Published in final edited form as:

Stat Methods Med Res. ; : 962280218819202. doi:10.1177/0962280218819202.

A multi-locus predictiveness curve and its summary assessment for genetic risk prediction

Changshuai Wei¹, Ming Li², Yalu Wen³, Chengyin Ye⁴, and Qing Lu⁵

¹Core Artificial Intelligence, Amazon.com Inc, Seattle, WA, USA ²Department of Epidemiology and Biostatistics, Indiana University at Bloomington, Bloomington, IN, USA ³Department of Statistics, University of Auckland, Auckland, New Zealand ⁴Department of Health Management, Hangzhou Normal University, Hangzhou, China ⁵Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA

Abstract

Genetic association studies using high-throughput genotyping and sequencing technologies have identified a large number of genetic variants associated with complex human diseases. These findings have provided an unprecedented opportunity to identify individuals in the population at high risk for disease who carry causal genetic mutations and hold great promise for early intervention and individualized medicine. While interest is high in building risk prediction models based on recent genetic findings, it is crucial to have appropriate statistical measurements to assess the performance of a genetic risk prediction model. Predictiveness curves were recently proposed as a graphic tool for evaluating a risk prediction model on the basis of a single continuous biomarker. The curve evaluates a risk prediction model for classification performance as well as its usefulness when applied to a population. In this article, we extend the predictiveness curve to measure the collective contribution of multiple genetic variants. We further propose a nonparametric, U-statistics-based measurement, referred to as the U-Index, to quantify the performance of a multi-locus predictiveness curve. In particular, a global U-Index and a partial U-Index can be used in the general population and a subpopulation of particular clinical interest, respectively. Through simulation studies, we demonstrate that the proposed U-Index has advantages over several existing summary statistics under various disease models. We also show that the partial U-Index can have its own uniqueness when rare variants have a substantial contribution to disease risk. Finally, we use the proposed predictiveness curve and its corresponding U-Index to evaluate the performance of a genetic risk prediction model for nicotine dependence.

Corresponding author: Qing Lu, Department of Epidemiology and Biostatistics, Michigan State University, 909 Fee Road, Room 601, East Lansing, 48824 MI, USA. qlu@epi.msu.edu.

Authors' Note

Changshuai Wei was affiliated with Department of Biostatistics and Epidemiology, University of North Texas Health Science Center, Fort Worth, TX, USA while working on the article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Keywords

High-dimensional data; genetic risk prediction; U-statistics; nicotine dependence

1 Introduction

Genome-wide association studies have discovered a large number of genetic variants associated with complex human diseases. Although the identified variants often have low to moderate effects, they can collectively have a significant impact on disease risk. Risk prediction models incorporating these genetic findings and other environmental risk factors hold great promise for potential clinical use. Meanwhile, it is crucial to comprehensively evaluate the performance of these prediction models to ensure their clinical validity and usefulness.

The receiver operating characteristic (ROC) curve has been a popular tool for evaluating the classification validity of a risk prediction model. It comprises all possible pairs of sensitivity and specificity and uses area under the curve (AUC) as a global summary statistic to summarize the discriminative ability of a risk prediction model. However, the ROC curve does not measure the predicted risk of disease for an individual and may have limited clinical utility.¹ As an alternative, the predictiveness curve was proposed to provide a comprehensive and cohesive assessment of both classification and risk prediction performance for a prediction model.² A predictiveness curve is constructed by plotting each individual's predicted risks versus its quantile in the population, or in other words, the transpose of the cumulative distribution function of the risk distribution. Therefore, the predictiveness curve would address the limitation of the ROC curve by displaying the predicted risk distribution on the population level.

Most existing studies have estimated the predictiveness curve on the basis of a single continuous biomarker and have evaluated its statistical property.^{3,4} Several summary statistics have been proposed, such as R-square, total gain, and average entropy.³⁻⁸ The R-square is defined as the ratio between the variance of predicted risks from the prediction model and that from a perfect predictor. The total gain and average entropy statistics can both be considered as generalized R-square statistics. The average entropy is based on the reduction in maximum loglikelihood, while the total gain statistic replaces the squared risk difference in the R-square with the absolute value of risk difference. These summary statistics have been very useful to evaluate the overall performance of a predictiveness curve. Intuitively, a larger variance of predicted risks indicates well-separated risks among individuals, and thus better prediction performance of the biomarker. Nevertheless, these summary statistics are all developed for a single continuous biomarker that has a monotonic relationship with disease risk. Complex human diseases are generally caused by multiple genetic variants, environmental factors, and their interactions. Therefore, there is a great need for developing appropriate statistics to assess the performance of a predictiveness curve for multiple categorical genetic variables, such as multi-locus genotypes.

U-Statistics are well known for flexibility and robustness,⁹ and have been recently applied to genetic association studies to evaluate the association between multiple genetic variants and

disease risk.^{10–15} For example, a Mann-Whitney U-statistic has been demonstrated to be an excellent summary statistic for ROC curves.¹⁶ In this article, we first propose a predictiveness curve based on multiple genetic variants with the consideration of possible interactions and further define a U-statistics-based measurement, the U-Index, to summarize its overall performance. In addition, we show that the proposed U-Index has a close connection with ROC curves and Lorenz curves and further explore the asymptotic property of the U-Index for statistical inference. Finally, we perform simulation studies to compare the proposed U-Index with other existing summary statistics and apply it to a real dataset to evaluate its performance for evaluating a risk prediction model of nicotine dependence.

2 Methods

2.1 Predictiveness curves based on a single continuous biomarker

Predictiveness curves have recently been proposed to graphically describe the contribution of a single continuous biomarker to a risk prediction model.² To construct a predictiveness curve, the risk of disease was estimated for each subject in the study, ordered from highest and lowest, and plotted against its risk percentile in the study population. While a commonly used ROC curve mainly focuses on classification accuracy, a predictiveness curve provides direct information of risk prediction. Its advantages have been discussed elsewhere.² Following a similar idea, we extended below the predictiveness curves for multiple categorical genetic variables.

2.2 Predictiveness curves based on multi-locus genotypes

For a single continuous biomarker, there exists an ordinal relationship between the biomarker and disease risks. For categorical genetic variables, it would be necessary to have an underlying order of multi-locus genotypes with respect to disease risks in a same manner as a continuous biomarker. To achieve this, we assume that the study population has N subjects, each genotyped for L single nucleotide polymorphisms (SNPs), for building a risk prediction model. Let $G_i = (S_{i,1}, S_{i,2}, \dots, S_{i,L})$ be the genotypes for the i -th subjects at L SNP loci, which forms a total of K multi-locus genotypes, i.e. $G_i \in \{g_1, g_2, \dots, g_K\}$. Further, let $p_k = P(g_k)$ denote the frequency of the k -th multi-locus genotype g_k ($1 \leq k \leq K$) and $r_k = P(D|g_k)$ be the predicted genetic risk for individuals carrying g_k . Assuming a disease prevalence of ρ in the population, we have

$$p_k = P(g_k|D)\rho + P(g_k|\bar{D})(1 - \rho) \text{ and } r_k = \frac{P(g_k|D)\rho}{P(g_k|D)\rho + P(g_k|\bar{D})(1 - \rho)}$$

where $P(g_k|D)$ and $P(g_k|\bar{D})$ are frequencies of multi-locus genotype g_k among the case and control populations, respectively. These frequencies, $P(g_k|D)$ and $P(g_k|\bar{D})$, can be estimated directly in a prospective cohort study as

$$P(g_k|D) = \frac{n_{D,k}}{n_D} \text{ and } P(g_k|\bar{D}) = \frac{n_{\bar{D},k}}{n_{\bar{D}}}$$

where n_D and $n_{\bar{D}}$ are the sample size of cases and controls, respectively, and $n_{D,k}$ and $n_{\bar{D},k}$ are the number of individuals carrying g_k in cases and controls, respectively. In population-based case-control studies where the controls are a representative group of the general population, the disease prevalence ρ can be obtained from the literature.

Without loss of generality, we assume that the order of multi-locus genotype g_k ($1 \leq k \leq K$) monotonically increases with that of its predicted genetic risk r_k (i.e. $r_k \leq r_{k'}$ for any $k < k'$). Otherwise, the multi-locus genotypes can be reordered to satisfy this assumption. The risk quantile of a multi-locus genotype can thus be calculated as

$$q_k = \sum_{t=1}^k p_t \quad \text{for } 1 \leq k \leq K$$

A predictiveness curve can then be constructed as a function that maps risk quantile q_k (x -axis) to its predicted risk r_k (y -axis). An example is provided in Figure 1.

2.3 U-Index: a summary statistic for a predictiveness curve

To measure the overall performance of a predictiveness curve, we propose a U-statistics-based summary statistic, referred to as the U-Index, as follows

$$U = 2 \sum_{1 \leq k < k' \leq K} p_k p_{k'} \psi(r_k, r_{k'}) \quad (1)$$

where a kernel $\psi(r_k, r_{k'})$ measures the difference of predicted risk for subjects carrying genotype profiles g_k and $g_{k'}$. Various forms of $\psi(r_k, r_{k'})$ can be chosen depending on the needs of the study (e.g. to incorporate cost of disease risk). In this article, we use a basic kernel function of risk difference

$$\psi(r_k, r_{k'}) = r_k - r_{k'}, \text{ for } 1 \leq k < k' \leq K$$

Equation (1) can also be generalized as

$$U = 2 \int_0^1 \int_0^y \psi(r(y), r(x)) dx dy = 2 \int_0^1 \int_0^y (r(y) - r(x)) dx dy \quad (2)$$

The U-Index in equation (2) could have a wide range and can be further standardized by dividing its theoretical maximum value (i.e. perfect prediction). As a result, the standardized U-Index ranges between 0 and 1 as

$$U^{st} = \frac{2 \int_0^1 \int_0^y (r(y) - r(x)) dx dy}{\rho(1 - \rho)} \quad (3)$$

When a multi-locus genotype g_k has N_k carriers in a population with N subjects, we have $p_k = \frac{N_k}{N}$ and the corresponding U-Index is equivalent to a one-sample U-statistic

$$U = \frac{\sum_{1 \leq k < k' \leq K} N_k N_{k'} \psi(r_k, r_{k'})}{N(N-1)/2} \quad (4)$$

Using Hoeffding projection, the variance of the above U-Index in the population is

$$\text{Var}(U) = \frac{4}{N} \text{Var}(E(\psi(r_1, r_2) | r_1)) = \frac{4}{N} \sum_{k=1}^K \frac{N_k}{N} \left(\sum_{k'=1}^K \frac{N_{k'}}{N} \psi(r_k, r_{k'}) - U \right)^2 \quad (5)$$

In clinical practice, it is often of particular interest to identify a subpopulation with a specific range of risk quantile, i.e. $q \in (q_0, q_1)$. For such a purpose, we further define a partial U-Index to evaluate the performance of a predictiveness curve in a subpopulation as

$$U_{pt} = 2 \int_0^{q_1} \int_{q_0}^y \psi(r(y), r(x)) dx dy = 2 \int_0^{q_1} \int_{q_0}^y (r(y) - r(x)) dx dy \quad (6)$$

The partial U-Index can also be standardized to be within the range of 0 and 1

$$U_{pt}^{st} = \frac{2 \int_0^{q_1} \int_{q_0}^y (r(y) - r(x)) dx dy}{\rho_{st}(1 - \rho_{st})}, \text{ where } \rho_{st} = \int_{q_0}^{q_1} r(x) dx \quad (7)$$

The variance of the partial U-Index can be calculated similar to that of the global U-Index in equation (5) with the consideration of constraint on the selected range of risk quantile.

2.4 Connection with the area under the ROC curve and the Lorenz curve

Figure 1 gives an example of a predictiveness curve, an ROC curve, and a Lorenz curve. As we discussed above, the predictiveness curve is constructed by plotting the predicted risk of multi-locus genotypes (r_k) against their corresponding quantile (q_k) in the population. The ROC curve has been popularly used to evaluate the classification accuracy of a risk prediction model.¹⁶ For a risk prediction model based on multi-locus genotypes that are ordered according to disease risk, we can define the threshold based on a multi-locus genotype g associated with a particular disease risk

$$\hat{Y}_G = \begin{cases} 1 & r_G > r_g \\ 0 & r_G \leq r_g \end{cases}$$

The ROC curve in Figure 1 is then constructed by plotting the sensitivity against $1 -$ specificity at various thresholds. The area under the ROC curve (AUC_R) is often used as a summary statistic for the ROC curve, where a higher value of AUC_R represents higher overall classification accuracy. In the Supplementary material Appendix, we showed that the proposed global U-Index has the following relationship with the AUC_R value,

$$U = 4\rho(1 - \rho)(AUC_R - 0.5) \quad (8)$$

Note that the value of $AUC_R - 0.5$ is equivalent to the area between the ROC curve and the diagonal line (i.e. corresponding to the ROC curve of a non-informative marker). Therefore, the U-Index in equation (8) can be interpreted as the net classification improvement of the ROC curve over random guessing multiplied by a factor related to disease prevalence.

The Lorenz curve is initially proposed to describe income inequality and can be extended to describe risk inequality in the population. The Lorenz curve in Figure 1 was constructed by plotting the percentage of cumulative risk ($h(q) = \int_0^q r(u)du\rho$) against the risk quantile (q) of all individuals. Different from the ROC curve, the Lorenz curve should be convex. The area under the Lorenz curve (AUC_L) is often used as a summary statistic for the level of risk inequality in the population, where a lower value of AUC_L represents higher level of risk inequality. In the Supplementary material Appendix, we show that the proposed U-Index has the following relationship with AUC_L value

$$U = 2\rho(0.5 - AUC_L)$$

Note that the value of $0.5 - AUC_L$ is equivalent to the area between Lorenz curve and the diagonal line that represents no risk inequality in the population (i.e. all individuals have the same risk). Therefore, the U-Index in equation (8) can also be interpreted as the level of risk inequality in the population multiplied by a factor related to disease prevalence.

2.5 Statistical inference of U-index

In practice, it is common to build a risk prediction model based on a training dataset and to further evaluate its performance on an independent testing dataset. To construct a predictiveness curve and estimate its corresponding U-Index, we will first determine the order of multi-locus genotypes using the training dataset and estimate the predicted risk for each multi-locus genotype in the testing dataset by using the same order. Note that the order of multi-locus genotypes only ensures the monotonically increasing disease risks in the training dataset but does not guarantee monotonically increasing disease risks in the testing dataset. Under the null hypothesis of no predictiveness ability, the U-Index on the testing dataset asymptotically follows a normal distribution with a mean of zero. In the Supplementary material Appendix, we show that the U-Index on a testing dataset from a population-based case-control study is equivalent to a two-sample U-statistic below,

$$\hat{U} = 2\rho(1 - \rho) \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \phi(G_i, G_j)$$

where G_i and G_j are the multi-locus genotypes for the i -th case subject and the j -th control subject, and n_D and $n_{\bar{D}}$ are the samples size of cases and controls, respectively. The kernel function $\phi(G_i, G_j)$ is defined as

$$\phi(G_i, G_j) = \begin{cases} 1 & r(G_i) > r(G_j) \\ 0 & r(G_i) = r(G_j) \\ -1 & r(G_i) < r(G_j) \end{cases}$$

The asymptotic variance of the U-Index can be calculated by

$$\text{Var}(\hat{U}) = 4\rho^2(1 - \rho)^2$$

$$\left[\frac{1}{n_D(n_D - 1)} \left(\sum_{i=1}^{n_D} \left(\frac{1}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \phi(G_i, G_j) - \hat{U} \right)^2 + \frac{1}{n_{\bar{D}}(n_{\bar{D}} - 1)} \sum_{j=1}^{n_{\bar{D}}} \left(\frac{1}{n_D} \sum_{i=1}^{n_D} \phi(G_i, G_j) - \hat{U} \right)^2 \right] \right]$$

To test the significance of a U-Index on the training data, the above asymptotic distribution is no longer valid, but an empirical p -value can be obtained by using the permutation method. A bootstrap approach can also be used to obtain the empirical confidence interval.

3 Results

3.1 Simulation studies

We first conducted simulation studies to evaluate the performance of the proposed U-Index, and compared the U-Index to three existing summary statistics: average entropy, R-square, and total gain. We further compared the performance of the global U-Index with that of the partial U-Index when causal variants are mainly rare variants.

3.1.1 Simulation I—We simulated a population of 1 million subjects. The disease status of each subject was simulated based on four common SNPs (i.e. minor allele frequency > 0.01) with complex interactions. We considered four underlying disease models by gradually increasing disease heritability. The underlying predictiveness curves for four models are presented in Figure 2. In each simulation replicate, we sampled 1000 subjects as the study population and calculated the corresponding U-Index and three other summary statistics as follows

$$\text{Average Entropy: } AE = \int r(x) \log r(x) dx + \int (1 - r(x)) \log(1 - r(x)) dx$$

$$R\text{-square: } R^2 = \int (r(x) - \rho)^2 dx$$

$$\text{TotalGain: } TG = \int |r(x) - \rho| dx$$

For each disease model, the simulations were repeated 1000 times. Because these summary statistics had distinct measurement scales, it was not straightforward to directly compare their performance on the basis of the mean and standard deviation. Instead, we calculated “% of Bias” and “% of Non-coverage” to compare the accuracy and variation of the summary statistics. We calculated “% of Bias” by taking the ratio of the bias and true value and “% of Non-coverage” by counting the percentage of the times that the true value lies outside the 95% confidence interval obtained through bootstrap sampling.

The simulation results are summarized in Table 1. Overall, the “% of Bias” of the U-Index was the lowest among the four summary statistics, except under the last model of relatively high heritability. In such a scenario, the “% of Bias” of the U-Index was slightly higher than that of total gain, both of which were lower than 1% (i.e. 0.17% vs. 0.46%). We found a similar trend for “% of Non-coverage.” As the heritability increased, TG, R2, and AE gained limited improvement, while the U-Index consistently had the lowest “% of Non-coverage.”

It is worthwhile to note that the U-Index takes the order of multi-locus genotypes into consideration and is robust when the predictiveness curve is non-monotone in the testing data. This can partially explain the advantage of the U-Index over the other three summary statistics. To have a fair comparison with the other summary statistics without such a property, we further use an isotonic regression to refit the predictiveness curve before calculating each summary statistic. By doing so, the predictiveness curve was ensured to be monotonic in testing data. The updated simulation results are summarized in Table 2. We found that the “% of bias” of the U-Index was still lowest among the four summary statistics for all disease models. In regard to “% of Non-coverage,” the performance of the U-Index was comparable with that of TG, both of which were much lower than those of R2 and AE. The “% of Non-coverage” of the U-Index was slightly lower than that of TG when the heritability of disease was relatively low (i.e. 2% and 5%) but was slightly higher when the heritability of disease was relatively high (i.e. 10% and 20%).

3.1.2 Simulation II—It is often of particular clinical interest to focus on a subpopulation within certain risk quantiles rather than the entire population. For such a purpose, we propose a partial U-Index. To compare its performance with that of the global U-Index, we simulated a source population of 1 million subjects and further considered three disease models, each involving two common variants and two rare variants (i.e. minor allele frequencies < 0.01). We fixed the effects of the common variants at a moderate level but varied the effects of rare variants from a moderate level to a high level. The underlying predictiveness curve is presented in Figure 3. As demonstrated in Figure 3, rare variants play

a major role for a subpopulation at a very high risk level. In each simulation, we used a partial U-Index to summarize the predictiveness curve for the highest 10% risk quantile and compared its performance to that of the global U-Index. The simulation was repeated 1000 times for each disease model.

The simulation results are summarized in Table 3. When the effect sizes of rare variants increased, the performance of both the global and the partial U-Index improved, leading to lower “% of Bias” and “% of Non-coverage.” The partial U-Index consistently had lower “% of Bias” and “% of Non-Coverage” than the global U-Index in all simulations. We believe that this result was due to the fact that the contributions of rare variants can only be identified within a subpopulation of high disease risk under these disease models. The partial U-Index would gain a unique advantage over the global U-Index by focusing on subjects within high risk quantiles.

3.2 Application to nicotine dependence

Tobacco-attributable morbidity and mortality represents a remarkable 21st century global burden of disease. In the USA, the lifetime prevalence of nicotine dependence (ND) is about 24%.¹⁷ The etiology of ND is highly complex, involving both genetic and environmental factors as well as their interactions. Family-based studies have estimated the heritability of ND to be around 59%.¹⁸ Despite a relatively high heritability, the currently identified SNPs have not contributed significantly to the risk prediction of ND. In this article, we combined genetic variants and environmental factors to build a risk prediction model for ND, using the large-scale dataset from the Study of Addiction: Genetics and Environment (SAGE). The participants were unrelated individuals selected from two large, complementary studies: the Family Study of Cocaine Dependence (FSCD) and the Collaborative Genetic Study of Nicotine Dependence (COGEN). We used FSCD as the training data to build a risk prediction model and further validated the model in COGEN. ND as defined in the *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) was used as a binary outcome for analysis. In particular, there are 486 cases and 684 controls in FSCD and 696 cases and 708 controls in COGEN. From the literature, we selected 37 SNPs that were found to associate with ND. A total of seven non-genetic risk factors, including sex, age, education, income, and three types of trauma experience, were also included in the risk prediction model. To determine a parsimonious model, we applied a forward selection algorithm with a built-in cross-validation procedure to form a risk prediction model based on the 44 predictors in the training data.¹⁶ The algorithm selected three risk predictors: income, physical trauma, and a genetic variant, rs2656073, for the final risk prediction model. The predictiveness curve of the formed risk prediction model can be visualized in Figure 4. The model has moderate prediction accuracy, with a global U-Index of 0.160. We further evaluated the performance of this risk prediction model in the testing data. The predictiveness curve from the testing data has a pattern similar to that of the training data, with a global U-Index of 0.115.

To further evaluate the performance of the method, we sequentially added two predictors into the risk prediction model through forward selection (Figure 4). Consistent with the 10-fold cross-validation results, the prediction accuracy calculated from the testing data

declined when additional risk predictors were added. For the 5-predictor model, the predictiveness curve based on the testing data maintained the increasing trend ($U = 0.112$), but with a lot of irregularity. For the 7-predictor model, we could hardly observe any increasing trend ($U = 0.0385$). The results showed that when the model was overfitted, the performance of the risk prediction model declined in the testing data. In Figure 5, we show the trend of four summary statistics (i.e. U , R^2 , AE , TG) when model complexity increases. The U -Index had a decreasing value with some vibrations due to model over-fitting. However, the other three summary statistics (i.e. R^2 , AE , and TG) had increasing values (i.e. suggesting a better model), even when the model was overfitted.

4 Discussion

The translation of human genome discoveries into public health practice is a major goal in the coming decades. The use of emerging findings from genome-wide association studies and sequencing studies for disease prediction, early intervention, and pharmacogenetics will advance genome medicine and lead to more effective personalized prevention and treatment strategies. A predicted risk of disease is helpful for physicians to make personalized clinical consultations. Meanwhile, to evaluate the impact of a risk prediction model in the population, one needs to consider the distribution of the predictors. However, the widely used tools, such as ROC curves, have been suggested to have limitations in providing information on a population level.^{19–21} As an alternative, predictiveness curves have been proposed to access the usefulness of a risk prediction model on the population level. In this article, we extended the predictiveness curve for multiple risk predictors. We also provided two summary statistics, a global U -Index and a partial U -Index, to assess the performance of a predictiveness curve. A global U -Index can be calculated to evaluate the prediction performance in the entire population, while a partial U -Index can be used to address the clinical interest in a subpopulation (e.g. high-risk population). Although we illustrate the U -Index based on a formed risk prediction model, the proposed U -Index can also be used for parameters tuning in the model building process. For instance, given the connection between the U -Index and the AUC, we can directly use the U -Index in the cross-validation process to determine the hyper-parameters (e.g. tree depth) and select the parsimonious tree-based model.²²

In this article, we used a risk difference kernel for the U -Index and compared it with three other commonly used summary statistics: R -square, average entropy, and total gain. Through simulation studies, we demonstrated that R^2 , AE , and TG were subject to bias, especially when the heritability is moderate or low, while the U -Index had robust performance. This is because the U -Index can capture the monotonicity of the predictiveness curve based on multi-locus genotypes. The predictiveness curve was first proposed to evaluate prediction performance for a single continuous biomarker assuming a monotonic relationship with disease risk. However, for categorical multi-locus genotypes, the order depends on the underlying disease model. A predictiveness curve might not be monotonically increasing in the validation dataset. The U -Index can capture the non-monotonic information, and thus it gains an advantage over the other summary statistics. Although one can use smoothing approaches, such as isotonic regression, to make a predictiveness curve monotonically increasing, this additional procedure often leads to loss of information.

The U-Index is constructed from U-statistics. Since introduced by Hoeffding, U-statistics have been widely used for nonparametric analysis and have been recently adopted in genetic association study. The proposed U-Index is supposed to be an unbiased estimator of the selected kernel function $\psi(r_k, r_{k'})$. By using a kernel function of risk difference, i.e., $\psi(r_k, r_{k'}) = r_k - r_{k'}$, the U-Index can be interpreted as the expected risk difference in the entire population. If a set of genetic variants can predict the disease status well, the risk difference between a group with high risk and a group with low risk is expected to be relatively large. The standardized U-Index ranges between 0 and 1, which measures the proportion of risk difference explained by the model. By taking the underlying order of genetic variants into account, it can also serve as a goodness-of-fit measure of the risk prediction model.

Although the U-Index aims to evaluate a risk prediction model at the population level, the results should be interpreted with caution. A large value for the U-Index obtained in one population may not always be generalized to another population. The genetic effects of variants could vary in different populations. The allele frequencies and genotype distributions of these variants might also differ, which could further influence the prediction accuracy. As a result, the values of the U-Index in two heterogeneous populations are expected to differ. Furthermore, a small value does not necessarily mean genetic variants are useless in disease prediction. In fact, many high penetrant and rare variants may not impact the overall risk distribution in the entire population but could have a high impact on the high-risk population. Through simulations, we demonstrated that a partial U-Index is more suitable for evaluating the contribution of rare variants.

The proposed U-Index can be directly applied to prospective cohort studies. It can also be applied to population-based case-control studies. In our real data application, we analyzed two case-control datasets by assuming a ND prevalence of 24%. However, one should be cautious when inferring population predictiveness from case-control studies. The study samples might not be representative of the entire population, due to eligibility criteria and sampling variability. Furthermore, the population substructure and the correlation between subjects can lead to false-positive results. Different studies on the same population might end with a different prediction model owing to heterogeneity in studies (e.g. different genetic array and quality control process).

Many kernel functions can be used to define a U-Index. The choice of kernel functions depends on the research questions of interest. For different diseases, the cost associated with the disease risk can vary. Heavy weights can be given to a high-risk group if the high-risk group has a big impact. Even for the same disease, the impact can vary according to the characteristics of the population (e.g. medical and economic condition). The information can be incorporated in the kernel function if needed. For example, a weighted kernel can be defined through a general form as

$$\psi(r_k, r_{k'}) = \omega(r_k)r_k - \omega(r_{k'})r_{k'}, \text{ for } 1 \leq k < k' \leq K,$$

where a weight function $\omega(\cdot)$ can be chosen based on the need of the study (e.g. cost or impact related to the disease risk). When $\omega(\cdot)$ monotonically increases, the above kernel function would give more weight to high-risk groups than to low-risk groups. In this article, we attempted to build a risk prediction model for ND. The primary interest was to find an optimal genetic model that can differentiate the high-risk group from the low-risk group in the population. We believe a kernel function of risk difference is suitable for this purpose. It should be noted, however, that the unweighted kernel function (i.e. $\omega(\cdot) = 1$) has a direct connection with the area under the ROC curve and the Lorenz curve. Such a connection may not exist if a different kernel is used.

We included 37 previously reported SNPs to build a risk prediction model for ND. To increase the prediction accuracy, we also included several environmental risk factors. The optimal model selected three predictors: income, physical trauma, and a genetic variant (rs2656073), which attained a moderate value of the U-Index in the testing dataset. Further investigation of the final model revealed that the genetic variant only had a small effect on the overall prediction. This is reasonable since individual variants usually have small marginal effects for psychiatric disorders. Due to the limitations of the available information in the dataset, only seven environmental risk predictors were included in the analysis. The risk prediction model can be further improved by including additional genetic and environmental risk factors from other studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported, in part, by the National Institute on Drug Abuse under Award Number R01DA043501, the National Library of Medicine under Award Number R01LM012848, the National Heart, Lung and Blood Institute under Award Number K01HL140333, and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development under Award Number R03HD092854. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Moons KGM and Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Acad Radiol* 2003; 10: 670–672. [PubMed: 12809422]
2. 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: 362–368. [PubMed: 17982157]
3. Huang Y, Sullivan Pepe M and Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics* 2007; 63: 1181–1188. [PubMed: 17489968]
4. Huang Y and Pepe MS. Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. *Stat Med* 2010; 29: 1391–1410. [PubMed: 20527013]
5. Shapiro AR. The evaluation of clinical predictions. A method and initial application. *N Engl J Med* 1977; 296: 1509–1514. [PubMed: 325407]
6. Bura E and Gastwirth JL. The binary regression quantile plot: Assessing the importance of predictors in binary regression visually. *Biometric J* 2001; 43: 5–21.

7. Mittlbock M and Schemper M. Explained variation for logistic regression. *Stat Med* 1996; 15: 1987–1997. [PubMed: 8896134]
8. Cox DR and Wermuth N. A comment on the coefficient of determination for binary responses. *Am Stat* 1992; 46: 1–4.
9. Hoeffding W A class of statistics with asymptotically normal distribution. *Ann Math Stat* 1948; 19: 293–325.
10. Schaid DJ, McDonnell SK, Hebring SJ, et al. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* 2005; 76: 780–793. [PubMed: 15786018]
11. Wei Z, Li M, Rebbeck T, et al. U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet* 2008; 72: 821–833. [PubMed: 18691161]
12. Li M, Ye C, Fu W, et al. Detecting genetic interactions for quantitative traits with U-statistics. *Genet Epidemiol* 2011; 35: 457–468. [PubMed: 21618602]
13. Lu Q, Wei C, Ye C, et al. A likelihood ratio-based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. *Genet Epidemiol* 2012; 36: 583–593. [PubMed: 22760990]
14. Li M, He Z, Schaid DJ, et al. A powerful nonparametric statistical framework for family-based association analyses. *Genetics* 2015; 200: 69–78. [PubMed: 25745024]
15. Wei C and Lu Q. A generalized association test based on U statistics. *Bioinformatics* 2017; 33: 1963–1971. [PubMed: 28334117]
16. Lu Q and Elston RC. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *Am J Hum Genet* 2008; 82: 641–651. [PubMed: 18319073]
17. Breslau N, Johnson EO, Hiripi E, et al. Nicotine dependence in the United States: prevalence, trends, and smoking persistence. *Arch Gen Psychiatry* 2001; 58: 810–816. [PubMed: 11545662]
18. Li MD, Cheng R, Ma JZ, et al. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* 2003; 98: 23–31. [PubMed: 12492752]
19. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115: 928–935. [PubMed: 17309939]
20. Pepe MS, Feng Z and Gu JW. Comments on ‘Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al.. *Stat Med* 2008; 27: 173–181. [PubMed: 17671958]
21. Gail MH and Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; 6: 227–239. [PubMed: 15772102]
22. Lu Q, Obuchowski N, Won S, et al. Using the optimal robust receiver operating characteristic (ROC) curve for predictive genetic tests. *Biometrics* 2010; 66: 586–593. [PubMed: 19508241]

Figure 1.
Examples of predictiveness curve, ROC curve, and Lorenz curve.

Figure 2.
The predictiveness curves of four disease models with varying heritability.
From left to right: the heritability of the disease increases from 2% to 20%.

Figure 3.

The predictiveness curves of three disease models with varying contributions from rare variants. From left to right: rare variants with moderate effects, moderate high effects, and high effects.

Figure 4.

Estimated predictiveness curves based on the SAGE data under varying levels of model complexities.

Left column from top to bottom: prediction models with three predictors, five predictors, and seven predictors in FSCD. Right column from top to bottom: prediction models with three predictors, five predictors, and seven predictors in COGEND.

Figure 5. Performance of U-Index and other three summary statistics (i.e. R-square, Total Gain, and Average Entropy) when disease prediction model has an increase in complexity (i.e. from 3 predictors to 15 predictors).

Table 1.

Performance of U-Index and 3 other summary statistics in simulation I.

Heritability	U	R2	TG	AE
0.02				
True Value	0.199	0.00223	0.142	0.0125
Mean	0.194	0.00666	0.175	0.0225
SD	0.017	0.00233	0.013	0.0031
% of Bias	2.51	198	22.9	79.9
% of Non-coverage	6.60	100	100	100
0.05				
True Value	0.312	0.00611	0.225	0.0317
Mean	0.305	0.01280	0.248	0.0427
SD	0.018	0.00373	0.014	0.0048
% of Bias	2.26	109	10.4	34.8
% of Non-coverage	9.70	100	91.2	100
0.10				
True Value	0.437	0.0145	0.328	0.0654
Mean	0.436	0.0244	0.346	0.0791
SD	0.016	0.0055	0.0138	0.0067
% of Bias	0.14	67.8	5.31	21
% of Non-coverage	7.40	99.8	65.6	89
0.20				
True Value	0.599	0.0410	0.475	0.138
Mean	0.596	0.0576	0.476	0.153
SD	0.0138	0.0108	0.0138	0.01
% of Bias	0.464	40.5	0.173	11.2
% of Non-coverage	7.20	99.3	13.8	88.5

AE: average entropy; R2: R-square; TG: total gain; U: U-Index. Values in bold highlight better performance of U-index than other summary statistics.

Table 2.

Performance of U-Index and three other summary statistics when the underlying predictiveness curve is restricted to being monotonic.

Heritability	U	R2	TG	AE
0.02				
True Value	0.199	0.00223	0.142	0.0125
Mean	0.202	0.00302	0.146	0.0141
SD	0.017	0.00103	0.014	0.0024
% of Bias	1.39	35.0	2.29	12.9
% of Non-coverage	7.70	39.2	9.10	26.2
0.05				
True Value	0.312	0.00611	0.225	0.0317
Mean	0.312	0.00795	0.228	0.0337
SD	0.018	0.00227	0.015	0.0042
% of Bias	0.006	30.0	1.39	6.42
% of Non-coverage	8.10	44.9	9.20	21.3
0.10				
True Value	0.437	0.0145	0.328	0.0654
Mean	0.442	0.0179	0.334	0.0694
SD	0.016	0.0039	0.015	0.0060
% of Bias	1.29	23.2	1.75	6.13
% of Non-coverage	11.9	42.4	9.90	25.8
0.20				
True Value	0.599	0.0410	0.475	0.138
Mean	0.601	0.0476	0.471	0.142
SD	0.014	0.0086	0.014	0.009
% of Bias	0.399	15.9	0.759	3.25
% of Non-coverage	8.90	32.1	7.90	18.7

AE: average entropy; R2: R-square; TG: total gain; U: U-Index.

Table 3.

Performance of global U-Index and partial U-Index when rare variants have a major contribution to disease risk.

	<u>Moderate effect of rare variants</u>		<u>Intermediate effect of rare variants</u>		<u>Large effect of rare variants</u>	
	U	U_{pt}	U	U_{pt}	U	U_{pt}
True Value	0.0298	0.103	0.0432	0.243	0.0527	0.312
Mean	0.0201	0.130	0.0325	0.251	0.0422	0.316
SD	0.0147	0.061	0.0142	0.0563	0.0147	0.054
% of Bias	32.6	26.4	24.7	3.3	20	1.38
% of Non-Coverage	67.7	41.2	35.2	1.6	16.7	0