Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(12) 3492–3510

DOI: 10.1177/0962280220929039
journals.sagepub.com/home/smm
SAGE

Criteria for evaluating risk prediction of multiple outcomes

Frank Dudbridge

Abstract

Risk prediction models have been developed in many contexts to classify individuals according to a single outcome, such as risk of a disease. Emerging "-omic" biomarkers provide panels of features that can simultaneously predict multiple outcomes from a single biological sample, creating issues of multiplicity reminiscent of exploratory hypothesis testing. Here I propose definitions of some basic criteria for evaluating prediction models of multiple outcomes. I define calibration in the multivariate setting and then distinguish between outcome-wise and individual-wise prediction, and within the latter between joint and panel-wise prediction. I give examples such as screening and early detection in which different senses of prediction may be more appropriate. In each case I propose definitions of sensitivity, specificity, concordance, positive and negative predictive value and relative utility. I link the definitions through a multivariate probit model, showing that the accuracy of a multivariate prediction model can be summarised by its covariance with a liability vector. I illustrate the concepts on a biomarker panel for early detection of eight cancers, and on polygenic risk scores for six common diseases.

Keywords

Risk prediction, multiplicity, screening, multivariate analysis, biomarkers, polygenic risk score

1 Introduction

Risk prediction is important in many medical contexts in which prediction models can guide decision making.[1] Examples include primary prevention, such as cholesterol reduction in subjects at risk of cardiovascular disease[2]; secondary prevention, such as the targeted enrolment of individuals into screening programmes[3]; allocation of treatment according to prognosis[4]; and differential diagnosis.[5] In general, models are constructed with the prediction of a single discrete outcome in mind. Thus models for identifying individuals at risk of, for example, breast cancer,[6] cardiovascular disease[7] and diabetes[8] have been developed by separate research communities with different study cohorts, although the models may share some variables and identify some of the same individuals as at risk. Consequently, evaluation of prediction models is also done according to single outcomes.

The emergence of "-omic" and other molecular biomarkers has raised the prospect of panels of features that can simultaneously predict multiple outcomes from a single biological sample. For example, a blood test called CancerSEEK has been proposed for early detection of eight cancers from circulating proteins and tumour DNA mutations.[9] Genome-wide association studies have garnered particular attention as many diseases are heritable and the DNA sequence is fixed throughout life. Because many diseases are influenced by numerous variants across

Department of Health Sciences, University of Leicester, Leicester, UK

Corresponding author:
Frank Dudbridge, Department of Health Sciences, University of Leicester, Leicester LE1 7RH, UK.
Email: frank.dudbridge@leicester.ac.uk

the entire genome, genetic risk can be efficiently measured with a generic micro-array,[10] and in principle could be calculated for multiple conditions at any point in life. Epigenetic variation may also provide useful risk stratification and has been advocated for the early detection of several cancers.[11] Furthermore, the emergence of large, broadly phenotyped cohorts such as UK Biobank[12] provides useful resources for developing and evaluating such models.

Apart from the practical efficiencies of conducting several assessments in parallel, simultaneous prediction has other potentially useful applications. Individuals may be more concerned about their risk across a range of conditions rather than of one in particular, a demand increasingly targeted by direct-to-consumer genetic testing companies.[13] Furthermore, some interventions may be effective for several conditions, and identification of individuals at increased risk of any of them may lead to greater impact of such interventions. As a simple example, body mass index is associated with several diseases with otherwise distinct causes, including coronary heart disease, type-2 diabetes, breast cancer and depression.[14] A weight loss intervention might be more effective when targeted to those at increased risk of any of those conditions. Similarly, evidence that aspirin usage could reduce the risk of various cancers[15] as well as of cardiovascular disease suggests that risk prediction for a set of diseases could be of benefit. More speculatively, forensic applications could utilise simultaneous prediction of phenotypes from anonymous DNA samples.[16,17]

Prediction of this nature is already done informally using recurrent risk factors such as age, gender, smoking and blood pressure. For example, in the UK the NHS Health Check is offered to individuals aged between 40 and 74 on account of the strong association of age with risk of stroke, kidney disease, heart disease, type 2 diabetes and dementia. For such risk factors, their strength of association and ease of measurement obviate any need for formal evaluation over many outcomes. But for emerging risk factors it is less clear whether their utility is enhanced by their potential to predict multiple outcomes. There are problems of multiplicity reminiscent of those in exploratory hypothesis testing, but a framework is lacking for addressing these issues in the context of risk prediction.

Prediction of multiple outcomes can be distinguished from prediction of a single composite outcome. Composite outcomes have been used to group related conditions, such as cardiovascular disease,[18] and to define outcomes of specific interest such as frailty and all-cause mortality.[19] Prediction of such outcomes may be viewed as a crude form of multiple outcome prediction: here I consider the composite evaluation of multiple predictions, rather than the evaluation of a single composite prediction. Composite evaluation may offer improved accuracy over a composite outcome; pragmatically it can use predictors developed individually for each outcome without the need to develop a specific predictor for their composite.

Several authors have studied the statistical modelling of a multivariate response, using methods such as partial least squares[20] and multivariate linear regression.[21–23] While it is recognised that prediction can be improved by exploiting correlation among responses, the literature has emphasised methods to improve model fitting, with accuracy typically measured by squared error metrics for each response marginally[22,24] or in total across responses.[21] This may be adequate in applications such as chemometrics and genetic selection where the responses are quantitative, but is less satisfying for prediction of discrete outcomes. Here I am not concerned with model fitting *per se* but in evaluating models, however estimated, in the context of their joint risk predictions. There is some work on mutually exclusive events, such as polytomous outcomes[25] and competing risks,[26] but general vectors of dichotomous outcomes have not been studied.

Here I propose definitions of some basic criteria for evaluating risk prediction models of multiple outcomes. The evaluation of single outcome models, while not a settled question, has at least a standard set of core criteria that serve as a basis for more nuanced assessment.[27] The present aim is to propose a similar set of core criteria as a starting point for the development of more refined approaches. I do not aim to give a complete account of multiple outcome prediction, but to identify and open discourse around some basic issues in this emerging area.

In section 2, I identify four senses in which multiple predictions can be evaluated, termed outcome-wise, joint, and weak and strong panel-wise. Examples are given in which each sense of prediction may be appropriate. I define sensitivity, specificity, concordance, and relative utility in each of these senses. In section 3, I develop analytical expressions for each of these quantities from a multivariate probit model. These show that the accuracy of a multivariate prediction model can be summarised by its covariance with a liability vector, and from this covariance matrix all the proposed criteria can be derived. Section 4 applies the results to some examples of current applications, and uses the model of section 3 to project their future performance as improved predictors are developed. Section 5 provides some discussion.

2 Definitions

2.1 Preliminaries

For individual $i = 1, \dots, N$, let $\boldsymbol{D}_i \in \{0, 1\}^m$ be a vector of binary indicators for m dichotomous outcomes. Say that outcome j *did occur* when the j-th element of $\boldsymbol{D}_i$ is 1, and the outcome *did not occur* when that element is 0. Similarly to Gail and Pfeiffer,[28] define the vector $\boldsymbol{\pi}_i$ whose j-th component is the probability of outcome j in individual i. Where necessary, components are identified by brackets: for example, $\pi_{i[j]}$ denotes the j-th component of $\boldsymbol{\pi}_i$. Let $\boldsymbol{X}_i$ be a vector of predictors and consider a *marginal risk prediction model* $r(\boldsymbol{x})$ as a mapping from the set Ω of possible values of $\boldsymbol{X}_i$ to $[0, 1]^m$. The model is understood as marginal in that, reflecting much current practice, $r(\boldsymbol{x})$ provides a risk prediction for each outcome but not for combinations of outcomes. In particular, correlations between outcomes may arise from comorbidity, competing risks or other sources, so that outcome-specific predictions may not be easily combined into predictions for groups of outcomes.

As for single outcome prediction, calibration is a desirable property of a risk predictor, and it will be generally useful for the predictor to be calibrated for all outcomes. Informally, calibration requires that predicted risks equal actual risks, but a distinction can be made between the risk among individuals with given predictors $\boldsymbol{x}$, and risk among individuals with given predictions $r(\boldsymbol{x})$. These quantities may differ if $r(\boldsymbol{x})$ has the same value for many values of $\boldsymbol{x}$, as in the case of a risk score formed as a linear combination of many predictors.[29]

Definition 1: *The risk prediction model* $r(\boldsymbol{x})$ *is strongly calibrated if* $E(\boldsymbol{D}|\boldsymbol{x}) = E(\boldsymbol{\pi}|\boldsymbol{x}) = r(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \Omega$. *The predictor is weakly calibrated if* $E(\boldsymbol{D}|r(\boldsymbol{x}) = \boldsymbol{r}^*) = E(\boldsymbol{\pi}|r(\boldsymbol{x}) = \boldsymbol{r}^*) = \boldsymbol{r}^*$ *for all* $\boldsymbol{x} \in \Omega$ *and* $\boldsymbol{r}^* \in [0, 1]^m$.

Calibration is usually assessed by plots or goodness-of-fit tests.[29–31] While these approaches could generalise to a multivariate setting, the following component-wise definition is sufficient for application to marginal prediction models, and can be assessed by applying univariate methods to each component of $r(\boldsymbol{x})$.

Definition 2: *The risk prediction model* $r(\boldsymbol{x})$ *is strongly component-wise calibrated if* $E(D_{[j]}|\boldsymbol{x}) = E(\pi_{[j]}|\boldsymbol{x}) = r_{[j]}(\boldsymbol{x})$ *for all* $j = 1, \cdots, m$ *and* $\boldsymbol{x} \in \Omega$. *The prediction model is weakly component-wise calibrated if* $E\Big(D_{[j]}|r_{[j]}(\boldsymbol{x}) = r^*\Big) = E(\pi_{[j]}|r_{[j]}(\boldsymbol{x}) = r^*) = r^*$ *for all* $j = 1, \cdots, m$, $\boldsymbol{x} \in \Omega$ *and* $r^* \in [0, 1]$.

Calibration implies component-wise calibration, but the converse need not hold. In the rest of the paper I assume that $r(\boldsymbol{x})$ is at least weakly component-wise calibrated.

Let $\boldsymbol{t} \in [0, 1]^m$ be a vector of risk thresholds. Each individual i is assigned to a high-risk category for each outcome j where $r_{[j]}(\boldsymbol{X}_i) \geq t_{[j]}$.

2.2 Outcome-wise criteria

A straightforward approach is to treat outcomes, rather than individuals, as the sampling units and then apply standard criteria to the vectorised outcomes. Such a view might be appropriate when the consequences of predicting or developing the outcomes are independent. This approach has been used in evaluating carrier screening panels for Mendelian disorders.[32] Another example might be in molecular screening for allergies.[33]

Definition 3: *Outcome-wise sensitivity is the probability of a positive prediction for an outcome that did occur. Over the joint sample space of* $\boldsymbol{D}$ *and* $\boldsymbol{X}$

$$sens_O(\boldsymbol{t}) = \frac{E_{D,X}\Big(\boldsymbol{D}'\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t})\Big)}{E_D(\boldsymbol{D}'\boldsymbol{1})}$$

where $\boldsymbol{I}$ is a vector of component-wise indicators and $\mathbf{1}$ is the vector with all elements equal to one. This is equivalent to the classical sensitivity when $m = 1$. However, whereas the classical sensitivity does not depend on the outcome probability $E(\pi)$, the outcome-wise sensitivity does depend on the relative outcome probabilities. To see this, write

$$\begin{aligned} sens_O(\boldsymbol{t}) &= \frac{E_{\boldsymbol{D},\boldsymbol{X}}\left(\sum_j D_{[j]} I(r_{[j]}(\boldsymbol{X}) \geq t_{[j]})\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})} = \sum_{j=1}^{m} \frac{E_{\boldsymbol{D},\boldsymbol{X}}\left(D_{[j]} I(r_{[j]}(\boldsymbol{X}) \geq t_{[j]})\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})} \\ &= \sum_{j=1}^{m} \frac{E_{\boldsymbol{D},\boldsymbol{X}}\left(D_{[j]} I(r_{[j]}(\boldsymbol{X}) \geq t_{[j]})\right)}{E_{\boldsymbol{D}}(D_{[j]})} \frac{E_{\boldsymbol{D}}(D_{[j]})}{E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})} = \sum_{j=1}^{m} \Pr(r_{[j]}(\boldsymbol{X}) \geq t_{[j]} | D_{[j]} = 1) \frac{\Pr(D_{[j]} = 1)}{\sum_k \Pr(D_{[k]} = 1)} \end{aligned}$$

The first term in the summand is the classical sensitivity for outcome j, so the outcome-wise sensitivity is the weighted sum of the individual outcome sensitivities, with the weights as the relative outcome probabilities. Therefore, a sample estimate of outcome-wise sensitivity may be subject to ascertainment bias, but a population estimate may be obtained by weighting the individual outcome sensitivities using external estimates of outcome probabilities.

Weights may be used to attach greater importance to the prediction of some outcomes. This may be done by generalising the outcome-wise sensitivity to

$$sens_O(\boldsymbol{t}) = \frac{E_{\boldsymbol{D},\boldsymbol{X}}\left(\boldsymbol{D}'\boldsymbol{W}\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t})\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\boldsymbol{W}\mathbf{1})}$$

where $\boldsymbol{W}$ is a diagonal matrix with positive entries. Again this is equivalent to a weighted sum of individual outcome sensitivities, with the weights as the relative outcome probabilities scaled by the respective diagonal elements of $\boldsymbol{W}$.

Definition 4: *Outcome-wise specificity is the probability of a negative prediction for an outcome that did not occur.*

$$spec_O(\boldsymbol{t}) = \frac{E_{\boldsymbol{D},\boldsymbol{X}}\left((\mathbf{1} - \boldsymbol{D})'\boldsymbol{I}(r(\boldsymbol{X}) < \boldsymbol{t})\right)}{E_{\boldsymbol{D}}\left((\mathbf{1} - \boldsymbol{D})'\mathbf{1}\right)}$$

Similarly to the sensitivity, the outcome-wise specificity is the weighted sum of the individual outcome specificities, with the weights as the relative probabilities of the complementary outcomes. General weights may be introduced as for the sensitivity.

A standard, if often criticised[34–36] summary of sensitivity and specificity is the area under the receiver operating characteristic (ROC) curve, which for a single outcome is constructed by plotting sensitivity against 1-specificity over the range of t. The idea of a ROC does not easily generalise to multiple outcomes because vectors $\boldsymbol{t}$ need not parameterise a one-to-one mapping of specificity to sensitivity. However, the C- (concordance) index,[37] which for a single outcome is equivalent to the area under the entire ROC curve, can be extended more readily.

The C-index for a single outcome is the probability that, given one individual with the outcome and one without, the prediction is higher for the former, i.e. $\Pr(r(X_{i_1}) > r(X_{i_o}) | D_{i_1} = 1, D_{i_o} = 0)$. An outcome-wise extension might be to evaluate the same probability over outcomes rather than individuals. However, this would compare the predicted risk for an outcome that did occur to the predicted risk of a different outcome that did not occur, which is difficult to interpret when the elements of $\boldsymbol{t}$ are unequal. Stated differently, if the aim is to quantify how well $r(\boldsymbol{x})$ discriminates outcomes that did occur from those that did not, it makes little sense to compare predictions for different outcomes when the risk thresholds for those outcomes may be different.

A more satisfactory approach is to compare a prediction for an outcome that did occur to a prediction for *the same* outcome when it did not occur. This just yields the C-index for that outcome, so the expected C-index for multiple outcomes is the weighted sum of individual outcome C-indices. For outcome j the probability of observing a discordant pair of outcomes is $E(D_{[j]})(1 - E(D_{[j]}))$ giving

Definition 5: *Outcome-wise C-index is the weighted sum of individual outcome C-indices.*

$$\sum_{j=1}^{m} \Pr(r_{[j]}(\boldsymbol{X}_{i_1}) > r_{[j]}(\boldsymbol{X}_{i_0}) | D_{i_1[j]} = 1, D_{i_o[j]} = 0) \frac{E(D_{[j]})(1 - E(D_{[j]}))}{\sum_k E(D_{[k]})(1 - E(D_{[k]}))}$$

One criticism of the ROC curve is that it treats sensitivity and specificity equally when they may entail different benefits and costs. The relative utility curve has been proposed to address this issue,[38,39] and is especially useful for comparing different risk prediction models. Here I summarise its derivation for one outcome before developing an outcome-wise extension.

Let b be the benefit of correctly predicting an outcome that did occur, and c the cost of incorrectly predicting an outcome that did not occur. Given a decision making risk threshold t, for an individual i with risk prediction $r(\boldsymbol{X_i}) = t$ the net benefit of a positive prediction is $b\Pr(D_i = 1|r(\boldsymbol{X_i}) = t) - c\Pr(D_i = 0|r(\boldsymbol{X_i}) = t)$ and this is positive when

$$\frac{\Pr\Big(D_i = 1|r(X_i) = t\Big)}{\Pr\Big(D_i = 0|r(\boldsymbol{X_i}) = t\Big)} > \frac{c}{b}$$

It follows that if the risk predictor is weakly calibrated, the net benefit is positive if $r(\boldsymbol{X_i}) > t$ where t is such that

$$\frac{t}{1-t} = \frac{\Pr\Big(D_i = 1|r(\boldsymbol{X}_i) = t\Big)}{\Pr\Big(D_i = 0|r(\boldsymbol{X}_i) = t\Big)} = \frac{c}{b}$$

Therefore, use of the threshold t implies a cost-benefit ratio of $t/(1-t)$. With this threshold, the expected net benefit over the population is

$$\begin{aligned}
&\Pr\Big(r(\boldsymbol{X}) \geq t\Big)\Big[b\Pr\Big(D = 1|r(\boldsymbol{X}) \geq t\Big) - c\Pr\Big(D = 0|r(\boldsymbol{X}) \geq t\Big)\Big] \\
&= b\Big[\Pr\Big(r(\boldsymbol{X}) \geq t|D = 1\Big)\Pr(D = 1) - \frac{c}{b}\Pr\Big(r(\boldsymbol{X}) \geq t|D = 0\Big)\Pr(D = 0)\Big] \\
&= b\Pr(D = 1)\Big[sens(t) - \frac{t}{1-t}\frac{\Pr(D = 0)}{\Pr(D = 1)}(1 - spec(t))\Big]
\end{aligned}$$

The relative utility is the ratio of this expectation to its theoretical maximum when sensitivity and specificity are both 1, thus

$$RU(t) = sens(t) - \frac{t}{1-t}\frac{\Pr(D = 0)}{\Pr(D = 1)}(1 - spec(t))$$

The net benefit is understood as resulting from taking action on a prediction, and so is relative to the result of taking no action. If the default, in the absence of risk prediction, is to take no action, then that is equivalent to a risk predictor with sensitivity 0 and specificity 1 at all thresholds. Conversely, if the default were always to take action then the sensitivity is 1 and the specificity is 0. A default of no action is rational when its relative utility is greater than under the default of always taking action. The definition of $RU(t)$ shows that this occurs when $t \geq \Pr(D = 1)$, termed the *relevant region* for evaluating relative utility.[38] On the other hand, if the default is to take action, then the analogous definition for $t \leq \Pr(D = 1)$ is

$$RU(t) = spec(t) - \frac{1-t}{t}\frac{\Pr(D = 1)}{\Pr(D = 0)}(1 - sens(t))$$

These expressions assume negligible cost of evaluating $r(\boldsymbol{X})$; more general derivations are provided elsewhere.[38] Turning now to multiple outcomes, let b_O and c_O represent common values of benefit and cost for all outcomes. (In practice these quantities may vary across outcomes, so they may be thought of here as average values.) Assume that benefits and costs are additive across outcomes within individuals. For an individual i with risk prediction $r(\boldsymbol{X}_i) = \boldsymbol{t}$, the net benefit of a positive prediction is now

$$b_O E_{\boldsymbol{D}_i}\Big(\boldsymbol{D}_i'\boldsymbol{1}|r(\boldsymbol{X}_i) = \boldsymbol{t}\Big) - c_O E_{\boldsymbol{D}_i}\Big((\boldsymbol{1} - \boldsymbol{D}_i)'\boldsymbol{1}|r(\boldsymbol{X}_i) = \boldsymbol{t}\Big)$$

and is positive when

$$\frac{E_{\boldsymbol{D}_i}\left(\boldsymbol{D}_i'\mathbf{1}|r(\boldsymbol{X}_i)=\boldsymbol{t}\right)}{E_{\boldsymbol{D}_i}\left((\mathbf{1}-\boldsymbol{D}_i)'\mathbf{1}|r(\boldsymbol{X}_i)=\boldsymbol{t}\right)} > \frac{c_O}{b_O}$$

If the risk predictor is weakly component-wise calibrated, then

$$E_{\boldsymbol{D}_i}\left(\boldsymbol{D}_i'\mathbf{1}|r(\boldsymbol{X}_i)=\boldsymbol{t}\right) = \boldsymbol{t}'\mathbf{1}$$

Therefore, the use of threshold vector $\boldsymbol{t}$ implies the cost–benefit ratio

$$\frac{\boldsymbol{t}'\mathbf{1}}{(\mathbf{1}-\boldsymbol{t})'\mathbf{1}} = \frac{c_O}{b_O}$$

Under additive benefits and costs, the expected net benefit over the population is

$$\begin{aligned}
&\sum_{j=1}^{m}\Pr\left(r_{[j]}(\boldsymbol{X}) \geq t_{[j]}\right)\left[b_O E\left(D_{[j]}|r_{[j]}(\boldsymbol{X}) \geq t_{[j]}\right) - c_O E\left(1 - D_{[j]}|r_{[j]}(\boldsymbol{X}) \geq t_{[j]}\right)\right] \\
&\quad = b_O\left[E_{\boldsymbol{D},\boldsymbol{X}}\left(\boldsymbol{D}'\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) - \frac{c_O}{b_O}E_{\boldsymbol{D},\boldsymbol{X}}\left((\mathbf{1}-\boldsymbol{D})'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right)\right)\right] \\
&\quad = b_O E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})\left[sens_O(\boldsymbol{t}) - \frac{\boldsymbol{t}'\mathbf{1}}{(\mathbf{1}-\boldsymbol{t})'\mathbf{1}}\frac{E_{\boldsymbol{D}}\left((\mathbf{1}-\boldsymbol{D})'\mathbf{1}\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})}(1 - spec_O(\boldsymbol{t}))\right]
\end{aligned}$$

Definition 6: *Outcome-wise relative utility for threshold vector* **t** *is*

$$RU_O(\boldsymbol{t}) = sens_O(\boldsymbol{t}) - \frac{\boldsymbol{t}'\mathbf{1}}{(\mathbf{1}-\boldsymbol{t})'\mathbf{1}}\frac{E_{\boldsymbol{D}}\left((\mathbf{1}-\boldsymbol{D})'\mathbf{1}\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})}(1 - spec_O(\boldsymbol{t}))$$

As before, a diagonal weight matrix $\boldsymbol{W}$ can be used to allow some outcomes to contribute more to the relative utility, giving

$$RU_O(\boldsymbol{t}) = sens_O(\boldsymbol{t}) - \frac{\boldsymbol{t}'\boldsymbol{W}\mathbf{1}}{(\mathbf{1}-\boldsymbol{t})'\boldsymbol{W}\mathbf{1}}\frac{E_{\boldsymbol{D}}\left((\mathbf{1}-\boldsymbol{D})'\boldsymbol{W}\mathbf{1}\right)}{E_{\boldsymbol{D}}(\boldsymbol{D}'\boldsymbol{W}\mathbf{1})}(1 - spec_O(\boldsymbol{t}))$$

where $sens_O(\boldsymbol{t})$ and $spec_O(\boldsymbol{t})$ are also used in their weighted versions. For multiple outcomes, relative utility defines a surface over the space of threshold vectors, $RU : [0,1]^m \mapsto (-\infty, 1]$. The relevant region is $\{\boldsymbol{t} : \boldsymbol{t}'\mathbf{1} \geq E_{\boldsymbol{D}}(\boldsymbol{D}'\mathbf{1})\}$ when the default, in the absence of risk prediction, is to take no action for any outcome. When there are outcomes for which the default is to take action, a pragmatic approach is to substitute the complementary outcomes, and thresholds, in the above definitions.

2.3 Joint criteria

An issue with outcome-wise measures is that actions are applied to individuals rather than outcomes. In many contexts, it is more appropriate to summarise risk predictions for each individual before taking action. To this end I now define individual-wise measures, which vary according to the definition of a true positive prediction. For joint measures, the aim is to predict the joint occurrence of all outcomes in an individual. An example might be in forensic identification from an anonymous DNA sample, where a profile could be constructed from several traits such as hair colour,[40] height[16] and weight,[41] each discretised into broad categories.

Definition 7: *Joint sensitivity is the probability of predicting all outcomes to occur, in an individual for which all outcomes did occur.*

$$sens_J(\boldsymbol{t}) = \Pr(r(\boldsymbol{X}) \geq \boldsymbol{t} | \boldsymbol{D} = \boldsymbol{1})$$

If the elements of $r(\boldsymbol{X})$ are jointly independent and the elements of $\boldsymbol{D}$ also are jointly independent, then

$$\begin{aligned} sens_J(\boldsymbol{t}) &= \frac{\prod_j \Pr(r_{[j]}(\boldsymbol{X}) \geq t_{[j]}, D_{[j]} = 1)}{\prod_j \Pr(D_{[j]} = 1)} \\ &= \prod_j \frac{\Pr(r_{[j]}(\boldsymbol{X}) \geq t_{[j]}, D_{[j]} = 1)}{\Pr(D_{[j]} = 1)} \\ &= \prod_j \Pr(r_{[j]}(\boldsymbol{X}) \geq t_{[j]} | D_{[j]} = 1) \end{aligned}$$

In this case, the joint sensitivity is the product of individual outcome sensitivities. However, in the general case of dependence between elements of $r(\boldsymbol{X})$ or $\boldsymbol{D}$, the joint sensitivity is not readily expressed in terms of the individual outcome sensitivities.

Definition 8: *Joint specificity is the probability of predicting at least one outcome not to occur, in an individual for which at least one outcome did not occur.*

$$spec_J(\boldsymbol{t}) = \Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) \neq \boldsymbol{1} | \boldsymbol{D} \neq \boldsymbol{1})$$

Note that this may depend on the distribution of $\boldsymbol{D}$ and therefore that an estimate of $spec_J(\boldsymbol{t})$ may be subject to ascertainment bias. When information is available on the distribution of $\boldsymbol{D}$, an unbiased estimate of $spec_J(\boldsymbol{t})$ could be obtained by weighting each observation in which $\boldsymbol{D} \neq \boldsymbol{1}$ by the inverse of its sampling probability.

To define joint concordance, note that $\boldsymbol{D} = \boldsymbol{1} \iff \min(\boldsymbol{D}) = 1$ and $\boldsymbol{D} \neq \boldsymbol{1} \iff \min(\boldsymbol{D}) = 0$.

Definition 9: *Joint C-index is the probability that, given one individual in which all outcomes did occur and one individual in which at least one outcome did not occur, the minimum risk prediction is higher in the former individual.*

$$C_J = \Pr(\min(r(\boldsymbol{X}_{i_1})) > \min(r(\boldsymbol{X}_{i_0})) | \boldsymbol{D}_{i_1} = \boldsymbol{1}, \boldsymbol{D}_{i_0} \neq \boldsymbol{1})$$

To define relative utility, let b_J be the benefit of predicting all outcomes to occur when all outcomes did occur, and c_J the cost of predicting all outcomes to occur when at least one outcome did not occur. For an individual i with risk prediction $r(\boldsymbol{X}_i) = \boldsymbol{t}$, the net benefit of a positive prediction is $b_J \Pr\left(\boldsymbol{D}_i = \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t}\right) - c_J \Pr(\boldsymbol{D}_i \neq \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t})$ and this is positive when

$$\frac{\Pr\left(\boldsymbol{D}_i = \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D}_i \neq \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t}\right)} > \frac{c_J}{b_J}$$

Therefore, use of the threshold vector $\boldsymbol{t}$ implies a cost–benefit ratio of $\frac{\Pr\left(\boldsymbol{D}_i = \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D}_i \neq \boldsymbol{1} | r(\boldsymbol{X}_i) = \boldsymbol{t}\right)}$. With this threshold, the expected net benefit in the population is

$$\begin{aligned} &\Pr\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right)\left[b_J \Pr\left(\boldsymbol{D} = \boldsymbol{1} | r(\boldsymbol{X}) \geq \boldsymbol{t}\right) - c_J \Pr\left(\boldsymbol{D} \neq \boldsymbol{1} | r(\boldsymbol{X}) \geq \boldsymbol{t}\right)\right] \\ &= b_J\left[\Pr\left(r(\boldsymbol{X}) \geq \boldsymbol{t} | \boldsymbol{D} = \boldsymbol{1}\right)\Pr(\boldsymbol{D} = \boldsymbol{1}) - \frac{c_J}{b_J}\Pr\left(r(\boldsymbol{X}) \geq \boldsymbol{t} | \boldsymbol{D} \neq \boldsymbol{1}\right)\Pr(\boldsymbol{D} \neq \boldsymbol{1})\right] \\ &= b_J \Pr(\boldsymbol{D} = \boldsymbol{1})\left[sens_J(\boldsymbol{t}) - \frac{\Pr\left(\boldsymbol{D} = \boldsymbol{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} \neq \boldsymbol{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \frac{\Pr(\boldsymbol{D} \neq \boldsymbol{1})}{\Pr(\boldsymbol{D} = \boldsymbol{1})}(1 - spec_J(\boldsymbol{t}))\right] \end{aligned}$$

Definition 10: *Joint relative utility for threshold* $\boldsymbol{t}$ *is*

$$RU_J(\boldsymbol{t}) = sens_J(\boldsymbol{t}) - \frac{\Pr\left(\boldsymbol{D} = \mathbf{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} \neq \mathbf{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \frac{\Pr(\boldsymbol{D} \neq \mathbf{1})}{\Pr(\boldsymbol{D} = \mathbf{1})} (1 - spec_J(\boldsymbol{t}))$$

In general $\Pr(\boldsymbol{D} = \mathbf{1} | r(\boldsymbol{X}) = \boldsymbol{t})$ must be estimated. As this may be difficult in practice, the following working definition may be useful. If risk predictions and outcomes both are jointly independent, and the risk predictor is weakly component-wise calibrated, then $\Pr(\boldsymbol{D} = \mathbf{1} | r(\boldsymbol{X}) = \boldsymbol{t}) = \prod_j t_{[j]}$ and

$$RU_J(\boldsymbol{t}) = sens_J(\boldsymbol{t}) - \frac{\prod_j t_{[j]}}{1 - \prod_j t_{[j]}} \frac{\Pr(\boldsymbol{D} \neq \mathbf{1})}{\Pr(\boldsymbol{D} = \mathbf{1})} (1 - spec_J(\boldsymbol{t}))$$

The relevant region is $\{\boldsymbol{t} : \Pr\left(\boldsymbol{D} = \mathbf{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right) \geq \Pr(\boldsymbol{D} = \mathbf{1})\}$ when the default, in the absence of risk prediction, is to take no action for at least one outcome.

2.4 Panel-wise criteria

For panel-wise criteria the aim is to predict the occurrence of at least one outcome in an individual. A correct prediction may, however, be defined in different ways according to whether the predicted outcomes are the same as those that did occur. Here I propose two senses of panel-wise prediction, called the weak and strong senses by analogy to family-wise errors in hypothesis testing.

Definition 11: *Weak panel-wise sensitivity is the probability of predicting at least one outcome to occur, in an individual for which at least one outcome did occur.*

$$sens_S(\boldsymbol{t}) = \Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) \neq \mathbf{0} | \boldsymbol{D} \neq \mathbf{0})$$

The subscript S stands for *screening* as explained later. Note that this may depend on the distribution of $\boldsymbol{D}$ and therefore that an estimate of $sens_S(\boldsymbol{t})$ may be subject to ascertainment bias. When information is available on the distribution of $\boldsymbol{D}$, an unbiased estimate of $sens_S(\boldsymbol{t})$ could be obtained by weighting each observation in which $\boldsymbol{D} \neq \mathbf{0}$ by the inverse of its sampling probability.

Definition 12: *Weak panel-wise specificity is the probability of predicting no outcomes to occur, in an individual for which no outcomes did occur.*

$$spec_S(\boldsymbol{t}) = \Pr(r(\boldsymbol{X}) < \boldsymbol{t} | \boldsymbol{D} = \mathbf{0})$$

Definitions 11 and 12 are complementary to the joint sensitivity and specificity, and similarly the weak panel-wise specificity is the product of the component-wise specificities in the case that risk predictions and outcomes both are jointly independent. The complement of weak panel-wise specificity is analogous to the weak sense of family-wise type-1 error rate in hypothesis testing. Similar arguments to the joint criteria give the following definitions of concordance and relative utility.

Definition 13: *Weak panel-wise C-index is the probability that, given one individual in which at least one outcome did occur and one individual in which no outcomes did occur, the maximum risk prediction is higher in the former individual.*

$$C_S = \Pr\left(\max\left(r(\boldsymbol{X}_{i_1})\right) > \max(r(\boldsymbol{X}_{i_0})) | \boldsymbol{D}_{i_1} \neq \mathbf{0}, \boldsymbol{D}_{i_0} = \mathbf{0}\right)$$

Definition 14: *Weak panel-wise relative utility for threshold vector* $\boldsymbol{t}$ *is*

$$RU_S(\boldsymbol{t}) = sens_S(\boldsymbol{t}) - \frac{\Pr\left(\boldsymbol{D} \neq \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} = \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \frac{\Pr(\boldsymbol{D} = \boldsymbol{0})}{\Pr(\boldsymbol{D} \neq \boldsymbol{0})} (1 - spec_S(\boldsymbol{t}))$$

If risk predictions and outcomes both are jointly independent, and the risk predictor is weakly component-wise calibrated, then

$$RU_S(\boldsymbol{t}) = sens_S(\boldsymbol{t}) - \frac{1 - \prod_j (1 - t_{[j]})}{\prod_j (1 - t_{[j]})} \frac{\Pr(\boldsymbol{D} = \boldsymbol{0})}{\Pr(\boldsymbol{D} \neq \boldsymbol{0})} (1 - spec_S(\boldsymbol{t}))$$

The relevant region is $\{\boldsymbol{t} : \Pr\left(\boldsymbol{D} \neq \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right) \geq \Pr(\boldsymbol{D} \neq \boldsymbol{0})\}$ when the default, in the absence of risk prediction, is to take no action for any outcome.

Turning to the strong sense definitions, the key difference is that the predicted and actual outcomes must coincide for at least one outcome that did occur.

Definition 15: *Strong panel-wise sensitivity is the probability that at least one outcome is correctly predicted to occur in an individual for which at least one outcome did occur.*

$$sens_P(\boldsymbol{t}) = \Pr(\boldsymbol{D}'\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) \neq \boldsymbol{0} | \boldsymbol{D} \neq \boldsymbol{0})$$

Estimates of $sens_P(\boldsymbol{t})$ may be subject to ascertainment bias, which could be adjusted for by weighting each observation where $\boldsymbol{D} \neq \boldsymbol{0}$ by the inverse of its sampling probability.

Definition 16: *Strong panel-wise specificity is the probability that all outcomes that did not occur are predicted not to occur in an individual for which at least one outcome did not occur.*

$$spec_P(\boldsymbol{t}) = \Pr\left((\boldsymbol{1} - \boldsymbol{D})'\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) = \boldsymbol{0} | \boldsymbol{D} \neq \boldsymbol{1}\right)$$

Definitions 15 and 16 complement each other in a different way to the weak sense definitions 15 and 16. The complement of strong panel-wise specificity is analogous to the strong sense of family-wise type-1 error in hypothesis testing. Note that an individual may count towards both sensitivity and specificity, a property shared with the outcome-wise measures.

Definition 17: *Strong panel-wise C-index is the probability that, given one individual in which at least one outcome did occur and one individual in which at least one outcome did not occur, the maximum risk prediction is greater among the outcomes that did occur in the former individual than among the outcomes that did not occur in the latter.*

$$C_P = \Pr(\max\left(\boldsymbol{D}_{i_1} \circ r(\boldsymbol{X}_{i_1})\right) > \max\left((\boldsymbol{1} - \boldsymbol{D}_{i_0}) \circ r(\boldsymbol{X}_{i_0})\right) | \boldsymbol{D}_{i_1} \neq \boldsymbol{0}, \boldsymbol{D}_{i_0} \neq \boldsymbol{1})$$

where $\circ$ denotes Hadamard product.

Note that under this definition an individual may appear on both sides of the inequality (i.e. $i_1 = i_0$) and, unlike C_J and C_S, C_P does not have a natural interpretation as a measure of discrimination. Furthermore, it need not equal 0.5 under random predictions. Nevertheless it corresponds to definitions of sensitivity and specificity in the same way as those other measures of concordance, and could be used as a summary measure for comparing different predictors of a set of outcomes.

Relative utility cannot be developed in the same manner as RU_J and RU_S, but the following working definition is analogous to that of the weak panel-wise sense.

Definition 18: *Strong panel-wise relative utility for threshold vector* $\boldsymbol{t}$ *is*

$$RU_P(\boldsymbol{t}) = sens_P(\boldsymbol{t}) - \frac{\Pr\left(\boldsymbol{D} \neq \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} \neq \boldsymbol{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \frac{\Pr(\boldsymbol{D} \neq \boldsymbol{1})}{\Pr(\boldsymbol{D} \neq \boldsymbol{0})} (1 - spec_P(\boldsymbol{t}))$$

with the relevant region

$$\left\{ \boldsymbol{t} : \frac{\Pr\left(\boldsymbol{D} \neq \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} \neq \boldsymbol{1} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \geq \frac{\Pr(\boldsymbol{D} \neq \boldsymbol{0})}{\Pr(\boldsymbol{D} \neq \boldsymbol{1})} \right\}$$

If risk predictions and outcomes both are jointly independent, and the risk predictor is weakly component-wise calibrated, then

$$RU_P(\boldsymbol{t}) = sens_P(\boldsymbol{t}) - \frac{1 - \prod_j (1 - t_{[j]})}{1 - \prod_j t_{[j]}} \frac{\Pr(\boldsymbol{D} \neq \boldsymbol{1})}{\Pr(\boldsymbol{D} \neq \boldsymbol{0})} (1 - spec_P(\boldsymbol{t}))$$

Which of the weak or strong measures is more appropriate will depend on the application. For example, if the same action would be performed for all outcomes, it is less important to predict specific outcomes. That might be the case when screening for a range of conditions with a common intervention, as is done say when measuring blood pressure with a view to prescribing anti-hypertensives. For this reason I suggest *screening*, with subscript S, as a shorthand for weak panel-wise, and *panel-wise* itself, subscript P, as a shorthand for strong panel-wise, and will use those terms in the rest of the paper. (Strong) panel-wise measures may be appropriate in early detection settings where the action depends on the specific outcomes predicted.

Figure 1 shows an example of four outcomes in eight individuals, showing which individuals count towards the different senses of sensitivity.

3 Multivariate probit model

For a single outcome, many measures of predictive accuracy can be expressed in terms of variance explained by the risk predictor, assuming a probit model for the outcome.[42] This allows any of the measures to be derived from reported values of any others, and argues for the use of variance explained as a fundamental measure of prediction accuracy without the caveats associated with, for example, ROC curves. Here this framework is extended to the prediction of multiple traits using a multivariate probit model for outcomes.[43]

Assume that individual i has a latent *liability* vector $\boldsymbol{L}_i$ distributed as multivariate normal with dimension m, mean vector $\boldsymbol{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}_L$ with diagonal entries 1. Define the threshold vector $\boldsymbol{\tau}$ such that outcome j occurs whenever $L_{[j]} \geq \tau_{[j]}$, thus $\tau_{[j]} = \Phi^{-1}(1 - \Pr(D_{[j]} = 1))$.

Assume that each outcome has a single normally distributed predictor, so that the predictor vector $\boldsymbol{X}_i$ is multivariate normal with dimension m, mean vector $\boldsymbol{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}_X$. Let the joint distribution of liability and predictor be multivariate normal with mean vector $\boldsymbol{0}$ and variance–covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_L & \boldsymbol{\Sigma}_{LX} \\ \boldsymbol{\Sigma}'_{LX} & \boldsymbol{\Sigma}_X \end{bmatrix}$$

where component $[jk]$ of $\boldsymbol{\Sigma}_{LX}$ is the covariance between liability for outcome j and predictor of outcome k. A notable special case is $\boldsymbol{\Sigma}_{LX} = \boldsymbol{\Sigma}_X$. Then the diagonal elements of $\boldsymbol{\Sigma}_X$ are the variances in each liability explained by

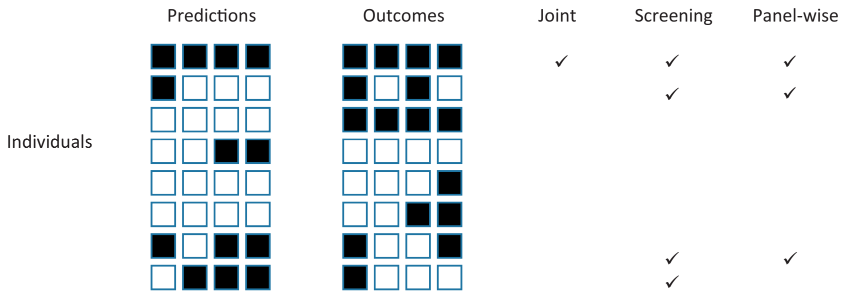

Figure 1. Example outcomes in eight individuals. Outcomes predicted to occur are shown in black on the left panel. Outcomes that did occur are shown in black on the right panel. Ticks show individuals counting in the numerator for each sense of sensitivity. Here the sample joint sensitivity is 1/2, the screening sensitivity is 4/7, and the panel-wise sensitivity is 3/7. The outcome-wise sensitivity is 7/16.

the corresponding predictors, and for each outcome, conditional on its own predictor there is no additional information from any other predictors.

The following expressions will be useful. If each element of $\boldsymbol{X}$ estimates the corresponding element of $\boldsymbol{L}$, the risk prediction for outcome j is given by

$$r_{[j]}(\boldsymbol{X}) = 1 - \Phi\left(\frac{\tau_{[j]} - X_{[j]}}{\sqrt{1 - \Sigma_{X[jj]}}}\right) \tag{1}$$

and the risk threshold $t_{[j]}$ is equivalent to the predictor threshold

$$\tilde{t}_{[j]} = \tau_{[j]} - (1 - \Sigma_{X[jj]})^{\frac{1}{2}}\Phi^{-1}(1 - t_{[j]})$$

Given outcomes $\boldsymbol{D} = \boldsymbol{d}$, the liability follows a multivariate truncated normal distribution, with truncation at $\boldsymbol{\tau}$ from below for the outcomes that did occur and from above for those that did not. Denote the conditional mean vector and variance–covariance matrix of the truncated liability by $\boldsymbol{\mu}_{L|\boldsymbol{D}=\boldsymbol{d}}$ and $\boldsymbol{\Sigma}_{L|\boldsymbol{D}=\boldsymbol{d}}$; these quantities may be computed numerically by the method of Tallis.[44,45] The Pearson-Aitken selection formulae[46] give the mean predictor in individuals with outcomes $\boldsymbol{d}$ as

$$E(\boldsymbol{X}|\boldsymbol{D} = \boldsymbol{d}) = \boldsymbol{\Sigma}_{LX}\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\mu}_{L|\boldsymbol{D}=\boldsymbol{d}} \tag{2}$$

and the variance–covariance matrix

$$var(\boldsymbol{X}|\boldsymbol{D} = \boldsymbol{d}) = \boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_{LX}(\boldsymbol{\Sigma}_L^{-1} - \boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Sigma}_{L|\boldsymbol{D}=\boldsymbol{d}}\boldsymbol{\Sigma}_L^{-1})\boldsymbol{\Sigma}'_{LX} \tag{3}$$

Assume that conditional on $\boldsymbol{d}$ the predictor follows the m-variate normal distribution with the above mean and variance–covariance. Furthermore since $\boldsymbol{L}$ has mean $\boldsymbol{0}$

$$\boldsymbol{\mu}_{L|\boldsymbol{D}\neq\boldsymbol{d}} = -\frac{\Pr(\boldsymbol{D} = \boldsymbol{d})}{\Pr(\boldsymbol{D} \neq \boldsymbol{d})}\boldsymbol{\mu}_{L|\boldsymbol{D}=\boldsymbol{d}}$$

and

$$\boldsymbol{\Sigma}_{L|\boldsymbol{D}\neq\boldsymbol{d}} = \frac{\boldsymbol{\Sigma}_L - \Pr(\boldsymbol{D} = \boldsymbol{d})\boldsymbol{\Sigma}_{L|\boldsymbol{D}=\boldsymbol{d}}}{1 - \Pr(\boldsymbol{D} = \boldsymbol{d})} - \left[\frac{\Pr(\boldsymbol{D} = \boldsymbol{d})}{1 - \Pr(\boldsymbol{D} = \boldsymbol{d})}\right]^2 \boldsymbol{\mu}'_{L|\boldsymbol{D}=\boldsymbol{d}}\boldsymbol{\mu}_{L|\boldsymbol{D}=\boldsymbol{d}}$$

from which $E(\boldsymbol{X}|\boldsymbol{D} \neq \boldsymbol{d})$ and $var(\boldsymbol{X}|\boldsymbol{D} \neq \boldsymbol{d})$ follow analogously to equations (2) and (3).

Finally assume that conditional on a prediction $\boldsymbol{X} = \boldsymbol{x}$ the liability follows the m-variate normal distribution with the mean and variance–covariance given by the Pearson-Aitken selection formulae as

$$E(\boldsymbol{L}|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{\Sigma}_{LX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x} \tag{4}$$

$$var(\boldsymbol{L}|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{\Sigma}_L - \boldsymbol{\Sigma}_{LX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}'_{LX} \tag{5}$$

The outcome-wise criteria can be expressed in terms of single outcome criteria, which are special cases of the joint criteria below and are therefore omitted for brevity.

3.1 Joint criteria

From Definition 7

$$\begin{aligned} sens_J(\boldsymbol{t}) &= \Pr(r(\boldsymbol{X}) \geq \boldsymbol{t}|\boldsymbol{D} = \boldsymbol{1}) \\ &= \frac{\Pr(r(\boldsymbol{X}) \geq \boldsymbol{t}, \boldsymbol{D} = \boldsymbol{1})}{\Pr(\boldsymbol{D} = \boldsymbol{1})} \\ &= \frac{\boldsymbol{\Phi}((-\boldsymbol{\tau}, -\tilde{\boldsymbol{t}})'; \boldsymbol{0}, \boldsymbol{\Sigma})}{\boldsymbol{\Phi}(-\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)} \end{aligned}$$

where $\boldsymbol{\Phi}(\cdot, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal cumulative distribution function with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$.

From Definition 8

$$\begin{aligned} spec_J(\boldsymbol{t}) &= \Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) \neq \boldsymbol{1}|\boldsymbol{D} \neq \boldsymbol{1}) \\ &= 1 - \frac{\Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) = \boldsymbol{1}, \boldsymbol{D} \neq \boldsymbol{1})}{\Pr(\boldsymbol{D} \neq \boldsymbol{1})} \\ &= 1 - \frac{\Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) = \boldsymbol{1}) - \Pr(\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) = \boldsymbol{1}, \boldsymbol{D} = \boldsymbol{1})}{1 - \Pr(\boldsymbol{D} = \boldsymbol{1})} \\ &= 1 - \frac{\boldsymbol{\Phi}(-\tilde{\boldsymbol{t}}; \boldsymbol{0}, \boldsymbol{\Sigma}_X) - \boldsymbol{\Phi}((-\boldsymbol{\tau}, -\tilde{\boldsymbol{t}})'; \boldsymbol{0}, \boldsymbol{\Sigma})}{1 - \boldsymbol{\Phi}(-\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)} \end{aligned}$$

Calculating joint concordance requires the distribution of the maximum element of the multivariate risk predictor. This has recently been derived analytically[47] but can be approximated by simulation. First simulate a predictor from the multivariate normal distribution conditional on $\boldsymbol{D} = \boldsymbol{1}$, given by equations (2) and (3), and convert each component to a risk using equation (1). Simulate a second predictor in the same way but conditional on $\boldsymbol{D} \neq \boldsymbol{1}$. Over a large number of simulations, the joint concordance is estimated as the proportion in which the minimum risk of the first predictor exceeds the minimum in the second.

From Definition 10, the joint relative utility is

$$RU_J(\boldsymbol{t}) = sens_J(\boldsymbol{t}) - \frac{\Pr\big(\boldsymbol{D} = \boldsymbol{1}|r(\boldsymbol{X}) = \boldsymbol{t}\big)}{\Pr\big(\boldsymbol{D} \neq \boldsymbol{1}|r(\boldsymbol{X}) = \boldsymbol{t}\big)} \frac{\Pr(\boldsymbol{D} \neq \boldsymbol{1})}{\Pr(\boldsymbol{D} = \boldsymbol{1})} (1 - spec_J(\boldsymbol{t}))$$

with $\Pr(\boldsymbol{D} = \boldsymbol{1}) = \boldsymbol{\Phi}(-\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)$, and $\Pr\big(\boldsymbol{D} = \boldsymbol{1}|r(\boldsymbol{X}) = \boldsymbol{t}\big) = \boldsymbol{\Phi}(-\boldsymbol{\tau}; -\boldsymbol{\mu}_{L|X=t}, \boldsymbol{\Sigma}_{L|X=t})$ where $\boldsymbol{\mu}_{L|\boldsymbol{X}=\boldsymbol{t}}$ and $\boldsymbol{\Sigma}_{L|\boldsymbol{X}=\boldsymbol{t}}$ are given by equations (4) and (5), respectively.

3.2 Screening criteria

Following analogous steps to the joint measures, from Definition 11

$$sens_S(\boldsymbol{t}) = 1 - \frac{\boldsymbol{\Phi}(\tilde{\boldsymbol{t}}; \boldsymbol{0}, \boldsymbol{\Sigma}_X) - \boldsymbol{\Phi}((\boldsymbol{\tau}, \tilde{\boldsymbol{t}})'; \boldsymbol{0}, \boldsymbol{\Sigma})}{1 - \boldsymbol{\Phi}(\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)}$$

From Definition 12

$$spec_S(\boldsymbol{t}) = \frac{\boldsymbol{\Phi}((\boldsymbol{\tau}, \tilde{\boldsymbol{t}})'; \boldsymbol{0}, \boldsymbol{\Sigma})}{\boldsymbol{\Phi}(\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)}$$

To estimate screening concordance, first simulate a predictor from the multivariate normal distribution conditional on $\boldsymbol{D} = \boldsymbol{0}$, given by equations (2) and (3), and convert each component to a risk using equation (1). Simulate a second predictor in the same way but conditional on $\boldsymbol{D} \neq \boldsymbol{0}$. Over a large number of simulations, the screening concordance is estimated as the proportion in which the maximum risk of the second predictor exceeds the maximum in the first.

From definition 14, the screening relative utility is

$$RU_S(\boldsymbol{t}) = sens_S(\boldsymbol{t}) - \frac{\Pr\left(\boldsymbol{D} \neq \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)}{\Pr\left(\boldsymbol{D} = \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right)} \frac{\Pr(\boldsymbol{D} = \boldsymbol{0})}{\Pr(\boldsymbol{D} \neq \boldsymbol{0})} (1 - spec_S(\boldsymbol{t}))$$

with $\Pr(\boldsymbol{D} = \boldsymbol{0}) = \boldsymbol{\Phi}(\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)$, and $\Pr\left(\boldsymbol{D} = \boldsymbol{0} | r(\boldsymbol{X}) = \boldsymbol{t}\right) = \boldsymbol{\Phi}(\boldsymbol{\tau}; \boldsymbol{\mu}_{L|X=t}, \boldsymbol{\Sigma}_{L|X=t})$.

3.3 Panel-wise criteria

Panel-wise measures can be evaluated by summing over outcome vectors $\boldsymbol{d}$. From Definition 15 the panel-wise sensitivity is

$$\begin{aligned} sens_P(\boldsymbol{t}) &= \Pr(\boldsymbol{D}'\boldsymbol{I}(r(\boldsymbol{X}) \geq \boldsymbol{t}) \neq \boldsymbol{0} | \boldsymbol{D} \neq \boldsymbol{0}) \\ &= 1 - \Pr\left(\boldsymbol{D}'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) = \boldsymbol{0} | \boldsymbol{D} \neq \boldsymbol{0}\right) \\ &= 1 - \sum_{\boldsymbol{d}:\boldsymbol{d} \neq \boldsymbol{0}} \Pr\left(\boldsymbol{D}'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) = \boldsymbol{0} | \boldsymbol{D} = \boldsymbol{d}\right) \Pr(\boldsymbol{D} = \boldsymbol{d} | \boldsymbol{D} \neq \boldsymbol{0}) \\ &= 1 - \frac{1}{1 - \boldsymbol{\Phi}(\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)} \sum_{\boldsymbol{d}:\boldsymbol{d} \neq \boldsymbol{0}} \Pr\left(\boldsymbol{D}'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) = \boldsymbol{0}, \boldsymbol{D} = \boldsymbol{d}\right) \end{aligned}$$

The probability in the summand is an integral of the multivariate normal density with mean vector $\boldsymbol{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. For components j where $d_{[j]} = 1$, the limits of integration are $[\tau_{[j]}, \infty]$ for the liability components and $[-\infty, \tilde{t}_{[j]})$ for the predictor components. For components j where $d_{[j]} = 0$, the limits are $[-\infty, \tau_{[j]})$ and $[-\infty, \infty]$, respectively.

From definition 16 the panel-wise specificity is

$$\begin{aligned} spec_P(\boldsymbol{t}) &= \Pr\left((\boldsymbol{1} - \boldsymbol{D})'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) = \boldsymbol{0} | \boldsymbol{D} \neq \boldsymbol{1}\right) \\ &= \frac{1}{1 - \boldsymbol{\Phi}(-\boldsymbol{\tau}; \boldsymbol{0}, \boldsymbol{\Sigma}_L)} \sum_{\boldsymbol{d}:\boldsymbol{d} \neq \boldsymbol{1}} \Pr\left((\boldsymbol{1} - \boldsymbol{D})'\boldsymbol{I}\left(r(\boldsymbol{X}) \geq \boldsymbol{t}\right) = \boldsymbol{0}, \boldsymbol{D} = \boldsymbol{d}\right) \end{aligned}$$

For components j where $d_{[j]} = 1$, the limits of integration are $[\tau_{[j]}, \infty]$ for the liability components and $[-\infty, \infty]$ for the predictor components. For components j where $d_{[j]} = 0$, the limits are $[-\infty, \tau_{[j]})$ and $[-\infty, \tilde{t}_{[j]})$, respectively.

To estimate panel-wise concordance, simulate liabilities $\boldsymbol{L}$ and predictors $\boldsymbol{X}$ from their joint multivariate normal distribution with mean vector $\boldsymbol{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Concordance is estimated according to Definition 17 using pairs of simulated $\boldsymbol{L}$ and $\boldsymbol{X}$ in which one has $\boldsymbol{D} \neq \boldsymbol{1}$ and the other has $\boldsymbol{D} \neq \boldsymbol{0}$.

The panel-wise relative utility can be calculated from Definition 18 using expressions given above.

All the criteria are now expressed in terms of the marginal outcome probabilities $\Pr(D_{[j]} = 1)$ and the joint variance–covariance matrix $\boldsymbol{\Sigma}$ of liability and predictor. A summary measure of the prediction accuracy is suggested by the multivariate analysis of variance, via Wilks' Λ

$$1 - \Lambda = 1 - \frac{\det(\boldsymbol{\Sigma}_L + \boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_{LX} - \boldsymbol{\Sigma}'_{LX})}{\det(\boldsymbol{\Sigma}_L)}$$

This is the proportion of variance of $\boldsymbol{L}$ explained by the predictor $\boldsymbol{X}$. For a single outcome, $1 - \Lambda$ equals the coefficient of determination from the regression of $\boldsymbol{L}$ on $\boldsymbol{X}$.[42]

4 Examples

4.1 CancerSEEK

CancerSEEK is a blood-based test of circulating proteins and tumour DNA mutations that are associated with the presence of cancer.[9] It has been proposed for early detection of cancers of the ovary, liver, stomach, pancreas, esophagus, colorectum, lung, or breast. A single test is applied, from which a positive result suggests the presence of one of these cancers. Given a positive test, a secondary algorithm identifies the likely site of the cancer.

CancerSEEK tests a composite outcome, and as such the standard univariate criteria correspond to screening criteria. However, the authors reported sensitivities for each cancer individually, at a risk threshold of 0.893, and reported their incidence-weighted average as 55%. This average corresponds to outcome-wise sensitivity (Definition 3), but it is also a screening sensitivity if at most one cancer is present in each subject. The screening specificity was reported as over 99%.

The in-sample screening sensitivity at this risk threshold was 62.2% and the area under the ROC curve (AUC) was 91% (Figure 2a in Cohen et al.[9]). However, as noted in Definition 11 these estimates are subject to ascertainment bias, in particular the under-sampling of breast cancers relative to other cancer cases, explaining the discrepancy between the in-sample and incidence-weighted sensitivities. I randomly resampled cases from each cancer (their Table S4) in proportion to their incidence rates (L. Danilova, personal communication). The in-sample screening sensitivity was now 55%, equal to the outcome-wise sensitivity, and the screening concordance reduced to 89%. This is the concordance that would be expected in a population screening context.

4.2 Polygenic risk scores

A polygenic risk score (PRS) is an aggregation of genetic risk, $\hat{\boldsymbol{\beta}}'\boldsymbol{G}$ where $\hat{\boldsymbol{\beta}}$ is a vector of estimated effects (e.g. log odds ratios) and $\boldsymbol{G}$ is a vector of coded genotypes (e.g. numbers of risk alleles) across many DNA sites, typically single nucleotide polymorphisms (SNPs).[48] A PRS can be computed for many diseases at once in the same individual, by forming products of different effect vectors with the fixed genotype vector.

PRS have been constructed for a number of diseases and have shown promise for risk prediction.[10] Table 1 shows six diseases for which PRS have been fitted using variants across the whole genome, as opposed to a limited number of associated SNPs. The reported AUCs were converted to liability variances explained using published formulae,[49] giving the diagonal elements of $\boldsymbol{\Sigma}_{LX}$. Assume that the correlation between pairs of estimated PRS equals the total genetic correlation of the diseases, which was obtained from the LD-Hub database[50] (Table 2) to give the off-diagonal elements of $\boldsymbol{\Sigma}_X$. This assumption is more tenable for these PRS, which include variants across the whole genome, than for PRS constructed from a limited number of associated SNPs. Assume further that the correlation between disease liabilities also equals the genetic correlation, giving $\boldsymbol{\Sigma}_L$ (Table 3). Finally assume that the PRS for disease j has no covariance with disease liability k conditional on the PRS for disease k, where $j \neq k$. Under this assumption $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_{LX}$ (Table 2).

Under the model developed in section 3, the event-wise concordance is 0.653, the screening concordance is 0.607, which is lower than all individual AUCs, and the joint concordance is 0.749. The panel-wise concordance is 0.49, compared to a value of 0.37 obtained when the correlation matrices are the same but all individual AUCs are set to 0.5.

For illustration, consider a screening application to identify, early in life, those at elevated risk of at least one of these diseases. Suppose the risk threshold vector is set equal to the prevalence, so that the predictor identifies individuals with above-average predicted risk for at least one disease. The screening sensitivity is 0.955, which is considerably higher that the individual sensitivities (Table 1). However, the screening specificity is much lower at 0.074. Similarly to multiple hypothesis testing, the prediction of multiple outcomes increases both the true-positive and false-positive rate at a given threshold vector, but the thresholds that reflect the cost–benefit ratio are different in the multiple prediction context than for the single predictions. The screening concordance of 0.607 suggests that, across all thresholds regarded equally, the sensitivity-specificity trade-off is not as good as for any disease

Table 1. Properties of fitted PRS for six common diseases.

Disease	AUC	Prevalence	Liability R^2	SNP h^2	Sensitivity	Specificity
Type-2 Diabetes	0.66	0.102	0.0856	0.196	0.630	0.599
Coronary Artery Disease	0.623	0.0461	0.0398	0.22	0.600	0.575
Crohn's Disease	0.75	0.005	0.103	0.26	0.701	0.666
Ulcerative Colitis	0.7	0.0025	0.0553	0.19	0.657	0.632
Schizophrenia	0.62	0.01	0.0254	0.235	0.595	0.576
Rheumatoid Arthritis	0.7	0.01	0.0732	0.18	0.661	0.629

Note: AUC and Prevalence, the reported values in the literature.[54–60] Liability R^2, the diagonal elements of Σ_{LX} derived from AUC and Prevalence [49]. SNP h^2, the liability variance explained by all genotyped SNPs, which is the maximum possible value of Liability R^2.[24,60–63] Sensitivity and sensitivity, their values when risk threshold equals the prevalence.

Table 2. Variance–covariance matrix Σ_X between PRS for the six diseases of Table 1.

	T2D	CAD	CD	UC	SCZ	RA
T2D	0.0856					
CAD	0.0225	0.0398				
CD	−0.0111	0.0347	0.102			
UC	−0.0086	0.0191	0.0409	0.0553		
SCZ	−0.00131	0	0.00679	0.00480	0.0254	
RA	−0.038	−0.034	−0.00251	0.00566	−0.00185	0.0732

Note: Assumed to equal the liability-PRS covariance matrix Σ_{LX}.

Table 3. Genetic correlations between the six diseases of Table 1.

	T2D	CAD	CD	UC	SCZ	RA
T2D	1					
CAD	0.384	1				
CD	−0.119	0.057	1			
UC	−0.125	0.038	0.543	1		
SCZ	−0.028	0	0.113	0.128	1	
RA	−0.048	−0.063	−0.029	0.089	−0.043	1

Note: Assumed to equal the correlations between their overall liabilities Σ_L.

individually. The screening relative utility is -0.004, suggesting that these PRS provide no benefit in a multiple screening application. The liability variance explained is $1 - \Lambda = 0.332$, which of itself is higher than the individual R^2 (Table 1) but, as just seen, leads to lower values of several criteria of accuracy.

In principle, PRS could be developed that explain greater proportions of liability[48] up to the so-called SNP heritability (Table 1). Under this scenario the liability variance explained increases to $1 - \Lambda = 0.765$, giving a screening concordance of 0.664 and relative utility of 0.275. This suggests that further progress in genetic prediction may lead to more useful applications in multiple screening contexts, especially if further combined with non-genetic risk factors.

5 Discussion

Standard concepts of sensitivity and specificity generalise naturally to the multivariate setting. Positive and negative predictive values generalise similarly, and for completeness their definitions are provided in the supplementary text. Although the ROC curve does not extend so easily, the related concept of concordance does so. However, in contrast to the single outcome setting, concordance is sensitive to the outcome probabilities, negating one perceived advantage of that criterion. In the strong panel-wise sense the concordance is unsatisfying because an individual can be regarded as being discordant with itself, and there is no natural interpretation in terms of discrimination. The range of panel-wise concordance depends upon the number of outcomes and the covariance of predictors and outcomes, and may fall below 0.5. In practice its minimum value can be estimated by simulation or theory, as in section 4.2, by setting the predictors to be independent of the outcomes while maintaining the correlation among predictors and among outcomes. Strong panel-wise measures have an intermediate position between outcome-wise and screening measures, in that prediction is evaluated at the individual level but the predictions of specific outcomes are taken into account. The proposed definitions are motivated by possible applications in early detection of disease, and have convenient analogies with family-wise error in hypothesis testing, but other approaches may be possible.

Relative utility, which is a useful summary of sensitivity and specificity when predicting a single outcome, presents some difficulties when predicting multiple outcomes. I propose definitions assuming common benefits and costs for all outcomes, which allow analogous development to that for a single outcome, but may lead to sub-optimal assessment of utility when the benefits and costs vary across outcomes. When outcomes are correlated, accurate calculation of relative utility may be difficult, so approximations are provided assuming independent predictors and outcomes. It remains to be seen how useful these definitions prove in practice, given their assumptions of common additive benefits and costs, and independent predictors and outcomes.

Some examples of screening have been discussed, but examples of outcome-wise or joint accuracy can also be envisaged. CancerSEEK is a recent example of molecular technology applied to early detection of multiple cancers. Its performance was reported in the screening sense, but the proposed definitions clarify that all quantities can be affected by ascertainment bias. The present criteria are more sensitive to incidence and sampling rates than the corresponding univariate measures.

I have only considered the accuracy of a given predictor, and have not considered how such predictors are constructed. Multivariate predictors could be constructed simply by concatenating univariate predictors. The example of PRS shows that this is feasible and pragmatic given that such scores are currently constructed from case/control studies of individual diseases. In future, given the increasing availability of extensive phenotyping in large cohorts, it will be possible to build prediction models with the optimisation of multiple outcome prediction as the direct objective. Methodology for such model building is a fertile area for future work.

Prediction models are often evaluated for their improvement over existing models. Evaluation of incremental performance remains a controversial subject when predicting a single trait. Among several proposed measures the net reclassification index has attained a default status among practitioners yet has received strong criticism.[51,52] Such issues are likely to be magnified when predicting multiple traits.

Given predictors for a set of outcomes, a natural question is whether there is some subset of outcomes for which risk prediction is most effective. Naïve comparison of, say, relative utilities for different groups of outcomes would be inappropriate without consideration of the relative benefits of predicting each group. Thus, the finding that the screening concordance of PRS is lower over six diseases than for each disease individually should not in itself argue against a screening application, because the benefits and costs of screening six diseases are different from those of screening one disease. Many authors have argued for decision-theoretic treatments of risk prediction.[28,53] Such approaches can also be developed for the multiple outcome setting and would put the comparison of predictors for different groups of outcomes on a more coherent footing.

Competing risks present a problem for mutually exclusive outcomes, such as diseases of later life. There is a distinction between accounting for competing risks in model building, and in model evaluation. The emphasis here is on evaluation, for which the proposed criteria could be adapted to account for competing risks. However, the explicit consideration of multiple outcomes may encourage more careful consideration of competing risks at the model building stage and lead to improved prediction in general.

An R library to calculate these criteria from empirical data, and to evaluate the multivariate probit formulae of section 3, is available from https://github.com/DudbridgeLab/multipred.

Acknowledgement

I am grateful to Paul Newcombe for reading the manuscript, and to Richard Morris, Alex Sutton and Angelica Ronald for their helpful suggestions.

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.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Frank Dudbridge https://orcid.org/0000-0002-8817-8908

Supplemental material

Supplemental material for this article is available online.

References

1. Steyerberg EW. *Clinical prediction models*. New York: Springer-Verlag, 2009.
2. Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014; **129**: S49–S73.
3. Pashayan N, Morris S, Gilbert FJ, et al. Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: a life-table model. *JAMA Oncol* 2018; **4**: 1504–1510.
4. Feldman M, Stanford R, Catcheside A, et al. The use of a prognostic table to aid decision making on adjuvant therapy for women with early breast cancer. *Eur J Surg Oncol* 2002; **28**: 615–619.
5. Bae JH, Park SH, Ye BD, et al. Development and validation of a novel prediction model for differential diagnosis between Crohn's disease and intestinal tuberculosis. *Inflamm Bowel Dis* 2017; **23**: 1614–1623.
6. Gail MH. Twenty-five years of breast cancer risk models and their applications. *J Natl Cancer Inst* 2015; **107. DOI**: 10.1093/jnci/djv042.
7. Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ* 2010; **341**: c6624.
8. Gray LJ, Taub NA, Khunti K, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabet Med* 2010; **27**: 887–895.
9. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018; **359**: 926–930.
10. Chatterjee N, Shi J and Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016; **17**: 392–406.
11. Widschwendter M, Jones A, Evans I, et al. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat Rev Clin Oncol* 2018; **15**: 292–309.
12. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**: e1001779.
13. Horton R, Crawford G, Freeman L, et al. Direct-to-consumer genetic testing. *BMJ* 2019; **367**: l5688.
14. Cefalu WT, Bray GA, Home PD, et al. Advances in the science, treatment, and prevention of the disease of obesity: reflections from a diabetes care editors' expert forum. *Diab Care* 2015; **38**: 1567–1582.
15. Rothwell PM, Price JF, Fowkes FG, et al. Short-term effects of daily aspirin on cancer incidence, mortality, and non-vascular death: analysis of the time course of risks and benefits in 51 randomised controlled trials. *Lancet* 2012; **379**: 1602–1612.
16. Lello L, Avery SG, Tellier L, et al. Accurate genomic prediction of human height. *Genetics* 2018; **210**: 477–497.
17. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013; **14**: R115.
18. Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. *Lancet* 2018; **391**: 1897–1907.
19. Hippisley-Cox J and Coupland C. Development and validation of QMortality risk prediction algorithm to estimate short term risk of death and assess frailty: cohort study. *BMJ* 2017; **358**: j4208.
20. Barker M and Rayens W. Partial least squares for discrimination. *J Chemometr* 2003; **17**: 166–173.
21. Breiman L and Friedman JH. Predicting multivariate responses in multiple linear regression. *J Roy Stat Soc B Met* 1997; **59**: 3–37.
22. Jia Y and Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 2012; **192**: 1513–1522.
23. Bottolo L, Chadeau-Hyam M, Hastie DI, et al. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet* 2013; **9**: e1003657.
24. Maier R, Moser G, Chen GB, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet* 2015; **96**: 283–294.
25. Van Calster B, Van Belle V, Vergouwe Y, et al. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Stat Med* 2012; **31**: 2610–2626.

26. Zheng Y, Cai T, Jin Y, et al. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* 2012; **68**: 388–396.
27. Hand DJ. Measuring diagnostic accuracy of statistical prediction rules. *Stat Neerl* 2001; **55**: 3–16.
28. Gail MH and Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**: 227–239.
29. Pepe MS and Janes H. Methods for evaluating prediction performance of biomarkers and tests. *UW Biostatistics Working Paper Series* 2012; 384.
30. 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.
31. Lemeshow S and Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982; **115**: 92–106.
32. Beauchamp KA, Muzzey D, Wong KK, et al. Systematic design and comparison of expanded carrier screening panels. *Genet Med* 2018; **20**: 55–63. 2017/06/24.
33. Canonica GW, Ansotegui IJ, Pawankar R, et al. A WAO - ARIA - GA(2)LEN consensus document on molecular-based allergy diagnostics. *World Allergy Organ J* 2013; **6**: 17.
34. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn* 2009; **77**: 103–123.
35. Li JL and Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. *J R Stat Soc C-Appl* 2010; **59**: 673–692.
36. Vickers AJ and Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making* 2006; **26**: 565–574.
37. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
38. Baker SG, Cook NR, Vickers A, et al. Using relative utility curves to evaluate risk prediction. *J R Stat Soc A Stat* 2009; **172**: 729–748.
39. Baker SG. Putting risk prediction in perspective: relative utility curves. *Jnci-J Natl Cancer I* 2009; **101**: 1538–1542.
40. Morgan MD, Pairo-Castineira E, Rawlik K, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat Commun* 2018; **9**: 5271.
41. Khera AV, Chaffin M, Wade KH, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 2019; **177**: 587–596.
42. So HC and Sham PC. A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. *Plos Genet* 2010; **6**: e1001230.
43. Chib S and Greenberg E. Analysis of multivariate probit models. *Biometrika* 1998; **85**: 347–361.
44. Leppard P and Tallis GM. Algorithm as-249 - evaluation of the mean and covariance of the truncated multinormal distribution. *Appl Stat-J Royal Stat C* 1989; **38**: 543–553.
45. Wilhelm S and Manjunath BG. tmvtnorm: Truncated multivariate normal and student t distribution. 1.4-10 ed. 2015, p. R package.
46. Aitken AC. Note on selection from a multivariate normal population. *Proc Edinburgh Math Soc* 1935; **4**: 106–110.
47. Nadarajah S, Afuecheta E and Chan S. On the distribution of maximum of multivariate normal random vectors. *Commun Stat – Theory Meth* 2019; **48**: 2425–2445.
48. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013; **9**: e1003348.
49. Wray NR, Yang J, Goddard ME, et al. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010; **6**: e1000864.
50. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017; **33**: 272–279.
51. Kerr KF, Wang Z, Janes H, et al. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* 2014; **25**: 114–121.
52. 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. *Stat Biosci* 2015; **7**: 282–295.
53. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–138.
54. Yaghootkar H, Bancks MP, Jones SE, et al. Quantifying the extent to which index event biases influence large genetic association studies. *Hum Mol Genet* 2017; **26**: 1018–1030.
55. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* 2018; **50**: 1505–1513.
56. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018; **72**: 1883–1893.
57. Chen GB, Lee SH, Montgomery GW, et al. Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. *BMC Med Genet* 2017; **18**: 94.

58. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; 511: 421-427.
59. Mak TSH, Porsch RM, Choi SW, et al. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol* 2017; **41**: 469–480.
60. Stahl EA, Wegmann D, Trynka G, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* 2012; **44**: 483–489.
61. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* 2018; **9**: 2941.
62. Verweij N, Eppinga RN, Hagemeijer Y, et al. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci Rep* 2017; **7**: 2761.
63. Chen GB, Lee SH, Brion MJ, et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum Mol Genet* 2014; **23**: 4710–4720.