Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic

Citation for published version:
McKeigue, P 2018, 'Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic', *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280218776989

Digital Object Identifier (DOI):
10.1177/0962280218776989

Published In:
Statistical Methods in Medical Research

Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic

Citation for published version:
McKeigue, P 2018, 'Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic' Statistical Methods in Medical Research. DOI: 10.1177/0962280218776989

Digital Object Identifier (DOI):
10.1177/0962280218776989

Published In:
Statistical Methods in Medical Research

Journal Title
XX(X):2–19

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Quantifying performance of a diagnostic test as the expected information for discrimination: relation to the C-statistic

Paul McKeigue

Abstract

Although the C-statistic is widely used for evaluating the performance of diagnostic tests, its limitations for evaluating the predictive performance of biomarker panels have been widely discussed. The increment in C obtained by adding a new biomarker to a predictive model has no direct interpretation, and the relevance of the C-statistic to risk stratification is not obvious. This paper proposes that the C-statistic should be replaced by the expected information for discriminating between cases and noncases (expected weight of evidence, denoted as Λ), and that the strength of evidence favouring one model over another should be evaluated by cross-validation as the difference in test log-likelihoods. Contributions of independent variables to predictive performance are additive on the scale of Λ. Where the effective number of independent predictors is large, the value of Λ is sufficient to characterize fully how the predictor will stratify risk in a population with given prior probability of disease, and the C-statistic can be interpreted as a mapping of Λ to the interval from 0.5 to 1. Even where this asymptotic relationship does not hold, there is a one-to-one mapping between the distributions in cases and noncases of the weight of evidence favouring case over noncase status, and the quantiles of these distributions can be used to calculate how the predictor will stratify risk. This proposed approach to reporting predictive performance is demonstrated by analysis of a dataset on the contribution of microbiome profile to diagnosis of colorectal cancer.

Keywords

diagnostic test, biomarkers, risk stratification, precision medicine, weight of evidence, cross-validation, C-statistic, Kullback-Leibler divergence, relative entropy, Bayesian

Introduction

The advent of platforms that can measure panels of hundreds or thousands of biomarkers presents new opportunities for developing diagnostic tests not only to detect disease, but to stratify people by risk and to predict response to therapy. It is widely expected that this will lead to a new era of "precision medicine" (1). The growth of research in this field has highlighted the limitations of current methods for evaluating the predictive performance of biomarker panels. There is no consensus on how to evaluate the incremental contribution of a biomarker panel to predictions based on clinical variables, and it is not clear how to use summary measures of predictive performance to evaluate the usefulness of a biomarker panel as a risk stratifier.

This paper is organized as follows. First, the limitations of current methods for quantifying performance of a diagnostic test are briefly reviewed. Next, the rationale for an alternative approach based on information theory and Bayesian inference is presented, and methods for calculating it are described. The proposed approach is demonstrated by applying it to a study that used a high-dimensional biomarker panel to distinguish cases and controls. The discussion section examines the relevance of other approaches to quantifying the information conveyed by an experiment or test, and recent guidelines for reporting predictive performance of diagnostic tests.

Limitations of current methods for quantifying performance of a classifier

The area under the receiver operating characteristic (ROC) curve or C-statistic is the most widely-used measure for evaluating the performance of a score in predicting a binary outcome. For simplicity, I denote the outcome as "disease", and the outcome categories as "case" and "control" though the argument applies more generally. Among the advantages of the C-statistic are that it does not require calibration and that it does not depend on the prevalence of disease, so that in principle an estimate obtained in a case-control study can be generalized to a clinical setting. With some additional assumptions, use of the C-statistic to

Usher Institute of Population Health Sciences and Informatics, University of Edinburgh

Corresponding author:
Paul McKeigue, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh EH8 9AG, UK
Email: paul.mckeigue@ed.ac.uk

evaluate the ranking of cases and controls is a proper scoring rule (2). This means that the assessed predictive performance is maximized by reporting the probabilities (or ranks) assigned by the forecaster. However the C-statistic also has serious limitations that have been widely discussed.

- It is not obvious what the C-statistic, defined as the probability of correctly classifying a case-control pair, tells us about the usefulness of a score for risk stratification in the population.
- The increment in the C-statistic obtained by adding new variables has no obvious interpretation. When a new predictor such as a biomarker is added to a baseline predictive model, the increment in C-statistic will depend upon what covariates have been included in the baseline model and on the extent to which these covariates have been matched between cases and controls (3; 4), even if the new predictor is uncorrelated with these covariates (5). The most efficient design in which to discover new biomarkers is a nested case-control study in which stored samples from cases are compared with controls matched for clinical covariates. When the predictive performance of a biomarker discovered in such a study is evaluated in a cohort study without matching for covariates, the increment in C-statistic obtained by adding the biomarker to this baseline model will be lower for reasons explained below. It is possible to work around this by standardizing the calculation of the ROC curve for covariates (6; 7), but this further complicates analysis and interpretation.
- The small increments in C-statistic that can be achieved by adding new variables to an baseline model that has a C-statistic of 0.9 or above have led to a mistaken belief that no useful increment in predictive performance can be obtained. "Researchers have observed that ΔAUC depends on the performance of the underlying clinical model. For example, good clinical models are harder to improve on, even with markers that have shown strong association" (8). Others have suggested that the problem lies in the interpretability of the C-statistic: "for models containing standard risk factors and possessing reasonably good discrimination, very large 'independent' associations of the new marker with the outcome are required to result in a meaningfully larger AUC" (9)

To supplement reporting of the C-statistic and the ROC curve, additional descriptors have been suggested. The cumulative distribution function $\mathcal{F}$ of the score values in controls can be estimated, and the distribution of the values returned by applying $\mathcal{F}$ to the score values in cases can be plotted as density of "percentile values" (10). The average of these values is equivalent to the C-statistic. To assess how the score will perform in a target population, the quantiles of predictive probability in that population can be plotted as a "predictiveness curve" (11); this however does not quantify predictive performance independently of the population in which the classifier is used.

To evaluate the incremental contribution of a new biomarker to prediction, alternative indices have been proposed, based on the proportion of individuals who are reclassified when the biomarker is added to the model: these include "integrated discrimination improvement" and "net reclassification index" (9). However these indices are not proper scoring rules (12); this means that adding a biomarker to the predictive model can apparently "improve" such an index even when that biomarker contains no predictive information (13; 14). The authors of a widely-quoted set of guidelines on reporting of multivariate models for diagnosis noted that "Identifying suitable measures for quantifying the incremental value of adding a predictor to an existing prediction model remains an active research area" (15).

Relation of C-statistic to expected information for discrimination

In a Bayesian framework, the weight of evidence favouring one hypothesis over another is the logarithm of the ratio of the likelihoods of the hypotheses given the data (16). This ratio of likelihoods of hypotheses is sometimes called the Bayes factor to distinguish it from the likelihood ratio tests used in classical statistics, which compare likelihoods at different values of a model parameter. The weight of evidence is not a scoring rule for comparison of classifiers: rather it is the difference between the logarithmic scores for the two hypotheses being compared (17). The C-statistic, defined as the probability of correctly classifying a case-control pair, is the probability that the weight of evidence in favour of the correct assignment of case-control status to this pair is greater than zero. We can calculate C, and also characterize the usefulness of the predictor for risk stratification, if we know the sampling distributions of the weight of evidence favouring case over control status in cases and controls.

Good and Toulmin (1968) (18) showed that for two alternative hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ the characteristic functions $\varphi_1(t)$, $\varphi_0(t)$ of the distributions of the weight of evidence $W_{1/0}$ favouring $\mathcal{H}_1$ over $\mathcal{H}_0$ when $\mathcal{H}_1$ is true and when $\mathcal{H}_0$ is true are related by the identity $\varphi_1(t+i) = \varphi_0(t)$, where i is the imaginary unit. This identity can be stated in an alternative form as $\exp\left(-W_{1/0}\right)p_1\left(W_{1/0}\right) = p_0\left(W_{1/0}\right)$, where $p_1\left(W_{1/0}\right)$ and $p_0\left(W_{1/0}\right)$ are the densities of $W_{1/0}$ when $\mathcal{H}_1$ is true and when $\mathcal{H}_0$ is true respectively. This result can be obtained simply by noting that at any value of W the ratio $p_1\left(W_{1/0}\right)/p_0\left(W_{1/0}\right)$ is the Bayes factor $\exp\left(W_{1/0}\right)$ favouring $\mathcal{H}_1$ over $\mathcal{H}_0$. This identity generalizes two results attributed to Turing (16):-

1. If the sampling distribution of the weight of evidence favouring a hypothesis $\mathcal{H}_1$ over a hypothesis $\mathcal{H}_0$ is Gaussian with mean Λ when $\mathcal{H}_1$ is true, its sampling distribution when $\mathcal{H}_0$ is true is Gaussian with mean $-\Lambda$, and both distributions have variance 2Λ (when natural logarithms are used).

2. The expected Bayes factor in favour of a wrong hypothesis is 1 (because $\exp\left(-W_{1/0}\right)p_1\left(W_{1/0}\right)$ integrates to 1). The practical implications of this result are examined in the Discussion section.

The sampling distribution of the weight of evidence is asymptotically Gaussian if there are many explanatory variables and their independent contributions are small (18). If this asymptotic distribution holds, the relation between the C-statistic and the expected weight of evidence Λ favouring true over false status is given by $C = 1 - \Phi\left(-\sqrt{\Lambda}\right)$ or $\Lambda = \left[\Phi^{-1}\left(C\right)\right]^2$ where $\Phi\left(\cdot\right)$ is the standard Gaussian cumulative distribution function (19). In this situation the C-statistic can be interpreted as a mapping of Λ (which can take values from 0 to infinity), to the interval from 0.5 to 1 as shown in Figure 1. A special case of this relation has been noted previously (20): with a single explanatory variable for which the class-conditional distributions in cases and controls are Gaussian with the same variance, $\Lambda = \frac{1}{2}\beta^2$ and $C = 1 - \Phi\left(-|\beta|/\sqrt{2}\right)$, where β is the standardized logistic regression coefficient of the outcome on the explanatory variable. More generally if the class-conditional distributions of the explanatory variables in cases and controls are Gaussian with the same covariance matrix, the sampling distribution of the weight of evidence favouring true over false status is Gaussian and the relation between C and Λ holds exactly (19).

The asymptotic relation between C and the expected weight of evidence Λ suggests that we might use Λ to report predictive performance. The statistic Λ has various alternative names: the expected information for discriminating between cases and controls; the Kullback-Leibler (KL) divergence from the class-conditional distribution $\mathcal{Q}$ of the predictors under incorrect case-control assignment to their distribution $\mathcal{P}$ under correct assignment; or the relative entropy of $\mathcal{P}$ with respect to $\mathcal{Q}$. As Λ is a KL divergence, it can take only non-negative values. The expected information for discrimination has a more intuitive interpretation than the C-statistic, because the mathematical definition of information as reduction in entropy corresponds closely to intuitive ideas of information (21). Improbable or surprising observations convey more information than unexceptional observations.

To facilitate intuitive interpretation of Λ, we can use logarithms to base 2, so that the expected information is expressed in bits. Figure 1 shows that a C-statistic of 0.7, sometimes cited as the threshold for "modest" predictive performance(22), is asymptotically equivalent to only 0.4 bits on the scale of Λ. More appropriate cutoffs for moderate and good prediction would be one bit and three bits, for which the asymptotically equivalent C values are respectively 0.8 and 0.925. Using Figure 1 we can explain how increments in the C-statistic may be misleading when used to evaluate the incremental contribution of a biomarker panel to predictive performance. For instance, in a case-control study where cases and controls have been matched for covariates so that the baseline model has a

C-statistic of 0.5, adding a biomarker that contributes one bit of information for discrimination would increase the asymptotically equivalent C-statistic from 0.5 to 0.8. When the same biomarker is evaluated in an unmatched cohort study in which the covariates contribute two bits of information, the baseline model will have a C-statistic of 0.88 and adding the biomarker will increase this only to 0.925. Whether or not the asymptotic relation between C and the expected weight of evidence Λ holds, contributions of independent predictors are additive on the scale of Λ; this supports using Λ instead of C to quantify predictive performance and the incremental contribution of additional biomarkers. In human genetics, the strength of the genetic effect on a disease is often quantified as the sibling recurrence risk ratio λ_S, defined as the ratio of disease risk in a sibling of an affected individual to average risk in the population. Under a polygenic model in which effects are additive on a logistic scale, $\Lambda = \log \lambda_S$ (23).

Evaluating the distributions of weight of evidence

To evaluate the performance of a predictive model, and the strength of evidence favouring one model over another over another, we require a test dataset with the observed case-control status y_i (coded as control = 0, case = 1) of the ith individual, the predicted probability p_i of disease in this individual generated by the model, and the prior probability of disease P given by the observed frequency of disease in the training dataset. This test dataset can be formed either by a single test/training split or by concatenating the N disjoint test folds used for N-fold cross-validation. Although the asymptotic properties discussed below are for leave-one-out cross-validation, it is not usually necessary in large datasets to proceed to the limit of leave-one-out; it is sufficient to start with $N = 10$ for N-fold cross-validation and to double N until the results do not change appreciably. For survival modelling where failure times are directly observed, the dataset can be rearranged with one observation per person-time interval, and the average taken over person-time intervals.

The weight of evidence w_i favouring correct over incorrect case-control assignment in the ith individual is calculated using Bayes theorem, by subtracting the log prior odds from the log posterior odds

$$w_i = (2y_i - 1)\left(\log \tfrac{p_i}{1-p_i} - \log \tfrac{P}{1-P}\right)$$

Λ is estimated as the average of w_i over all cases and controls in the test dataset.

The distributions of weight of evidence in cases and controls can then be examined. If these distributions have the asymptotic form derived by Turing, the expectation Λ contains all the information we need to compute quantiles of weight of evidence favouring case over control status in cases and controls. Otherwise to compute these quantiles we have to estimate these distributions from the data. For these estimated distributions to be consistent, they should be constrained so

that at each value of W the ratio of density in cases to density in controls is $\exp(W)$. The densities in cases and controls can be obtained by multiplying the geometric mean of these densities (as a function of W) by $\exp\left(\frac{1}{2}W\right)$ and $\exp\left(-\frac{1}{2}W\right)$ respectively. The problem of estimating a consistent pair of densities can thus be reduced to the problem of estimating this geometric mean function. A workable procedure for this is described below, where $f_0(W)$ and $f_1(W)$ denote estimated densities of the weight of evidence W favouring case over control status in cases and controls respectively.

1. Fit smoothed kernel densities $f_1(W)$, $f_0(W)$ to the values of W in the case and control samples respectively over a grid of values of W.
2. Estimate the geometric mean of the densities in cases and controls as a function of W as a weighted average of $f_1(W)\exp\left(-\frac{1}{2}W\right)$ and $f_0(W)\exp\left(\frac{1}{2}W\right)$. Weights for cases and controls are proportional to the expected numbers of cases and controls at each value of W: number of cases $\times \exp\left(\frac{1}{2}W\right)$, number of controls $\times \exp\left(-\frac{1}{2}W\right)$ respectively. This reduces to evaluating the arithmetic mean of the case and control densities as a function of W.
3. Calculate the adjusted densities $g_1(W)$, $g_0(W)$ in cases and controls by multiplying this estimated geometric mean function by $\exp\left(\frac{1}{2}W\right)$ and $\exp\left(\frac{1}{2}W\right)$ respectively.
   For the ratio $g_1(W)/g_0(W)$ to be exactly $\exp(W)$, these adjusted densities must have the same normalizing constant. This requires a slight reweighting of the unadjusted densities $f_1(W)$ and $f_0(W)$. The weighting function is $\exp\left(\pm\theta\left(w_i-\bar{w}\right)^2\right)$ where $\bar{w}$ is the sample mean of the weight of evidence and the sign before θ is positive in cases and negative in controls. The optimal value of θ is determined by using an optimization algorithm such as the *optim* function in the *R* package to minimize an objective function defined as the absolute value of the logarithm of the ratio of the two normalizing constants. The optimal value of θ is usually very close to zero - in other words, only very slight reweighting is required to ensure that the adjusted densities have the same normalizing constant.

Relation of the distributions of weight of evidence to the receiver operating characteristic curve

Johnson (24) noted a simple relationship between the distributions of weight of evidence W favouring case over control status in cases and controls and the ROC curve generated from these distributions. If the quantiles of W in controls and cases are q_0 and q_1 respectively, the sensitivity is $(1-q_1)$ and the specificity is q_0 and the ROC is the curve obtained by plotting $(1-q_1)$ as a function of $(1-q_0)$. The gradient of this function is

$$\frac{dq_1}{dq_0} = \frac{dq_1/dW}{dq_0/dW} = \frac{g_1(W)}{g_0(W)} = \exp(W)$$

As $q_0(W)$ increases with W, it follows that the gradient of this model-based ROC curve is a monotonic decreasing function of $(1 - q_0)$, unlike the crude ROC curve calculated from ranking the scores of cases and controls. This model-based ROC curve generated from the adjusted distributions of W in cases and controls contains the same information as a plot of the distributions, but is more difficult to use to quantify how the score will behave as a risk stratifier because the likelihood ratio cannot be read off a logarithmic scale on the axis but instead is represented as the gradient of the curve. A plot of the adjusted cumulative distributions of W in cases and controls is the most useful graphical representation of how the classifier can be used as a risk stratifier.

Evaluating the strength of evidence that adding one or more biomarkers improves prediction

To evaluate the strength of evidence that adding a biomarker or a panel of biomarkers improves prediction, we can evaluate the difference in log-likelihoods of the corresponding models given the test data. The log-likelihood of the model given test data on the ith individual is

$$\log \mathcal{L} = \sum_i \left[y_i \log p_i + (1 - y_i) \log (1 - p_i) \right]$$

Model comparison based on the test log-likelihood is equivalent to using the logarithmic scoring rule, which is strictly proper. In a Bayesian framework, the difference in log-likelihoods of models can be interpreted directly as the weight of evidence favouring one model over another, without having to evaluate its sampling distribution. It is possible to construct a test based on the distribution of the C-statistic over hypothetical repeated sampling of test datasets (25), but this is not the same as a classical p-value based on the distribution of the test statistic over hypothetical repeated sampling of training datasets (26). It is of interest to compare the relationship of these classical tests to inference based on test log-likelihoods. For leave-one-out cross-validation, the difference in test log-likelihoods of models is asymptotically equivalent to the difference in the values of the Akaike Information Criterion (27) (evaluated in natural log units rather than deviance units) on the training data, and $2(\Delta \log \mathcal{L} + k)$ has asymptotically a chi-square distribution with k degrees of freedom. where $\Delta \log \mathcal{L}$ is the difference in test log-likelihoods (in natural log units) of models with and without the extra biomarkers, and k is the effective number of extra parameters. Thus for a single extra variable, a test log-likelihood ratio of 20, which might be considered moderately strong evidence that a biomarker improves prediction, is asymptotically equivalent to a p-value of 0.0047 on the training dataset.

Example: incremental contribution of microbiome profile to detection of colorectal cancer

To demonstrate this approach to reporting the incremental contribution of a biomarker panel to prediction, these methods were applied to analysis of a publicly available dataset from a study of detection of colorectal cancer in symptomatic individuals, using fecal microbiome profile in addition to the standard fecal immunochemical test (FIT) for blood (28). The dataset consisted of quantitative FIT results and microbiome profiles on 101 cases of cancer and 141 controls (after excluding those with adenoma). For the predictive modelling, the number of variables in the microbiome profiles was restricted to 201 operational taxonomic units (OTUs) that had nonzero values in at least 20% of individuals. The Bayesian program *Stan* (29) was used to generate the posterior distribution of predictive probabilities from two alternative logistic regression models: a baseline model with FIT only and an uninformative prior on the effect parameter, and a model with FIT plus the microbiome markers, with a hierarchical shrinkage prior on the microbiome variables that allows the algorithm to learn that most effect sizes are near zero (30). The prediction of colorectal cancer in test data was evaluated by 40-fold cross-validation, with predictive probabilities evaluated as the average of 2000 posterior samples on each test fold. The densities were adjusted as described above to make them consistent, with reweighting parameter $\theta = 0.00018$.

Table 1 compares the model with FIT + microbiome profile to the model with FIT only. Including the microbiome profile increases the C-statistic from 0.892 to 0.932. This result might be misinterpreted as showing that the microbiome profile makes only a small incremental contribution to prediction when compared with a baseline model using FIT only. However the expected information for discrimination is approximately doubled from 3 to 6.5 bits when the microbiome profile is added to the model. The strength of evidence that this improves prediction can be evaluated as the difference in test log-likelihood, which is 60.2 bits.

In this example where one variable(FIT) accounts for half the expected information and the class-conditional distributions of this variable are far from Gaussian (most FIT values in controls are zero), we would not necessarily expect the weight of evidence to follow its asymptotic Gaussian distribution. Figure 2 shows the unadjusted estimates of the densities in cases and controls of the weight of evidence favouring case over control status are skewed, together with the densities adjusted as described above to make them consistent. The main effect of this adjustment is to shrink the left tail of the density in cases and the right tail of the density in cases. Thus, for instance at $W = -6$ bits where the true case/control density ratio is 1:64 and the unadjusted ratio is about 1:7, adjustment shrinks the density in cases and increases the density in controls

slightly. The model-based estimates of Λ and C, based on the adjusted densities, are higher than the crude estimates.

Figure 3 shows the adjusted cumulative frequency distributions. These can be used to evaluate how a combined test based on FIT and microbiome profile could be used for risk stratification in a clinical setting (for illustrative purposes only, not as a policy recommendation). For instance suppose that in a setting in which the prior probability of colorectal cancer in symptomatic individuals referred from primary care is 5% (prior odds 1:19), a threshold of at least 1% risk of cancer (posterior odds 1:99) has been set as the criterion for further investigation by colonoscopy. From the adjusted cumulative frequency distributions we can estimate that using this risk threshold (weight of evidence favouring case over noncase status $\log_2 19/99$ = -2.38 bits) with a combined test based on FIT and microbiome profile would exclude 2% of cancer cases and 88% of noncases as having posterior probability of cancer less than 1%.

This study illustrates also how the the projection predictive method (31; 32) can be used to select the most predictive variables. After evaluating predictive performance by cross-validation, 2000 posterior samples of the fitted values of the linear predictor were generated from a model with FIT + microbiome profile fitted to the full data and forward selection was performed using the projection predictive method. The increment in predictive information contributed by each additional biomarker was evaluated as the reduction in KL divergence of full-model fitted values from their projection on to the subspace of microbiome variables selected. Figure 4 shows that the predictive information in the microbiome profile is contributed by many variables of small effect.

Discussion

Although the expected information for discrimination (expected weight of evidence) is a natural measure of the information content of a test or experimental design that contrasts two alternative hypotheses, it has not been widely used for this purpose in biostatistics, except in genetic linkage analysis during the pre-genome era where the weight of evidence (lod score) was used to quantify support for linkage, and the expectation of the lod score (ELOD) was the accepted measure of the information content of a study design (33). Lee (1999) (34) suggested reporting the expected information for discrimination in cases and controls separately to quantify the performance of a test score, but assumed that likelihood ratios would be evaluated by tabulating frequencies of scores grouped into ordinal categories, rather than by using the predictive probabilities output by the classifier to evaluate the likelihood ratio as the ratio of posterior odds to prior odds. In practice, estimates of probability ratios based on grouping scores into bins would be unstable: if only an uncalibrated test score were available, it would

be better to fit a model (such as a logistic regression) that outputs predictive probabilities before computing the expected weight of evidence.

An alternative approach to quantifying the information content of an experiment or test is to calculate the expected gain of information on the outcome (35). In the context of a diagnostic test, this would be the KL divergence from the prior to the posterior distribution of case-control status, rather than the expected information for discrimination which is the KL divergence from the distribution of the predictors given incorrect assignment to their distribution given correct assignment of case-control status(36). Unlike the expected information for discrimination, the expected gain of information about the outcome is not additive for independent biomarkers, and depends on the prevalence of disease so cannot be generalized from one setting to another.

A key requirement for quantifying predictive performance is that it should be evaluated not on the training data used to learn the model but on test data not seen before. Unless a very large dataset is available in which a single test / training split provides both a training dataset adequate to learn an optimal predictive model and a test dataset large enough to estimate predictive performance accurately, the most efficient way to evaluate performance will be through cross-validation. Without internal validation (through cross-validation or a single test/train split), it is not possible to evaluate whether poor performance on a test dataset is attributable to lack of generalizability or to lack of predictive information in the original dataset. Several groups have recently produced guidelines for reporting the evaluation of risk predictors or diagnostics using biomarkers: REMARK (37), GRIPS (38), STARD (39), and TRIPOD (15), Although evaluation of predictive performance by cross-validation is mentioned in supplementary materials, the summary recommendations and checklists do not emphasize this critical point. Even where studies report using cross-validation to evaluate predictive performance, it is not always clear that the test data have not been used at some earlier stage to learn the model. A common malpractice is to use the full dataset for variable selection, before the split into test/training folds (40). The wider adoption of reproducible research requirements (41), may make it easier for readers to determine whether correct practice was followed.

As long as the learning algorithm generates predictive probabilities, the expected information for discrimination can be evaluated just as easily on "black-box" predictors such as kernel-based learning algorithms as on simple logistic regression models. However unlike the C-statistic which depends only upon how the predictor ranks cases and controls, the expected information for discrimination depends on calibrating the predictor so that the predictive probabilities equate to the observed frequencies of cases at each level of the predictors in the test dataset. For a linear model with likelihood in the exponential family, maximizing the likelihood guarantees that the model is correctly calibrated to the training data (21). Thus where the test and test and training datasets are random subsamples

of the original dataset, formed either by a single test/training split or by cross-validation, calibration is unlikely to be a problem. If a predictor is to be evaluated in a different setting to that in which it was developed, it will usually be necessary to recalibrate it by adding an intercept term on the scale of log odds so as to equate the observed and predicted number of cases.

The expected weight of evidence can be given an intuitive interpretation: for instance an expected weight of evidence of 3 bits implies that a "typical" result would be for the posterior odds in favour of the true case-control status to be eight times the prior odds. For a predictor that is based on a large number of independent biomarkers of small effect, the asymptotic distribution derived by Turing will hold and the expected information for discrimination will be enough to characterize fully the distributions of the weight of evidence in cases and controls. Means and variances of the estimated distributions of the weight of evidence in cases and controls, together with a plot of these distributions, should be reported to allow the reader to determine whether this asymptotic distribution holds. Even where it does not hold, the other advantages of using the expected weight of evidence - additivity of effects of independent predictors, and its intuitive interpretation - support its use as a summary measure of predictive performance. However to evaluate how the predictor will perform as a risk stratifier, the reader will need the distributions in cases and controls if these distributions do not have their asymptotic form. A plot of these distributions is thus more useful than a conventional plot of the ROC curve.

Visualizing these distributions shows something not widely appreciated: that however good the classifier, the distribution of the weight of evidence in favour of the wrong hypothesis has a tail that extends well to the right of zero. This is a corollary of Turing's result that the expectation of the Bayes factor in favour of the wrong hypothesis is 1: the distribution of this Bayes factor becomes more right-skewed as the expectation of the log Bayes factor (weight of evidence) becomes more negative (16). A practical and disconcerting consequence is that if a classifier has high performance, it will not often be wrong but when it is wrong it may be wildly wrong, giving a high likelihood ratio in favour of the wrong hypothesis. Thus if the weight of evidence has its asymptotic distribution, a diagnostic test that has an expected weight of evidence of 4 bits (equivalent to C-statistic of 0.95) will generate a likelihood ratio more than 8 to 1 in favour of the wrong assignment of disease status in 2% of individuals tested. While this may be acceptable for risk stratification, failure to appreciate the fallibility of the multivariate *in vitro* diagnostic tests now coming into use could have serious consequences in clinical practice.

Online resources

An R script to estimate the procedure described for estimating the distribution of weights of evidence is available at http://www.homepages.ed.ac.uk/pmckeigu/

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to this article.

References

[1] Wu PY, Cheng CW, Kaddi CD et al. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Transactions on Biomedical Engineering* 2017; 64(2): 263–273. DOI:10.1109/TBME.2016.2573285.

[2] Byrne S. A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electronic Journal of Statistics* 2016; 10(1): 380–393. DOI:10.1214/16-EJS1109. URL https://projecteuclid.org/euclid.ejs/1455715967.

[3] Pencina MJ, D'Agostino RB, Pencina KM et al. Interpreting incremental value of markers added to risk prediction models. *American Journal of Epidemiology* 2012; 176(6): 473–481. DOI:10.1093/aje/kws207.

[4] Pepe MS, Fan J, Seymour CW et al. Biases introduced by choosing controls to match risk factors of cases in biomarker research. *Clin Chem* 2012; 58(8): 1242–1251. DOI:10.1373/clinchem.2012.186007.

[5] Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008; 54(1): 17–23. DOI:10.1373/clinchem.2007.096529.

[6] Janes H, Longton G and Pepe M. Accommodating Covariates in ROC analysis. *The Stata Journal* 2009; 9(1): 17–39.

[7] Huang Y. Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case-control studies. *Biostatistics (Oxford, England)* 2016; 17(3): 499–522. DOI:10.1093/biostatistics/kxw003.

[8] Parikh CR and Thiessen-Philbrook H. Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease. *Journal of the American Society of Nephrology : JASN* 2014; 25(8): 1621–1629. DOI:10.1681/ASN.2013121300.

[9] Pencina MJ, D'Agostino RB, D'Agostino RB et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008; 27(2): 157–72; discussion 207–12. DOI:10.1002/sim.2929.

[10] Huang Y and Pepe MS. Biomarker evaluation and comparison using the controls as a reference population. *Biostatistics (Oxford, England)* 2009; 10(2): 228–244. DOI:10.1093/biostatistics/kxn029.

[11] Pepe MS, Feng Z, Huang Y et al. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol* 2008; 167(3): 362–368. DOI:10.1093/aje/kwm305.

[12] Hilden J and Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med* 2014; 33(19): 3405–3414. DOI:10.1002/sim.5804.

[13] Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol* 2011; 173(11): 1327–1335. DOI:10.1093/aje/kwr013.

[14] Pepe MS, Fan J, Feng Z et al. The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Statistics in Biosciences* 2015; 7(2): 282–295. DOI:10.1007/s12561-014-9118-0.

[15] 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. DOI:10.1186/s12916-014-0241-z.

[16] Good IJ. Weight of evidence: a brief survey. In Bernardo JM, DeGroot MH, Lindley DV et al. (eds.) *Bayesian Statistics*. Elsevier, 1985. pp. 249–270.

[17] Gneiting T and Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 2007; 102(477): 359–378. DOI:10.1198/016214506000001437. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214506000001437.

[18] Good IJ and Toulmin GH. Coding theorems and weight of evidence. *Journal of the Institute of Mathematics and Applications* 1968; 4.

[19] McKeigue P. Sample size requirements for learning to classify with high-dimensional biomarker panels. *Statistical Methods in Medical Research* 2017; : 962280217738807DOI:10.1177/0962280217738807.

[20] Austin PC and Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012; 12: 82. DOI:10.1186/1471-2288-12-82.

[21] Mackay DJ. *Information theory, inference and learning algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[22] Kansagara D, Englander H, Salanitro A et al. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011; 306(15): 1688–1698. DOI:10.1001/jama.2011.1515.

[23] Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS genetics* 2009; 5(7): e1000540. DOI:10.1371/journal.pgen.1000540.

[24] Johnson NP. Advantages to transforming the receiver operating characteristic (ROC) curve into likelihood ratio co-ordinates. *Statistics in Medicine* 2004; 23(14): 2257–2266. DOI:10.1002/sim.1835.

[25] DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44(3): 837–845.

[26] Chen W, Samuelson FW, Gallas BD et al. On the assessment of the added value of new predictive biomarkers. *BMC Medical Research Methodology* 2013; 13: 98. DOI:10.1186/1471-2288-13-98. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733611/.

[27] Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B (Methodological)* 1977; : 44–47.

[28] Baxter NT, Ruffin MT, Rogers MAM et al. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine* 2016; 8. DOI:10.1186/s13073-016-0290-3. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823848/.

[29] Carpenter B, Gelman A, Hoffman M et al. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 2017; 76(1): 1–32. DOI:10.18637/jss.v076.i01. URL https://www.jstatsoft.org/v076/i01.

[30] Piironen J and Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* 2017; 11(2): 5018–5051. DOI:10.1214/17-EJS1337SI. URL http://arxiv.org/abs/1707.01694. ArXiv: 1707.01694.

[31] Goutis C and Robert CP. Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika* 1998; 85(1): 29–37.

[32] Piironen J and Vehtari A. Projection predictive variable selection using Stan+R. *arXiv:150802502 [stat]* 2015; URL http://arxiv.org/abs/1508.02502. ArXiv: 1508.02502.

[33] Ott J. Major strengths and weaknesses of the lod score method. *Advances in Genetics* 2001; 42: 125–132.

[34] Lee WC. Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance. *International Journal of Epidemiology* 1999; 28(3): 521–525.

[35] Lindley DV. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics* 1956; 27: 986–1005.

[36] Hughes G. Information graphs for epidemiological applications of the Kullback-Leibler divergence. *Methods of Information in Medicine* 2014; 53(1): IV–VI.

[37] McShane LM, Altman DG, Sauerbrei W et al. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005; 97(16): 1180–1184. DOI:10.1093/jnci/dji237.

[38] Janssens ACJW, Ioannidis JPA, van Duijn CM et al. Strengthening the reporting of Genetic RIsk Prediction Studies: the GRIPS Statement. *PLoS Med* 2011; 8(3): e1000420. DOI:10.1371/journal.pmed.1000420.

[39] Bossuyt PM, Reitsma JB, Bruns DE et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ (Clinical Research ed)* 2015; 351: h5527. DOI:10.1136/bmj.h5527.

[40] Varma S and Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006; 7: 91. DOI:10.1186/1471-2105-7-91.

[41] Iqbal SA, Wallach JD, Khoury MJ et al. Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS Biology* 2016; 14(1): e1002333. DOI:10.1371/journal.pbio.1002333.

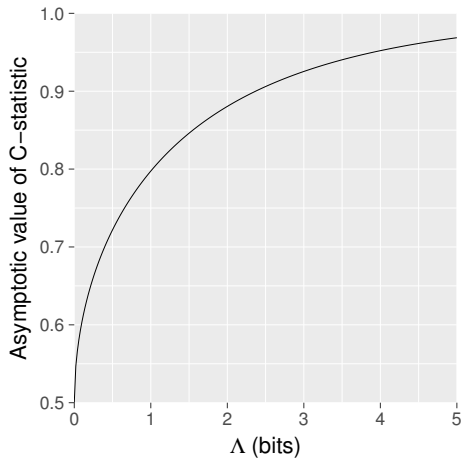

Figure 1. Asymptotic relationship of C-statistic to expected information for discrimination Λ

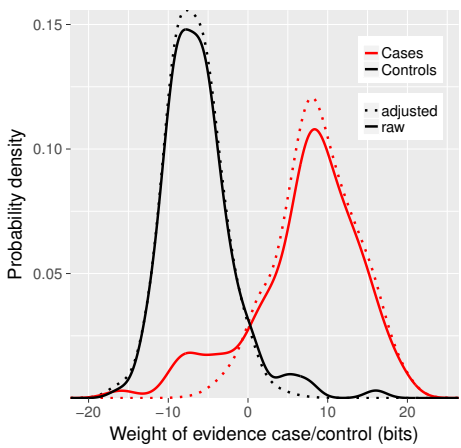

Figure 2. Distributions in cases and controls of weights of evidence favouring case over control status, from model combining FIT test with microbiome profile. Weights of evidence were computed on test folds by 40-fold cross-validation. Unadjusted densities were smoothed with a Gaussian kernel using bandwidth chosen by the Sheather-Jones algorithm. Adjusted densities were calculated from the mean of the unadjusted case and control densities as described in the text.

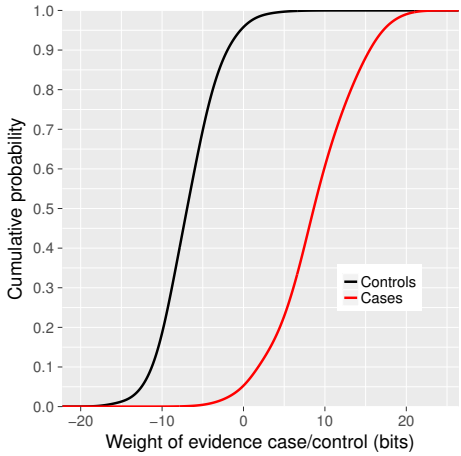

Figure 3. Adjusted cumulative distributions in cases and controls of weight of evidence.

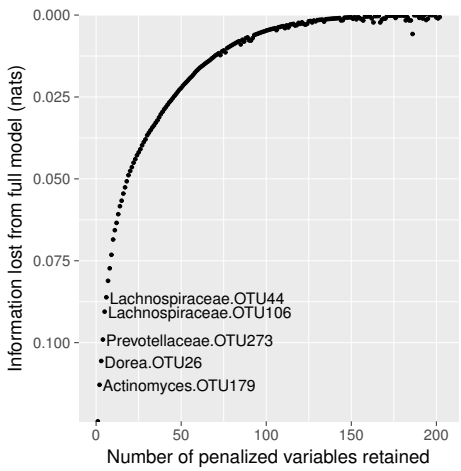

Figure 4. Proportion of total predictive information in microbiome profile obtained by forward selection of variables, using projective predictive method with posterior samples

Model	Crude C-statistic	Crude Λ (bits)	Adjusted C-statistic	Adjusted Λ (bits)	$\Delta \log \mathcal{L}$ (bits)
FIT only	0.892	3.0	0.930	3.0	0
FIT + microbiome	0.932	6.5	0.990	7.3	60.2

Table 1. Incremental contribution of microbiome profile to detection of colorectal cancer, compared with baseline model using faecal immunochemical test (FIT) only