modeling results in order to make classifications. The researcher first estimates the parameters of a probability model, which may then be used to estimate the predicted probability that an event of interest occurs, such as a disorder onset, side effect symptoms, or readmission. A classification rule is applied to allow the

Table 6. Examples of predictive measures.

No.	Predictive measures	Description
1	Mean Absolute Deviation (MAD)	Average of absolute deviations from a measure of central tendency (e.g. mean, median, mode) [257].
2	Mean Absolute Percentage Error (MAPE)	Measure of predictive accuracy expressed as a percentage of the average absolute error between predicted and observed values [258].
3	Mean Squared Predicted Error (MSPE)	The mean squared differences between predicted and out-of-sample observed values [5,7,33].
4	Predicted Residual Sums of Squares (PRESS)	Summary measure of the predicted residual errors between the predicted and out-of-sample values [259,260].
5	Residual Sum of Squares (RSS)	Sum of the squares of residuals, also known as SSE [5–7,33].
6	Root Mean Squared Error (RMSE)	Square root of the average of the squares of residuals [5,33].
7	R-squared (R^2)	R-squared (coefficient of determination) is the proportion of the variance of the dependent variable explained by the regression model [5–7,33].
8	Sum of Squared Errors (SSE)	Sum of squared differences between predicted and observed values [7,33].

decision maker to interpret the estimated probability as an outcome. One possible classification rule or strategy, known as the minimum probability of error (MPE), is to simply decide the patient has the disorder if the estimated probability the patient has the disorder is greater than the estimated probability that the patient does not have the disorder. However, in many health-related decision scenarios selecting the ‘most probable’ outcome may be inappropriate: when the cost of wrongly declaring the patient has a disorder is high (false positives), such as with expensive and potentially harmful treatments; or when the cost of wrongly declaring the patient healthy is high (false negatives), such as with many cancer diagnoses where efficacious treatments often require early detection and treatment during early stages of the disease; or when the researcher desires to maximize both sensitivity and specificity [261]. Further, such classification decisions require incorporating additional information to select an optimal outcome, which differs from applying the MPE rule. In this situation a classification [12,253–256,262] threshold may be utilized to predict a probability that an event will occur. In this approach, a specified value ranging from 0 to 1 (‘cut value’) is compared with estimated posterior outcome probabilities from the fitted model to classify the individual into a response category. The model’s predicted probability of the event is assigned to a positive response category (1) if it is above the decision threshold value, or to a negative response category (0) if it is at or below the threshold value. Thus, the researcher translates a predicted probability that an event will occur into the binary value that the event occurred or did not occur (e.g. inferring from the predicted probability the patient has a given disorder into whether the patient is considered to meet or not meet criteria to have the disorder). These results are commonly summarized via a classification table that displays the results of cross-classifying the actual binary outcome variable with the dichotomized response variable whose values are derived from the estimated logistic probabilities [12,33,261] shown in Table 7.

In some situations small improvements in the estimation of outcome probabilities can have significant consequences in terms of medical decision making and health care resource allocation. For example, classification decisions regarding the likelihood of rare events such as heart trauma or suicide must be as optimal as possible. Further, MPE decisions as well as more sophisticated decision-making strategies that attempt to improve

Table 7. A classification table for a binary logistic regression model depicts results for actual (observed) versus dichotomized predicted responses for diabetes^{a-d}.

Predicted	Observed		Total
	Present	Absent	
Present	183	60	243
Absent	85	440	525
Total	268	500	768

^aOutcome (diabetes/none).

^bAccuracy = 81.1%, Sensitivity = 68.28%, Specificity = 88.00%. Positive Predictive Value = 75.31%, Negative Predictive Value = 83.81%.

^cAUROC = .8726, 95%CI [0.8437, 0.9016].

^dPima Indians dataset (768 subjects), UCI Machine Learning Repository [263].

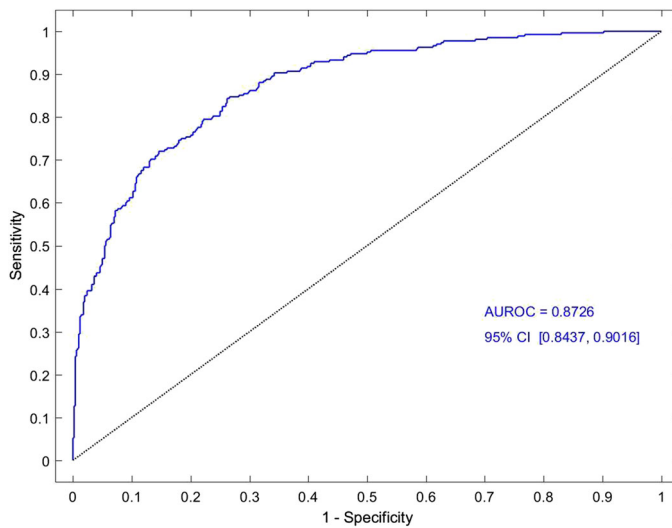


Figure 1. Receiver Operating Characteristic (ROC) curve for a binary logistic regression model depicts classifier performance for diabetes [263] prediction.

sensitivity and/or specificity [12,33,261] are dependent upon the calculation of decision threshold (cut value), which in turn are dependent upon the accuracy of outcome probability estimation. This necessitates that the statistical model which is generating the probabilities be correctly specified [18,79–81] (i.e. well-calibrated) in order to support optimal decision making that may also incorporate other factors. The receiver operating characteristic (ROC) curve [12,33,261,264–268] depicts the measure of a classifier’s (model) predictive performance across values of its discrimination threshold. It shows the true positive rate against the false positive rate across all threshold values. Figure 1 depicts a ROC curve with the area under the ROC curve (AUROC) [261,268,269] computed for the statistical model classification results shown in Table 7.

7.1.1. Bayesian classification

When a researcher is faced with situations where multiple mechanisms are required to represent the DGP [58,183,184,226,230], or involving rare events and small sample sizes

[270], then an approach to outcome probability estimation called the ‘Bayesian predictive density’, should be considered. Given a collection of fitted models the Bayesian [271] predictive density method estimates a conditional outcome probability for each model given a particular observation. Next, the set of estimated conditional outcome probabilities are combined using a weighted average where the weight associated with a particular conditional outcome probability corresponds to an estimate of the probability (or ‘preference’) for that model given the observed data. This model averaging algorithm not only has the potential of improving outcome probability estimates for rare events and small samples by looking for consensus among a large group of fitted models, it also has the potential to handle situations where outcome probabilities for two different subsets of observations are best estimated by using two disjoint subsets of fitted models.

Theoretical analyses [272–275] and empirical studies [270,272–275] have shown that when using the Kullback-Leibler Information Criterion (KLIC) cross-entropy measure as the measure of predictive performance, the Bayesian predictive density outcome probability estimator is superior in the estimation of outcome probabilities to the maximum likelihood outcome probability estimator methodology in situations involving rare events and insufficiently large sample sizes. Because averaging the outcome probabilities given the predictors over all possible parameter values in the parameter space is typically computationally intractable, simulation methods and analytic approximations have been explored to obtain workable approximations.

7.2. Explanation

A variety of statistical modeling paradigms [6,7,27,33,254,276–278] exist to perform classification that include discriminative classifiers, which model the conditional probability of the outcome given the observation (e.g. categorical regression), and generative classifiers that learn the model of the joint probability for the predictors and outcome (e.g. naive Bayes classifier). Both paradigms are utilized when the objective is classification. Classification results for statistical models with binary outcomes are often presented in 2×2 contingency tables with associated statistics (e.g. accuracy, sensitivity, specificity, etc.) as shown in Table 7. Results for categorical regression models having more than two outcomes are represented similarly, in an extended format. Additional measures of performance include receiver operating characteristic (ROC) curves [261,266] and Youden Index (J) [262,279–281]. Examples of classification performance measures [262] are presented in Table 8.

7.3. Recommendations

The challenge for performing classification is to determine and apply the appropriate measure for the researcher’s statistical model. These measures should be identified as part of the researcher’s data analysis plan, with the type of classification for data analysis dependent on the particular statistical modeling being utilized. The following modeling guidelines are recommended.

Table 8. Examples of classification measures.

No.	Classification measures	Description
1	Accuracy (% correct)	Proportion of total correct positive and negative classifications [7,254].
2	Error Rate (% incorrect)	Proportion of total incorrect classifications [7].
3	Balanced Error Rate (BER)	Average of Sensitivity and Specificity.
4	Sensitivity	Proportion of positive classifications that are predicted positive [12,261]. Also referred to true positive rate or recall.
5	Specificity	Proportion of negative classifications that are predicted negative [12,261]. Also referred to as true negative rate.
6	Positive Predictive Value (PPV)	Proportion of predicted positives that are actual positives [261,282]. Also referred to as precision.
7	Negative Predictive Value (NPV)	Proportion of predicted negative that are actual negatives [261,282].
8	ROC	Receiver Operator Characteristic (ROC) curve graphs true positive rates against false positive rates [12,261,265,266].
9	AUROC (or AUC)	Area under the Receiver Operating Characteristic curve where Receiver Operator Characteristic (ROC) curves plots true positive rates against false positive rates [12,261,265,266].
10	Youden Index (J)	Computes performance for a binary classifier as Sensitivity + Specificity – 1 [279,280].
11	Cohen's Kappa coefficient (κ)	Measure of inter-rater agreement for categorical items that corrects for chance agreement [283].
12	Matthews correlation coefficient (MCC)	Balanced measure that accounts for true positives, true negatives, false positives, and false negatives [284].
13	Phi (ϕ) Coefficient	Symmetric statistic that measures the association between two binary variables [285].
14	F1 Score	Measure of accuracy that is the harmonic average of precision and recall [286,287].

Recommendations: Classification Performance

1. *Minimum Acceptable Statistical Modeling* – Classification analyses that include model validation methods (e.g. k -fold, LOOCV, bootstrapping, etc.) are performed on the researcher's final model to determine whether the model is likely to accurately classify responses. *Researcher* specifies the appropriate model classification measures (e.g. accuracy, Kappa [283], AUROC [12,261,265,266], Youden Index (J) [279,280]) and a validation method for checking the classifier reliability in the data analysis plan.
 2. *Advanced Statistical Modeling* – Classification analyses that also include model validation methods with estimation of decision thresholds [288] are performed to evaluate different aspects of final model fit and classification performance. *Researcher* specifies all appropriate model classification measures, decision thresholds requiring optimization, and validation methods in data analysis plan.
-

8. Summary

In this article, recommendations were provided for supporting the development and evaluation of statistical models with an emphasis on regression modeling. Such methods are widely used in clinical science, epidemiology, and health services research to analyze and interpret data collected in interventional and observational studies, and thus have considerable bearing on critical decision making. In particular, methods were reviewed to support strategies for: (1) assessing model fit, (2) representing data, (3) identifying a single model or collection of models from a pool of covariates, (4) model validation, (5) evaluating model prediction, and (6) evaluating model classification. In addition to dealing with the problem of developing models with good model fit, special attention was also provided for the purpose of addressing problems of model misspecification and multicollinearity.

These recommendations and methods are applicable to GLMs and extensions such as GAMs, as well as the exponential family of nonlinear models. Further, nonparametric methods were presented that are also supportive of model development. A major focus of this article was devoted to the critical nature of developing correctly specified models by utilizing more advanced statistical approaches. We presented commonly used methods with additional recommendations for more advanced modeling strategies to support practical development of improved statistical models. Such models lead to better diagnostic and prognostic inferences that inform researchers as well as practitioners, administrators, and policy makers, who use research findings to make decisions that impact patients, their quality of life, and health care costs.

Acknowledgements

The authors wish to gratefully acknowledge the support of the National Institute of General Medical Sciences (NIGMS), National Institute of Mental Health (NIMH), and National Institute on Drug Abuse (NIDA). This paper reflects the authors' views and not necessarily the opinions or views of the NIGMS, NIMH, or NIDA.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was made possible by grants from the National Institute of General Medical Sciences (NIGMS) (R43GM123831, PI: S.S. Henley; R43GM114899, PI: S.S. Henley; R43GM106465, PI: S.S. Henley), National Institute of Mental Health (NIMH) (R43MH105073, PI: S.S. Henley), and National Institute on Drug Abuse (NIDA) (R43DA047224, PI: S.S. Henley).

Notes on contributors

Steven S. Henley, M.S. is a Computational Statistician, Research Professor of Medicine at the Loma Linda University School of Medicine, and President of Martingale Research Corporation. Professor Henley has extensive experience in statistical modeling, algorithm development, machine learning, systems engineering, and R&D project management. His research focuses on the development and evaluation of new theory, algorithms, and statistical software for clinical, epidemiologic, and health services research. Prof. Henley is a principal investigator for National Institutes of Health (NIH) Small Business Innovation Research (SBIR) sponsored research that develops advanced statistical modeling technologies.

Richard M. Golden, Ph.D., M.S.E.E., B.S.E.E. is a Mathematical Psychometrician, Professor of Cognitive Science and Electrical Engineering, and Program Head of the Undergraduate Cognitive Science and Graduate Applied Cognition and Neuroscience Programs in the School of Behavioral and Brain Sciences at the University of Texas at Dallas (UTD). He was an action editor of the *Journal of Mathematical Psychology*, author of *Mathematical Methods for Neural Network Analysis and Design* (MIT Press, 1996), and has published numerous papers on assessing model quality and robust statistical inference in the possible presence of model misspecification. Dr. Golden has been a principal investigator on National Science Foundation (NSF) funded research and collaborated on National Institutes of Health (NIH) Small Business Innovation Research (SBIR) projects that develop novel statistical modeling methods.

T. Michael Kashner, Ph.D., J.D. is a Health Econometrician and Senior Scientist with the Veterans Health Administration's Office of Academic Affiliations (OAA) in the Department of Veterans

Affairs (VA), Washington, DC, and Research Professor of Medicine at the Loma Linda University School of Medicine in Loma Linda, CA. As a well-funded and published health services researcher since 1983, Dr. Kashner applies advanced analytic methods to large education, clinical, cost, and administrative databases to make robust causal inferences in health professions education and mental health services for clinicians, administrators, and policy decision-makers.

References

- [1] Agresti A. *Categorical data analysis*. 2nd ed. New York: Wiley-Interscience; 2002.
- [2] Rencher AC, Christensen WF. *Methods of multivariate analysis*. 3rd ed. Hoboken (NJ): Wiley; 2012.
- [3] Christensen R. *Log-linear models and logistic regression*. New York: Springer-Verlag; 1997.
- [4] Fox J. *Applied regression analysis and generalized linear models*. 3rd ed. Los Angeles (CA): SAGE; 2015.
- [5] Draper NR, Smith H. *Applied regression analysis*. 3rd ed. New York: Wiley; 1998.
- [6] Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer; 2001.
- [7] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer; 2009.
- [8] Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. 1st; Reprint ed. New York: Springer; 2010.
- [9] Royston P, Sauerbrei W. *Multivariate model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. New York: John Wiley & Sons; 2008.
- [10] Hastie T, Tibshirani R. *Generalized additive models*. 1st ed. New York: Chapman and Hall; 1990.
- [11] McCullagh P, Nelder JA. *Generalized linear models*. 2nd ed. London; New York: Chapman and Hall; 1989.
- [12] Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. 3rd ed. New York: Wiley-Interscience; 2013.
- [13] Hosmer DW, Lemeshow S, May S. *Applied survival analysis: regression modeling of time-to-event data*. 2nd ed. Hoboken (NJ): Wiley-Interscience; 2008.
- [14] Zar JH. *Biostatistical analysis*. 5th ed. New York: Pearson; 2009.
- [15] Gentle JE. *Elements of computational statistics*. New York: Springer-Verlag; 2002.
- [16] Vittinghoff E, Glidden DV, Shiboski SC, et al. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. 2nd ed. New York: Springer; 2012.
- [17] Daniel WW, Cross CL. *Biostatistics: a foundation for analysis in the health sciences*. 11th ed. Hoboken (NJ): John Wiley & Sons; 2019.
- [18] White H. *Estimation, inference, and specification analysis*. Cambridge; New York: Cambridge University Press; 1994.
- [19] White H. *Asymptotic theory for econometricians*. Revised ed. New York: Academic Press; 2001.
- [20] Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task force-7. *Value Health*. 2012;15(6):843–850.
- [21] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13:1.
- [22] Depaoli S, van de Schoot R. Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychol Methods*. 2017;22(2):240–261.
- [23] These MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochem Med*. 2015;25(1):5–11.

- [24] Caro JJ, Briggs AH, Siebert U, et al. Modeling good research practices—overview: a report of the ISPOR-SMDM modeling good research practices task force–1. *Value Health*. 2012;15(6):796–803.
- [25] Wei B. *Exponential family nonlinear models*. New York: Springer; 1998.
- [26] Bishop C. *Neural networks for pattern recognition*. New York: Oxford University Press; 1995.
- [27] Golden RM. *Mathematical methods for neural network analysis and design*. Cambridge (MA): MIT Press; 1996.
- [28] Duda RO, Hart PE, Stork DG. *Pattern classification*. 2nd ed. New York: Wiley-Interscience; 2000.
- [29] Bishop CM. *Pattern recognition and machine learning*. 1st (reprint) ed. New York: Springer; 2016.
- [30] Murphy KP. *Machine learning: a probabilistic perspective*. 1st ed. Cambridge (MA): The MIT Press; 2012.
- [31] Zheng A, Casari A. *Feature engineering for machine learning: principles and techniques for data scientists*. 1st ed. Sebastopol (CA): O’Reilly Media; 2018.
- [32] Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.
- [33] Kuhn A, Johnson K. *Applied predictive modeling*. New York: Springer; 2013.
- [34] Collett D. *Modelling binary data*. 2nd ed. Boca Raton (FL): Chapman & Hall/CRC; 2003.
- [35] Dobson AJ, Barnett A. *An introduction to generalized linear models*. 3rd ed. Boca Raton (FL): Chapman & Hall/CRC; 2008.
- [36] Seber GAF, Wild CJ. *Nonlinear regression*; Hoboken (NJ): Wiley-Interscience; 2003.
- [37] Hardin JW, Hilbe JM. *Generalized linear models and extensions*. 3rd ed. College Station (TX): Stata Press; 2012.
- [38] Hilbe JM. *Logistic regression models*. New York: Chapman and Hall; 2009.
- [39] Weisberg S. *Applied linear regression*. 4th ed. Hoboken (NJ): Wiley; 2013.
- [40] Zhou X-H, Zhou C, Lui D, et al. *Applied missing data analysis in the health sciences*. 1st ed. Hoboken (NJ): Wiley; 2014.
- [41] Little RJA, Rubin DB. *Statistical analysis with missing data*. 2nd ed. Hoboken (NJ): Wiley-Interscience; 2002.
- [42] Schafer JL. *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC; 1997.
- [43] Allison PD. *Missing data*. Thousand Oaks (CA): Sage; 2001.
- [44] Golden RM, Henley SS, White H, et al. Consequences of model misspecification for maximum likelihood estimation with missing data. *Econometrics manuscript prepared (invited article)*.
- [45] Zhou Z-H. Challenges and strategies in analysis of missing data. *Biostat Epidemiol*. 2019. Submitted.
- [46] Cook RD, Weisberg S. *Applied regression including computing and graphics*. 1st ed. New York: Wiley-Interscience; 1999.
- [47] Cook RD, Weisberg S. *Residuals and influence in regression*. 1st ed. New York: Chapman and Hall/CRC; 1983.
- [48] Fox J. *Regression diagnostics: an introduction (quantitative applications in the social sciences)*. 1st ed. Newbury Park (CA): SAGE; 1991.
- [49] Hamilton LC. *Regression with graphics: a second course in applied statistics*. Pacific Grove, CA: Brooks/Cole; 1992.
- [50] Belsley DA, Kuh E, Welsch RE. *Regression diagnostics: identifying influential data and sources of collinearity*. 1st ed. New York: Wiley-Interscience; 1980.
- [51] Wasserman L. *All of nonparametric statistics*. New York: Springer; 2007.
- [52] Corder GW, Foreman DI. *Nonparametric statistics: a step-by-step approach*. 2nd ed. Hoboken (NJ): Wiley; 2014.
- [53] Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*. 3rd ed. New York: John Wiley & Sons; 2014.
- [54] Kashner TM, Henley SS, Golden RM, Zhou Z-HA. Making causal inferences about treatment effect sizes from observational datasets. *Biostat Epidemiol*. 2019. Submitted.
- [55] Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82:669–688.

- [56] Pearl J. Causality: models, reasoning, and inference. Cambridge: University of Cambridge Press; 2000.
- [57] VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol.* 2007;166(9):1096–1104.
- [58] Hoeting JA, Madigan D, Raftery AE, et al. Bayesian model averaging: a tutorial. *Stat Sci.* 1999;14(4):382–401.
- [59] Garcia-Perez MA. Statistical conclusion validity: some common threats and simple remedies. *Front Psychol.* 2012;3(325). doi:10.3389/fpsyg.2012.00325
- [60] Cook TD, Campbell DT. Quasi-experimentation: design & analysis issues for field settings. 1st ed. Boston (MA): Houghton Mifflin; 1979.
- [61] Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer; 2002.
- [62] Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res.* 2004;33(2):261–304.
- [63] Claeskens G, Hjort NL. Model selection and model averaging. Cambridge; New York: Cambridge University Press; 2008.
- [64] Steckler A, McLeroy KR. The importance of external validity. *Am J Public Health.* 2008;98(1):9–10.
- [65] Persaud N, Mamdani MM. External validity: the neglected dimension in evidence ranking. *J Eval Clin Pract.* 2006;12(4):450–453.
- [66] Maronna RA, Martin RD, Yohai VJ, et al. Robust statistics: theory and methods (with R). 2nd ed. Hoboken (NJ): John Wiley & Sons; 2019.
- [67] Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat Sci.* 2014;29(4):579–595.
- [68] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* 1999;130(6):515–524.
- [69] Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med.* 2000;19(4):453–473.
- [70] Khorsan R, Crawford C. How to assess the external validity and model validity of therapeutic trials: a conceptual approach to systematic review methodology. *Evid Based Compl Altern Med.* 2014;2014:1–12.
- [71] Steyerberg EW, Harrell FE, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* 2001;54(8):774–781.
- [72] Brakenridge SC, Phelan HA, Henley SS, et al. Early blood product and crystalloid volume resuscitation: risk association with multiple organ dysfunction after severe blunt traumatic injury. *J Trauma.* 2011;71(2):299–305.
- [73] Brakenridge SC, Henley SS, Kashner TM, et al. Comparing clinical predictors of deep venous thrombosis vs. pulmonary embolus after severe blunt injury: a new paradigm for post-traumatic venous thromboembolism? *J Trauma.* 2013;74(5):1231–1237.
- [74] Westover AN, Kashner TM, Winhusen TM, et al. A systematic approach to subgroup analyses in a smoking cessation trial. *Am J Drug Alcohol Abuse.* 2015;41(6):498–507.
- [75] Henley SS, Kashner TM, Golden RM, et al. Response to letter regarding “a systematic approach to subgroup analyses in a smoking cessation trial”. *Am J Drug Alcohol Abuse.* 2016;42(1):112–113.
- [76] Gorelick DA, McPherson S. Improving the analysis and modeling of substance use. *Am J Drug Alcohol Abuse.* 2015;41(6):475–478.
- [77] Volkow ND. Director’s report to the national advisory council on drug abuse, September 2015. US National Institutes of Health: National Institute on Drug Abuse; 2015.
- [78] Winhusen TM, Somoza EC, Brigham GS, et al. Impact of attention-deficit/hyperactivity disorder (ADHD) treatment on smoking cessation intervention in ADHD smokers: a randomized, double-blind, placebo-controlled trial. *J Clin Psychiatry.* 2010;71(12):1680–1688.
- [79] Golden RM, Henley SS, White H, et al. Generalized information matrix tests for detecting model misspecification. *Econometrics.* 2016;4(4):46.

- [80] Golden RM, Henley SS, White H, et al. New directions in information matrix testing: eigen-spectrum tests. In: Chen X, Swanson NR, editors. *Causality, prediction, and specification analysis: recent advances and future directions: essays in honor of Halbert L. White, Jr. (Festschrift Hal White Conference)*. New York: Springer; 2013. p. 145–178.
- [81] White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982;50(1): 1–25.
- [82] Bates DM, Watts DG. *Nonlinear regression analysis and its applications*. New York: Wiley-Interscience; 2007.
- [83] Hemmert GAJ, Schons LM, Wieseke J, et al. Log-likelihood-based pseudo- R^2 in logistic regression: deriving sample-sensitive benchmarks. *Sociol Methods Res*. 2016;47(3):507–531.
- [84] Kullback S, Leibler R. On information and sufficiency. *Ann Math Stat*. 1951;22:79–86.
- [85] Akaike H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*; 1973; Budapest.
- [86] Linhart H, Zucchini W. *Model selection*. New York: Wiley; 1986.
- [87] Konishi S, Kitagawa G. *Information criteria and statistical modeling*. New York: Springer; 2008.
- [88] Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. *Biometrika*. 1989;76(2):297–307.
- [89] Hurvich CM, Tsai CL. A corrected Akaike information criterion for vector autoregressive model selection. *J Time Ser Anal*. 1993;14:271–279.
- [90] Schwartz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461–464.
- [91] Ando T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*. 2007;94(2):443–458.
- [92] Ando T. Predictive Bayesian model selection. *Am J Math Manage Sci*. 2011;31:13–38.
- [93] Bozdogan H. Akaike's information criterion and recent developments in information complexity. *J Math Psychol*. 2000;44(1):62–91.
- [94] Takeuchi K. Distribution of information statistics and a criterion of model fitting for adequacy of models. *Math Sci*. 1976;153:12–18.
- [95] Djuric PM. Asymptotic MAP criteria for model selection. *IEEE Trans Signal Process*. 1998;46:2726–2735.
- [96] Lv J, Liu JS. Model selection principles in misspecified models. *J R Stat Soc Series B Stat Methodol*. 2014;76(1):141–167.
- [97] Poskitt DS. Precision, complexity and Bayesian model determination. *J R Stat Soc Series B Methodol*. 1987;49(2):199–208.
- [98] Cavanaugh JE. A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat Probab Lett*. 1999;42:333–343.
- [99] Seghouane AK. A note on overfitting properties of KIC and KICc. *Signal Process*. 2006;86:3055–3060.
- [100] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817–838.
- [101] White H. Consequences and detection of misspecified nonlinear regression models. *J Am Stat Assoc*. 1981;76:419–433.
- [102] Begg MD, Lagakos S. On the consequences of model misspecification in logistic regression. *Environ Health Perspect*. 1990;87:69–75.
- [103] Lehmann EL. Model specification: the views of Fisher and Neyman, and later developments. *Stat Sci*. 1990;5:160–168.
- [104] Godfrey LG. *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches*. Cambridge, MA: Cambridge University Press; 1989.
- [105] Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, MA: Cambridge University Press; 2006.
- [106] Davison AC, Tsai CL. Regression model diagnostics. *Int Stat Rev*. 1992;60:337–353.
- [107] Cook RD. Detection of influential observation in linear regression. *Technometrics*. 1977;19(1):15–18.

- [108] Hodge VJ, Austin J. A survey of outlier detection methodologies. *Artif Intell Rev.* 2004;22(2):85–126.
- [109] Rousseeuw P, Leroy A. Robust regression and outlier detection. 3rd ed. New York: John Wiley & Sons; 1996.
- [110] White H. Specification testing in dynamic models. *Advances in Econometrics - Fifth World Congress*; 1987.
- [111] Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat.* 1938;9:60–62.
- [112] Davies N, Petrucci JD. Detecting non-linearity in time series. *J R Stat Soc Series D Stat.* 1986;35(2):271–280.
- [113] Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health.* 1991;81:1630–1635.
- [114] Hosmer DW, Hosmer T, Le Cessie S, et al. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med.* 1997;16:965–980.
- [115] Allison PD. Measures of fit for logistic regression. Paper 1485-2014. *SAS Global Forum*; 2014; Washington (DC).
- [116] Xie XJ. Goodness-of-fit tests for logistic regression models: evaluating logistic model fit when continuous covariates are present. Riga: VDM Verlag Dr. Mueller E.K.; 2008.
- [117] Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med.* 2002;21(18):2723–2738.
- [118] Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. *Stat Med.* 2002;21:3789–3801.
- [119] Cho JS, Phillips PCB. Pythagorean generalization of testing the equality of two symmetric positive definite matrices. *J Econom.* 2018;202(1):45–56.
- [120] Prokhorov A, Schepsmeier U, Zhu Y. Generalized information matrix tests for copulas. *Econom Rev.* 2019; in press.
- [121] Huang W, Prokhorov A. A goodness-of-fit test for copulas. *Econom Rev.* 2014;33(7):751–771.
- [122] Cho J, Phillips PCB. Testing equality of covariance matrices via Pythagorean means. *Cowles Foundation Discussion Paper No.* 1970; 2014.
- [123] Ibragimov R, Prokhorov A. Heavy tails and copulas: topics in dependence modelling in economics and finance. Hackensack (NJ): World Scientific Publishing; 2017.
- [124] Schepsmeier U. Efficient information based goodness-of-fit tests for vine copula models with fixed margins: a comprehensive review. *J Multivar Anal.* 2015;138:34–52.
- [125] Schepsmeier U. A goodness-of-fit test for regular vine copula models. *Econom Rev.* 2016;38(1):25–46.
- [126] Hall A. The information matrix test for the linear model. *Rev Econ Stud.* 1987;54:257–263.
- [127] Nelder JA, Wedderburn RWM. Generalized linear models. *J R Stat Soc Series A.* 1972;135:370–384.
- [128] Golden RM. Making correct statistical inferences using a wrong probability model. *J Math Psychol.* 1995;39(1):3–20.
- [129] Archer KJ, Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata J.* 2006;6:97–105.
- [130] Copas JB. Unweighted sum of squares test for proportions. *Appl Stat.* 1989;38:71–80.
- [131] Deng X, Wan S, Zhang B. An improved goodness-of-test for logistic regression models based on case-control data by random partition. *Commun Stat Simul Comput.* 2009;38:233–243.
- [132] Hosmer DW, Lemeshow S, Klar J. Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small. *Biom J.* 1988;30(7):1–14.
- [133] Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commun Stat.* 1980;9:1043–1069.
- [134] Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika.* 1997;84:609–618.
- [135] Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. *Biometrika.* 1980;67:250–251.

- [136] Zhang B. A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1999;86:531–539.
- [137] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.
- [138] Breusch TS, Pagan AR. Simple test for heteroscedasticity and random coefficient variation. *Econometrica*. 1979;47(5):1287–1294.
- [139] Durbin J, Watson GS. Testing for serial correlation in least squares regression. *Biometrika*. 1971;58(1):1–19.
- [140] White KJ. The Durbin-Watson test for autocorrection in nonlinear models. *Rev Econ Stat*. 1992;74(2):370–373.
- [141] Hendry DF, Richard J-F. On the formulation of empirical models in dynamic econometrics. *J Econom*. 1982;20:3–33.
- [142] Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat Med*. 2013;32(1):67–80.
- [143] Davidson J. *Econometric theory*. Malden (MA): Wiley-Blackwell; 2000.
- [144] Ljung GM, Box GEP. On a measure of a lack of fit in time series models. *Biometrika*. 1978;65(2):297–303.
- [145] Golden RM. Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika*. 2003;68(2):229–249.
- [146] Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989;57:307–333.
- [147] Golden RM. Statistical tests for comparing possibly misspecified and nonnested models. *J Math Psychol*. 2000;44(1):153–170.
- [148] Rivers D, Vuong Q. Model selection tests for nonlinear dynamic models. *Economet J*. 2002;5:1–39.
- [149] Feng ZD, McCulloch CE. Using bootstrap likelihood ratios in finite mixture models. *J R Stat Soc Series B Methodol*. 1996;58(3):609–617.
- [150] Tekle FB, Gudicha DW, Vermunt JK. Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Adv Data Anal Classif*. 2016;10(2):209–224.
- [151] Gray HL, Baek J, Woodward WA, et al. A bootstrap generalized likelihood ratio test in discriminant analysis. *Comput Stat Data Anal*. 1996;22(2):137–158.
- [152] McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *J R Stat Soc Series C Appl Stat*. 1987;36(3):318–324.
- [153] Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika*. 2001;88(3):767–778.
- [154] Gibbons JD, Chakraborti S. *Nonparametric statistical inference*. 4th ed. New York: Chapman and Hall/CRC; 2003.
- [155] Serfling RJ. *Approximation theorems of mathematical statistics*. 2nd ed. New York: Wiley-Interscience; 1980.
- [156] Bagdonavicius V, Kruopis J, Nikulin MS. *Non-parametric tests for complete data*. London & Hoboken: Wiley-ISTE; 2011.
- [157] Kanji GK. *100 statistical tests*. 3rd ed. Thousand Oaks, CA: SAGE; 2006.
- [158] Barnett V, Lewis T. *Outliers in statistical data*. 3rd ed. New York: Wiley; 1994.
- [159] Iglewicz B, Hoaglin DC. *How to detect and handle outliers*. Milwaukee (WI): American Society for Quality Control Press; 1993.
- [160] Anderson TW, Darling DA. Asymptotic theory of certain “goodness-of-fit” criteria based on stochastic processes. *Ann Math Stat*. 1952;23:193–212.
- [161] Justel A, Peña D, Zamar R. A multivariate Kolmogorov–Smirnov test of goodness of fit. *Stat Probab Lett*. 1997;35(3):251–259.
- [162] Kolmogorov AN. Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell’Istituto Italiano degli Attuari*. 1933;4:83–91.

- [163] Smirnov N. Table for estimating the goodness of fit of empirical distributions. *Ann Math Stat.* 1948;19:279–281.
- [164] Zhao D, Bu L, Alippi C, et al. A Kolmogorov-Smirnov test to detect changes in stationarity in big data. *IFAC PapersOnLine.* 2017;50(1):14260–14265.
- [165] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3–4):591–611.
- [166] Kendall M. Rank correlation methods. Oxford: Charles Griffin & Company Limited; 1948.
- [167] Kendall M. A new measure of rank correlation. *Biometrika.* 1938;30(1–2):81–93.
- [168] Well AD, Myers JL. Research design and statistical analysis. 3rd ed. New York: Routledge; 2010.
- [169] Cook RD. Regression graphics: ideas for studying regressions through graphics. 1st ed. New York: Wiley-Interscience; 1998.
- [170] Kashner TM, Hinson R, Holland G, et al. A data accounting system for clinical investigators. *J Am Med Inform Assoc.* 2007;14(4):394–396.
- [171] Osborne JW. Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data. 1st ed. Thousand Oaks (CA): SAGE Publications; 2013.
- [172] Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Series B.* 1964;26(2): 211–252.
- [173] Kashner TM, Henley SS, Golden RM, et al. Assessing the preventive effects of cognitive therapy following relief of depression: a methodological innovation. *J Affect Disord.* 2007;104(1–3):251–261.
- [174] Kashner TM, Henley SS, Golden RM, et al. Studying the effects of ACGME duty hours limits on resident satisfaction: results from VA learners’ perceptions survey. *Acad Med.* 2010;85(7):1130–1139.
- [175] Jaccard J, Turris R. Interaction effects in multiple regression. (NY): SAGE; 2003.
- [176] Box GEP. Do interactions matter? *Qual Eng.* 1990;2:365–369.
- [177] Balli HO, Sørensen BE. Interaction effects in econometrics. *Empir Econ.* 2013;45(1):583–603.
- [178] Aiken LS, West SG. Multiple regression: testing and interpreting interactions. Thousand Oaks (CA): Sage; 1991.
- [179] Cox DR. Interaction. *Int Stat Rev / Revue Internationale de Statistique.* 1984;52(1):1–24.
- [180] Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med.* 1983;2(2):243–251.
- [181] Hayes AF, Matthes J. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behav Res Methods.* 2009;41(3):924–936.
- [182] Stone-Romero EF, Anderson LE. Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderator effects. *J Appl Psychol.* 1994;79:354–359.
- [183] George EI, Clyde M. Model uncertainty. *Stat Sci.* 2004;19(1):81–94.
- [184] Draper D. Assessment and propagation of model uncertainty. *J R Stat Soc Series B Methodol.* 1995;57(1):45–97.
- [185] Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health.* 1989;79(3):340–349.
- [186] George EI. The variable selection problem. *J Am Stat Assoc.* 2000;95(452):1304–1308.
- [187] Kosorok MR, Ma S. Marginal asymptotics for the “large p, small n” paradigm: with applications to microarray data. *Ann Stat.* 2007;35(4):1456–1486.
- [188] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodol.* 2008;70(5):849–911.
- [189] Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. *J R Stat Soc Series B Stat Methodol.* 2013;75(3):531–552.
- [190] Wasserman L, Roeder K. High-dimensional variable selection. *Ann Stat.* 2009;37(5A):2178–2201.
- [191] Buhlmann P, van de Geer S. Statistics for high-dimensional data: methods, theory, and applications. New York: Springer; 2011.

- [192] West M. Bayesian factor regression models in the “large p , small n ” paradigm. In: Bernardo JM, Bayarri MJ, Berger JO, et al., editors. *Bayesian statistics*. London: Oxford University Press; 2003. p. 723–732.
- [193] Johnstone IM, Titterton DM. Statistical challenges of high-dimensional data. *Philos Trans Series A Math Phys Eng Sci*. 2009;367(1906):4237–4253.
- [194] Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Series B*. 1996;58:267–288.
- [195] Hastie T, Tibshirani R, Tibshirani RJ. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv:1707.08692v2 [stat.ME]; 2017.
- [196] Efron MA. Multiple regression analysis. In: Ralston A, Wilf HS, editors. *Mathematical methods for digital computers*. New York: Wiley; 1960. p. 191–203.
- [197] Miller AJ. *Subset selection in regression*. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2002.
- [198] Furnival GM. All possible regressions with less computation. *Technometrics*. 1971;13(2): 403–408.
- [199] Furnival GM, Wilson RW. Regression by leaps and bounds. *Technometrics*. 1974;16:499–511.
- [200] Gatu C, Kontoghiorghe EJ. Branch-and-bound algorithms for computing the best-subset regression models. *J Comput Graph Stat*. 2006;15(1):139–156.
- [201] Hand DJ. Branch and bound in statistical data. *J R Stat Soc Series D Stat*. 1981;30(1):1–13.
- [202] Hans C, Dobra A, West M. Shotgun stochastic search for “large p ” regression. *J Am Stat Assoc*. 2007;102(478):507–516.
- [203] Cai A, Tsay RS, Chen R. Variable selection in linear regression with many predictors. *J Comput Graph Stat*. 2009;18(3):573–591.
- [204] George EI, McCulloch RE. Variable selection via Gibbs sampling. *J Am Stat Assoc*. 1993;88:881–889.
- [205] Han C, Carlin BP. Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *J Am Stat Assoc*. 2001;96(455):1122–1132.
- [206] Hocking R. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32:1–49.
- [207] Garside MJ. The best sub-set in multiple-regression analysis. *R Stat Soc Series C Appl Stat*. 1965;14(2–3):196–200.
- [208] Schatzoff M, Tsao R, Fienberg S. Efficient calculation of all possible regressions. *Technometrics*. 1968;10(4):769.
- [209] Bertsimas D, King A, Mazumder R. Best subset selection via a modern optimization lens. *Ann Stat*. 2016;44(2):813–852.
- [210] Bertsimas D, King A. Logistic regression: from art to science. *Stat Sci*. 2017;32(3):367–384.
- [211] Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. *Biometrics*. 1989;45(4):1265–1270.
- [212] Edwards D, Havranek T. A fast model selection procedure for large families of models. *J Am Stat Assoc*. 1987;82(397):205–213.
- [213] Edwards D, Havranek T. A fast procedure for model search in multidimensional contingency tables. *Biometrika*. 1985;72(2):339–351.
- [214] Lawless JF, Singhal K. Efficient screening of nonnormal regression models. *Biometrics*. 1978;34(2):318–327.
- [215] LaMotte LR, Hocking R. Computational efficiency in the selection of regression variables. *Technometrics*. 1970;12(1):83–93.
- [216] Brusco MJ, Stahl S. *Branch-and-Bound applications in Combinatorial data analysis*. New York: Springer-Verlag; 2005.
- [217] Gatu C, Kontoghiorghe EJ. Efficient strategies for deriving the subset VAR models. *Comput Manage Sci*. 2005;2(4):253–278.
- [218] Gatu C, Kontoghiorghe EJ, Gilli M, et al. An efficient branch-and-bound strategy for subset vector autoregressive model selection. *J Econ Dynam Control*. 2008;32(6):1949–1963.
- [219] Winkler G. *Image analysis, random fields, and dynamic Monte Carlo methods*. New York: Springer-Verlag; 1991.

- [220] Gilks WR, Richardson S, Spiegelhalter DJ, editors. *Markov Chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC; 1996.
- [221] Geyer C. Practical Markov chain Monte Carlo. *Stat Sci*. 1992;7(4):473–483.
- [222] Gamerman D, Lopes HF. *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. 2nd ed. New York: Chapman & Hall/CRC; 2006.
- [223] Neal RM. Probabilistic inference using Markov Chain Monte Carlo methods, technical report CRG-TR-93-1. Toronto: Department of Computer Science, University of Toronto; 1993.
- [224] Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc*. 1990;85(410):398–409.
- [225] George EI, McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin*. 1997;7:339–373.
- [226] Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam’s window. *J Am Stat Assoc*. 1994;89(428):1535–1546.
- [227] Raftery AE, Zheng Y. Long-run performance of Bayesian model averaging. Seattle: University of Washington; 2003.
- [228] Wang D, Lertsithichai P, Nanchahal K, et al. Risk factors of coronary heart disease: A Bayesian model averaging approach. *J Appl Stat*. 2003;30(7):813–826.
- [229] Wang D, Zhang W, Bakhai A. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Stat Med*. 2004;23(22):3451–3467.
- [230] Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. *Stat Med*. 2001;20(21):3215–3230.
- [231] Genell A, Nemes S, Steineck G, et al. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. *BMC Med Res Methodol*. 2010;10:108.
- [232] Wilson MA, Iversen ES, Clyde MA, et al. Bayesian model search and multilevel inference for SNP association studies. *Ann Appl Stat*. 2010;4(3):1342–1364.
- [233] Volinsky CT, Madigan D, Raftery AE, et al. Bayesian model averaging in proportional hazard models. assessing the risk of a stroke. *Appl Stat J R Stat Soc Series C*. 1997;46(4):433–448.
- [234] Regal RR, Hook EB. The effects of model selection on confidence-intervals for the size of a closed population. *Stat Med*. 1991;10(5):717–721.
- [235] Vock DM, Atchison EA, Legler JM, et al. Accounting for model uncertainty in estimating global burden of disease. *Bull World Health Organ*. 2011;89(2):112–120.
- [236] Montgomery JM, Nyhan B. Bayesian model averaging: theoretical developments and practical applications. *Polit Anal*. 2010;18(2):245–270.
- [237] Ando T, Tsay R. Predictive likelihood for Bayesian model selection and averaging. *Int J Forecast*. 2010;26(4):744–763.
- [238] Prost L, Makowski D, Jeuffroy M. Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. *Ecol Modell*. 2008;219(1–2):66–76.
- [239] Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. *Biometrics*. 2005;61(2):498–506.
- [240] Jackson CH, Thompson SG, Sharples LD. Accounting for uncertainty in health economic decision models by using model averaging. *J R Stat Soc Series A Stat Soc*. 2009;172:383–404.
- [241] Morales KH, Ibrahim JG, Chen C-J, et al. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *J Am Stat Assoc*. 2006;101(473):9–17.
- [242] Clyde M. Bayesian model averaging and model search strategies. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. *Bayesian statistics 6*. London: Oxford University Press; 1999. p. 157–185.
- [243] Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc*. 1997;92:179–191.
- [244] Burnham KP, Anderson DR. *Model selection and inference: a practical information-theoretic approach*. New York: Springer; 1998.
- [245] Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics*. 1997;53(2):603–618.

- [246] Wang H, Zhang X, Zou G. Frequentist model averaging estimation: a review. *J Syst Sci Complex*. 2009;22(4):732–748.
- [247] Ullah A, Wang H. Parametric and nonparametric frequentist model selection and model averaging. *Econometrics*. 2013;1:157–179.
- [248] Barr J, Cavanaugh J. Forensics: assessing model goodness: a machine learning view. *Encyclopedia Semant Comput Robot Intell*. 2018;2(2):1850015.
- [249] Russell SJ, Norvig P. *Artificial intelligence: a modern approach*. New Jersey: Prentice Hall; 2003.
- [250] Picard R, Cook D. Cross-validation of regression models. *J Am Stat Assoc*. 1984;79(387):575–583.
- [251] Stone M. Cross-validated choice and assessment of statistical predictions. *J R Stat Soc Series B*. 1974;36:111–133.
- [252] Efron B, Tibshirani R. *An introduction to the bootstrap*. New York: Chapman & Hall/CRC; 1993.
- [253] Anderson TW. *An introduction to multivariate statistical analysis*. New York: John Wiley & Sons; 2003.
- [254] Gordon AD. *Classification*. 2nd ed. New York: Chapman & Hall/CRC; 1999.
- [255] Huberty CJ. *Applied discriminant analysis*. New York: John Wiley & Sons; 1994.
- [256] Krzanowski WJ. *Principles of multivariate analysis: a user's perspective*. New York: Oxford University Press; 2000.
- [257] Huber PJ, Ronchetti EM. *Robust statistics*. 2nd ed. New York: John Wiley & Sons; 2009.
- [258] de Myttenaere A, Golden B, Le Grand B, et al. Mean absolute percentage error for regression models. *Neurocomputing*. 2016;192:38–48.
- [259] Tarpey T. A note on the prediction sum of squares statistic for restricted least squares. *Am Stat*. 2000;54(2):116–118.
- [260] Allen DM. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*. 1974;16:125–127.
- [261] Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press; 2004.
- [262] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inform Process Manage*. 2009;45(4):427–437.
- [263] UCI machine learning repository. Irvine (CA): University of California, School of Information and Computer Science; 2017. Available from: <http://archive.ics.uci.edu/ml>.
- [264] Wickens TD. *Elementary signal detection theory*. New York: Oxford University Press; 2002.
- [265] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27:861–874.
- [266] Metz CE. Basic principles of ROC analyses. *Semin Nucl Med*. 1978;8:283–298.
- [267] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39(4):561–577.
- [268] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
- [269] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
- [270] King G, Zeng L. Logistic regression in rare events data. *Polit Anal*. 2001;9:137–163.
- [271] Berger JO. *Statistical decision theory and Bayesian analysis*. 2nd ed. New York: Springer-Verlag; 1985.
- [272] Aitchison J. Goodness of prediction fit. *Biometrika*. 1975;62:547–554.
- [273] Harris IR. Predictive fit for natural exponential families. *Biometrika*. 1989;76(4):675–684.
- [274] Vidoni P. Improved predictive model selection. *J Stat Plan Inference*. 2008;138:3713–3721.
- [275] Fushiki T, Komaki F, Aihara K. Nonparametric bootstrap prediction. *Bernoulli*. 2005;11(2):293–307.
- [276] Ng A, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *NIPS*; 2001.

- [277] Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization and beyond. Cambridge (MA): MIT Press; 2001.
- [278] Smola AJ, Bartlett PJ, Schölkopf B, et al. Advances in large-margin classifiers. Cambridge, MA: MIT Press; 2000.
- [279] Perkins NJ, Schisterman EF. The Youden index and the optimal cut-point corrected for measurement error. *Biom J*. 2005;47(4):428–441.
- [280] Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
- [281] Baldi P, Brunak S, Chauvin Y, et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412–424.
- [282] Fletcher RH, Fletcher SW. Clinical epidemiology: the essentials. 4th ed. Baltimore (MD): Lippincott Williams & Wilkins; 2005.
- [283] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
- [284] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405(2):442–451.
- [285] Cramér H. Mathematical methods of statistics. Princeton: Princeton University Press; 1946.
- [286] Sasaki Y. The truth of the F-measure. Manchester: School of Computer Science, University of Manchester; 2007.
- [287] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011;2(1):37–63.
- [288] Magder LS, Fix AD. Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *J Clin Epidemiol*. 2003;56(10):956–962.