Eur J Epidemiol (2012) 27:761–770
DOI 10.1007/s10654-012-9733-3

# Assessing the discriminative ability of risk models for more than two outcome categories

**Ben Van Calster · Yvonne Vergouwe · Caspar W. N. Looman · Vanya Van Belle · Dirk Timmerman · Ewout W. Steyerberg**

Received: 14 May 2012 / Accepted: 14 September 2012 / Published online: 7 October 2012

**Abstract** The discriminative ability of risk models for dichotomous outcomes is often evaluated with the concordance index (*c*-index). However, many medical prediction problems are polytomous, meaning that more than two outcome categories need to be predicted. Unfortunately such problems are often dichotomized in prediction research. We present a perspective on the evaluation of discriminative ability of polytomous risk models, which may instigate researchers to consider polytomous prediction models more often. First, we suggest a "discrimination plot" as a tool to visualize the model's discriminative ability. Second, we discuss the use of one overall polytomous *c*-index versus a set of dichotomous measures to summarize the performance of the model. Third, we address several aspects to consider when constructing a polytomous *c*-index. These involve the assessment of concordance in pairs versus sets of patients, weighting by outcome prevalence, the value related to models with random performance, the reduction to the dichotomous *c*-index for dichotomous problems, and interpretation. We illustrate these issues on case studies dealing with ovarian cancer (four outcome categories) and testicular cancer (three categories). We recommend the use of a discrimination plot together with an overall *c*-index such as the Polytomous Discrimination Index. If the overall *c*-index suggests that the model has relevant discriminative ability, pairwise *c*-indexes for each pair of outcome categories are informative. For pairwise *c*-indexes we recommend the 'conditional-risk' method which is consistent with the analytical approach of the multinomial logistic regression used to develop polytomous risk models.

**Keywords** Polytomous risk prediction · Discrimination · *c*-index · Discrimination plot

## Introduction

Many diagnostic and prognostic risk models in epidemiology and clinical medicine are dichotomous, in that they predict an outcome with two complementary categories (event vs non-event) [1]. A generic example is the prediction of whether a tumor is benign or malignant. However, many prediction problems are in fact polytomous, meaning that the outcome has more than two categories. Particularly when the outcome categories have different treatment consequences, the polytomous nature should be taken into account in the risk model. Unfortunately, this is often not done [2].

The discriminative ability of dichotomous risk models is commonly expressed with the concordance probability or *c*-index [3, 4]. This measure assesses the probability that the model gives a higher score (e.g. probability of event) for a true event than for a true non-event, and equals the

B. Van Calster (✉) · D. Timmerman
Department of Development and Regeneration,
KU Leuven, University of Leuven, Herestraat 49 Box 7003,
3000 Leuven, Belgium
e-mail: ben.vancalster@med.kuleuven.be

B. Van Calster · Y. Vergouwe · C. W. N. Looman ·
E. W. Steyerberg
Department of Public Health, Erasmus MC, PO Box 2040,
3000 CA Rotterdam, The Netherlands

V. Van Belle
Department of Electrical Engineering, KU Leuven,
Kasteelpark Arenberg 10, 3001 Leuven, Belgium

V. Van Belle
IBBT- Future Health Department, KU Leuven,
Kasteelpark Arenberg 10, 3001 Leuven, Belgium

area under the ROC curve (AUC) [5]. Several discrimination measures have been suggested for polytomous risk models [6–14], but so far no standard measure exists. In this article, we aim to provide a perspective on how to assess the discriminative ability of polytomous risk models. We hope that this paper will help researchers to resort more easily to polytomous models in prediction research. We realize that other important aspects of model performance exist, such as calibration, explained variation, and decision-analytical measures [15]. However, these are beyond the scope of this article.

We use two case studies to illustrate important aspects that are related to the assessment of the discriminative ability of polytomous risk models. First, we visualize polytomous risk predictions. Next, we contrast a single polytomous discrimination measure with a set of dichotomous measures and discuss the properties a polytomous measure should contain.

## Case studies

### Ovarian tumor diagnosis

The treatment of women presenting with an ovarian tumor depends on the tumor type. Benign tumors can often be treated conservatively [16]. Borderline tumors may be treated with low-level surgical procedures that are beneficial when fertility preservation is an issue [17]. Primary invasive cancer is typically managed with laparotomy for staging, interval debulking surgery or cytoreduction, and carboplatin-based chemotherapy [18]. For metastatic invasive cancer, the optimal treatment depends on the primary cancer site. Existing models have focused on the dichotomous diagnosis of the tumor as benign vs malignant [19–22]. We consider a dataset containing 3,511 women with an adnexal mass [23, 24], to develop a polytomous prediction model using multinomial logistic regression [25] that includes nine predictors (Table 1) [26]. The dataset contains 72.9 % patients with a benign, 5.3 % with a borderline, 18.4 % with a primary invasive, and 3.4 % with a metastatic invasive tumor.

### Residual mass diagnosis after chemotherapy for testicular cancer

After cis-platin based chemotherapy in men with metastatic non-seminomatous testicular germ cell cancer, enlarged retroperitoneal lymph nodes may be detected. These can either consist of necrosis or fibrosis (benign tissue), mature teratoma, or viable cancer tissue. Because resection of benign tissue has no therapeutic value, existing models have focused on the dichotomous diagnosis of residual mass as benign tissue [27–30]. However, mature teratoma is non-invasive and is therefore less threatening than viable cancer. This impacts on the preferred treatment (with or without chemotherapy) such that it is useful to keep these outcomes separate when developing a prediction model. We used the same approach as for the previous case study to develop a polytomous prediction model, this time including five predictors on data from 1,094 patients from several studies [30]. Benign tissue was found in 38.9 % of the patients, mature teratoma in 48.9 %, and viable cancer in 12.2 % (Table 2).

## Defining the outcome: number of categories and measurement scale

The number of outcome categories should be realistic, and it is sensible that this strongly depends on treatment/management consequences. For the ovarian cancer study, tumors were categorized into 21 histology groups [24], which was reduced to four following clinical considerations. Further, the distinction between nominal (unordered) and ordinal (ordered) outcomes, although sharp in theory, is not always clear in practice. For example, we can define outcomes as ordinal if measured on an ordered scale, and as nominal otherwise. An example is the Glasgow Outcome Scale to describe recovery from traumatic brain injury [31], a five-level ordinal scale ranging from good recovery to death. However, in medical applications outcomes not measured on an ordinal scale can often be ordered in terms of perceived severity or invasiveness of related treatments, such as the mature teratoma and viable cancer categories in the testicular cancer study. There is no clear-cut separation, but it seems acceptable to be strict and assume ordinality only if measurements are on an ordinal scale. In this paper, we mainly refer to nominal outcomes but most issues also apply to ordinal outcomes.

## Visualizing the risk predictions for different outcome categories

For dichotomous models, the overall risk differentiation is sometimes visualized with box plots, that show the distribution of the risk of event separately for patients with the event and patients without the event [1]. We extend this approach to the polytomous situation using a *discrimination plot* (Fig. 1). The plot shows the distributions of predicted risks for each outcome category stratified by the actually observed outcome category. The prevalences of outcome categories are added as horizontal lines. The discrimination plot visualizes the extent to which discrimination between outcome categories is achieved.

**Table 1** Descriptive statistics for the ovarian tumor data

| | Benign N = 2,560, 72.9 % | Borderline N = 186, 5.3 % | Primary invasive N = 645, 18.4 % | Metastatic invasive N = 120, 3.4 % | Polytomous model: Odds ratios[a] |
|---|---|---|---|---|---|
| *Continuous variables, median ($p_{25}$-$p_{75}$)* | | | | | |
| Age, in years | 41 (31–52) | 48 (35–64) | 57 (50–67) | 55 (47–66) | 1.017; 1.041; 1.038 |
| Maximum diameter of the lesion, in mm | 61 (44–84) | 84 (52–160) | 88 (60–130) | 77 (54–137) | 1.014; 1.015; 1.014 |
| Normalized diameter of solid tumor part*, in % | 0 (0–8) | 20 (10–42) | 55 (37–82) | 56 (36–98) | 1.017; 1.045; 1.050 |
| *Dichotomous variables, n (%)* | | | | | |
| History of ovarian cancer | 20 (0.8 %) | 21 (11.3 %) | 7 (1.1 %) | 6 (5.0 %) | 18.1; 2.76; 13.0 |
| Acoustic shadows | 411 (16.1 %) | 6 (3.2 %) | 22 (3.4 %) | 5 (4.2 %) | 0.22; 0.12; 0.13 |
| Irregular cyst walls | 746 (29.1 %) | 141 (75.8 %) | 460 (71.3 %) | 73 (60.8 %) | 3.22; 4.33; 3.81 |
| Papillary structures with detectable flow | 94 (3.7 %) | 79 (42.5 %) | 191 (29.6 %) | 17 (14.2 %) | 6.81; 5.11; 2.30 |
| Ascites | 46 (1.8 %) | 17 (9.1 %) | 268 (41.6 %) | 49 (40.8 %) | 2.40; 9.70; 9.38 |
| Tumors on both ovaries | 358 (14.0 %) | 26 (14.0 %) | 206 (31.9 %) | 33 (27.5 %) | 1.29; 2.78; 2.26 |

* The normalized diameter of the solid tumor part is computed as the ratio of the maximum diameter of the solid part and the maximum diameter of the lesion

[a] The polytomous model is a multinomial logistic regression model that uses ‘benign’ as the reference outcome category. The reported odds ratios thus represent, in this order, the odds ratio for borderline versus benign, primary invasive versus benign, and metastatic versus benign

**Table 2** Descriptive statistics for the testicular cancer data

| | Benign N = 425, 38.9 % | Mature teratoma N = 535, 48.9 % | Viable cancer N = 134, 12.2 % | Polytomous model: Odds ratios[a] |
|---|---|---|---|---|
| *Continuous variables, median ($p_{25}$-$p_{75}$)* | | | | |
| Maximum diameter of residual mass after chemotherapy, in mm | 18 (10–28) | 30 (20–64) | 41 (25–100) | 1.04; 1.22[b] |
| Reduction in mass size after chemotherapy, in % | 60 (42–80) | 20 (0–59) | 34 (0–60) | 0.81; 0.86 |
| *Dichotomous variables, n (%)* | | | | |
| Elevated alpha-fetoprotein | 225 (52.9 %) | 423 (79.1 %) | 107 (79.9 %) | 0.35; 0.37 |
| Elevated human chorionic gonadotrophin | 241 (56.7 %) | 381 (71.2 %) | 94 (70.2 %) | 0.62; 0.70 |
| Teratoma elements in primary tumor | 146 (34.4 %) | 365 (68.2 %) | 80 (59.7 %) | 0.28; 0.46 |

[a] The polytomous model is a multinomial logistic regression model that uses ‘benign’ as the reference outcome category. The reported odds ratios thus represent, in this order, the odds ratio for teratoma versus benign, and viable cancer versus benign

[b] This variable was entered in the model using a square root transformation [30]

For example, for the ovarian model the predicted risk of a benign tumor should be higher for patients with a benign tumor compared to patients with another tumor type. The four leftmost box plots in Fig. 1a show that this was very often the case.

## What do we want to assess? One overall measure versus a set of measures

Whereas discrimination plots nicely visualize the discriminative ability of polytomous risk models, quantification is also desired. We distinguish between two general strategies to assess the discriminative ability for polytomous models: one overall measure or a set of dichotomous measures. In what follows, we use $k$ to denote the number of outcome categories.

### A set of dichotomous $c$-indexes to summarize polytomous discrimination

Summarizing performance through a set of dichotomous $c$-indexes is possible in at least three ways. First, the outcome can be dichotomized for each category by putting all patients that belong to another category into a rest group. In this way $k$ dichotomous $c$-indexes can be computed, one for each category. Using the ovarian tumor diagnosis example, we could place all patients with a borderline, primary invasive, or metastatic invasive tumor into a ‘non-benign’ rest group, and a $c$-index for this dichotomization can be

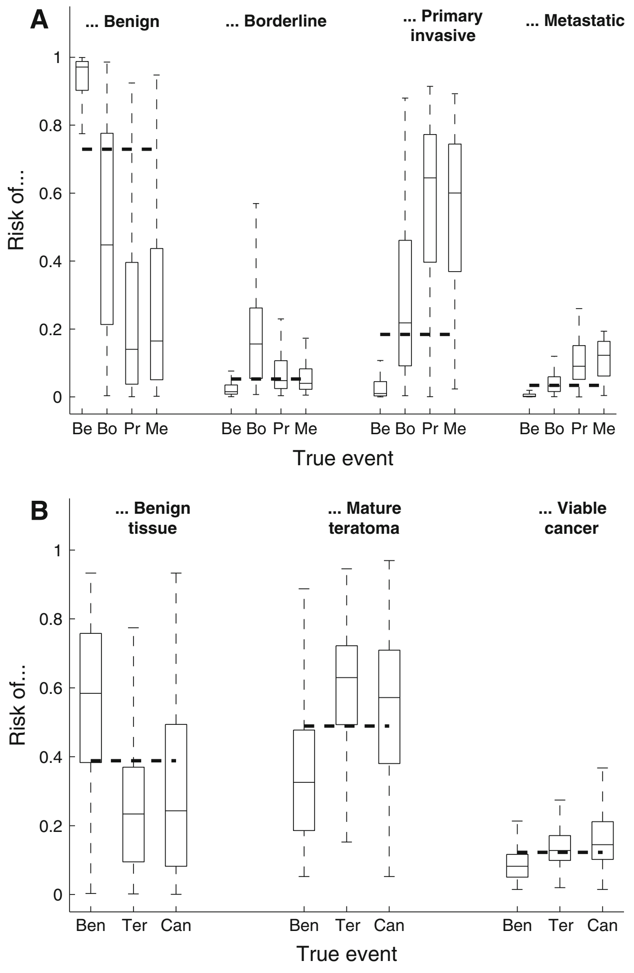

**Fig. 1** Discrimination plots of predictions for the ovarian tumor (**a**) and testicular cancer (**b**) models. The *dashed horizontal lines* show the prevalence of each outcome category. The *upper whiskers* extend to the largest data point that is not larger than the third quartile + 1.5 times the interquartile range, the *lower whiskers* extend the smallest data point that is not smaller than the first quartile − 1.5 times the interquartile range. Data points beyond the *whiskers* are not shown

computed using the risk of a benign tumor. This type of dichotomous comparisons can be labeled 1-versus-rest. Second, we can compare each pair of categories using only those patients that belong to one of the two categories at hand. There are $0.5*k*(k-1)$ pairs of categories for which a pairwise *c*-index can be computed. For the ovarian cancer case study there are four categories, hence six pairs of categories. Third, a stepwise or tree approach can be useful in some situations. In the first step, the *k* categories are split up into two groups of interest. Next, groups are further divided into subgroups until each subgroup contains only one category. For example, ovarian tumors can be divided into benign versus non-benign tumors. Next, non-benign tumors can be divided into borderline versus invasive tumors. Finally, invasive tumors are divided into primary and metastatic. This approach uses $k-1$ *c*-indexes to summarize performance, i.e. three for the ovarian cancer example.

In our opinion, the pairwise approach is superior to the 1-versus-rest approach. The comparison of a category of interest (e.g. primary invasive tumors) with a rest category that contains all other cases (e.g. all patients with a benign, borderline, or metastatic invasive tumor) may obscure relevant results. For example, excellent or poor discrimination between the category of interest and another category in the rest group may go unnoticed, particularly if the category in the rest group has low prevalence. Generally, the category in the rest group with highest prevalence will dominate the results. This drawback is avoided with pairwise *c*-indexes. Let us illustrate this on the ovarian tumor model. Pairwise discrimination between primary invasive and metastatic invasive tumors is low, as visualized in Fig. 1a by the two rightmost box plots and the fifth and sixth box plots from the right. Nevertheless, the discrimination between primary invasive tumors and other tumors is very good, as visualized by the box plots in Fig. 2. The reason is that 89 % of the patients in the rest category have a benign tumor such that the 1-versus-rest *c*-index will almost entirely reflect the discrimination between primary invasive and benign tumors.

The pairwise approach poses a practical problem. Because there are more than two categories, the estimated risks of two categories A and B for an individual patient do not sum to one. Hence there are multiple ways to compute a pairwise *c*-index: (1) as the average of the *c*-index obtained with the estimated risk for category A ($P_A$) and the *c*-index obtained with the estimated risk for category B ($P_B$) ('avg' method) [7], (2) as the *c*-index obtained with the difference between both risks ($P_A - P_B$) ('diff' method) [8], or (3) as the *c*-index obtained using the conditional risk of category A [$P_A/(P_A + P_B)$] ('conditional-

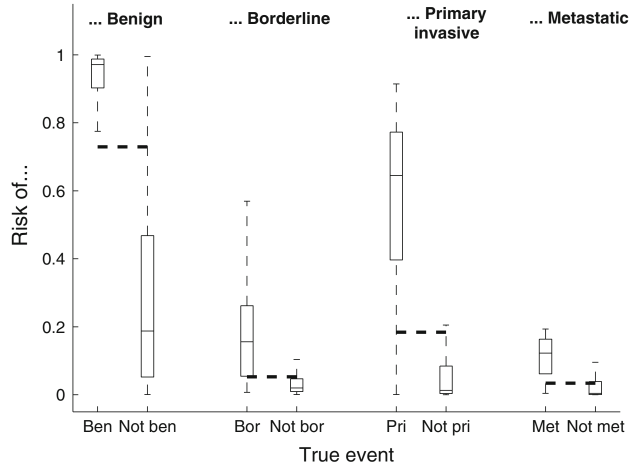

**Fig. 2** *Box plots* with predictions for each category versus all other categories (1-vs-rest) for the ovarian tumor model

risk' method). The 'conditional-risk' method appears most sensible, and corresponds to the setup of the standard 'baseline-category' multinomial logistic regression model [25]. This model compares each category with a reference category using a regression equation. For the testicular cancer study, the baseline-category model derives one regression equation for benign tissue versus viable cancer and one for mature teratoma versus viable cancer. To compare benign tissue and mature teratoma, the regression equation for mature teratoma versus viable cancer is subtracted from the regression equation for benign tissue versus viable cancer. Using these regression equations to compute pairwise *c*-indexes coincides with the 'conditional-risk' method.

### Polytomous discrimination and an overall polytomous *c*-index

A polytomous *c*-index aims to quantify the extent to which the model is able to discriminate between all *k* categories, as visualized in the discrimination plot, into a single number. Discriminatory ability of a polytomous model in essence refers to the situation where, to some extent, estimated probabilities for a specific category are higher for patients that belong to this category than for other patients.

The aim of a polytomous risk model is usually to discriminate between all categories. Hence we consider a polytomous *c*-index to be the first step as it provides an overall summary. Starting with a set of dichotomous results would give a scattered impression of model performance. If the overall *c*-index indicates that the model improves on pure chance discrimination, dichotomous measures can be investigated to obtain detailed information. Sometimes, however, the goal of the polytomous model may not be the investigation of simultaneous discrimination between all categories. For example, there may be specific interest into the discrimination between a reference category and each of the other categories separately. For this 'umbrella' situation [13], one overall polytomous *c*-index may not be desired. Instead pairwise *c*-indexes involving the reference category may suffice, or else a single *c*-index that joins these pairwise indexes [13].

## Aspects of polytomous *c*-indexes

### Determining concordance using sets or pairs of patients

For dichotomous outcomes, the *c*-index equals the proportion of concordant pairs among all comparable pairs of patients. A pair is comparable if the patients belong to different outcome categories, i.e. one event and one non-event. The pair is concordant if the estimated risk of event is highest for the patient with the event. There are two natural ways to extend the *c*-index to a polytomous outcome. The first is based on the investigation of pairs of two patients from different categories, the second is based on the investigation of sets of patients containing one patient from each category. The former is pragmatically more attractive whilst the latter is in our opinion conceptually preferable for the assessment of simultaneous discrimination between all categories.

Pair approaches suggested in the literature boil down to averaging pairwise [7, 8] or 1-vs-rest [9] *c*-indexes. The latter approach yields an intrinsically prevalence-dependent polytomous *c*-index because a 1-vs-rest *c*-index is dominated by highly prevalent categories in the rest group.

Polytomous *c*-indexes based on the set approach have been advocated as well [6, 11]. Here, the specific challenge is how to evaluate a set of patients, i.e. the definition of concordance. Several answers can be given to the question when a set of *k* patients is correctly predicted by the model [6, 32]. However, as described earlier, the definition of concordance is also not clear-cut for approaches that average pairwise *c*-indexes.

### Weighting by prevalence

Whether or not to weight a polytomous *c*-index based on category prevalences is a controversial topic. The standard dichotomous *c*-index is designed to be prevalence-independent. Even though this property of the *c*-index is often considered an advantage, this idea does not extend unequivocally to polytomous models. Existing measures for polytomous discrimination are sometimes weighted [8, 9] and sometimes not [6, 7, 11]. We believe that a prevalence-independent polytomous *c*-index is a natural extension of the standard dichotomous *c*-index.

A weighted index incorporates information related to the consequences of the predictions. Models that cannot discriminate a rare category from other categories will not produce a large absolute number of misclassifications whereas failure to discriminate a common category from other categories will. In contrast, an unweighted index is informative of sheer discrimination between categories as rare events are given the same importance as common events. Consider a hypothetical example with four categories (Fig. 3): category D (prevalence 4 %) can be excellently discriminated from the other categories (pairwise *c*-indexes with every other category are 0.95). Categories A-C have higher prevalence (32 %) but cannot be discriminated from one another (pairwise *c*-indexes are 0.50). An unweighted polytomous *c*-index based on the average of all pairwise *c*-indexes (a pair approach) is 0.73. When we weight pairwise *c*-indexes with the product of the prevalences of the two events, the polytomous *c*-index is

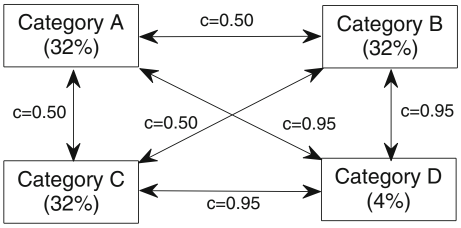

**Fig. 3** Schematic representation of a fictitious polytomous prediction problem with four categories. The figure shows the prevalence of each category (%) as well as the pairwise *c*-index between all pairs of categories (*c*)

0.55. Thus, if practical consequences related to category prevalence are taken into account overall discrimination is close to 0.5, the value indicating random performance for this measure. If prevalences are ignored, we would better detect the discrimination between category D and categories A-C.

Thus, when weighting by prevalence we implicitly express interest in the consequences of the model predictions. Utility, however, is based on a combination of prevalence and misclassification costs. A simple prevalence-weighted *c*-index is thus only partly utility-based. In addition, prevalence and misclassification costs are often inversely related in medical applications. In our opinion, an initial phase of performance evaluation should focus on sheer discrimination. A subsequent phase would then involve an evaluation of utility that takes both prevalence and misclassification costs into account [33, 34].

On a related note, the stability of the *c*-index over settings with varying prevalence can be questioned. The prevalence-independent property of the dichotomous *c*-index only holds when the conditional risk distributions remain invariant. In practice, prevalence differences across studies are often associated with different case-mix, different or imperfect reference standards, or distorted inclusion criteria [35–38]. In the testicular cancer example, a higher prevalence of viable cancers may go together with larger residual mass sizes for patients in this category.

### Random performance

A useless model cannot discriminate at all between the outcome categories, and is no different from making random predictions. For some existing polytomous *c*-indexes the value for random performance is 0.5 irrespective of $k$ [7–9], whereas for others it is $1/k$ or $1/k!$ [6, 11]. Not accidentally, the former indexes are based on pairs and the latter on sets. A fixed value of 0.5, in analogy with the dichotomous *c*-index may seem convenient. Still, depending on the number of categories the discrimination quality of a model will evolve differently from random to perfect. For higher values of $k$, it will be more difficult to obtain high values for the index. If the random performance value depends on $k$, this difference is made explicit as the distance between random and perfect performance increases with $k$. Nevertheless, $1/k!$ rapidly decreases with increasing $k$: with four categories the random performance value is only 0.04. If random performance has value $1/k$, the decrease is much slower (e.g. 0.25 with four categories).

### Interpretation

A reasonable property of a polytomous *c*-index is that it reduces to the standard dichotomous *c*-index when the number of categories is two. The dichotomous *c*-index has a clear and interesting interpretation, even though it has been criticized as artificial and inconsistent with clinical practice [39, 40]. It is essential to consider the interpretation of any polytomous *c*-index extension.

## Existing measures

In the previous sections we have referred to several measures that have been suggested in the literature for nominal outcomes. We will now shortly discuss these (see Table 3 for a summary [11]). More information is given elsewhere [11].

The *M*-index [7] is a pair approach that averages all pairwise *c*-indexes. It estimates the probability to correctly distinguish between a pair of patients from two randomly chosen categories. The index proposed by Obuchowski and colleagues [8] is suggested as a weighted measure but that, if unweighted, is similar to the *M*-index. The difference is that the *M*-index computes pairwise *c*-indexes using the 'avg' method whereas Obuchowski's index uses the 'diff' method. The ad hoc approach used by Provost and Domingos [9] is a pair approach that averages all 1-vs-rest *c*-indexes. It estimates the probability to correctly distinguish between a patient from a randomly chosen category and a patient from any other category. Provost and Domingos use prevalence-weighting, which is unnecessary given that 1-vs-rest *c*-indexes implicitly depend on category prevalences.

The volume under the surface (VUS) [6] and Polytomous Discrimination Index (PDI) [11] are set approaches. VUS estimates the probability correctly distinguish between a set of $k$ patients, with random performance at $1/k!$. PDI estimates the probability that a patient from a randomly chosen category is correctly identified within a set of patients. For PDI $1/k$ reflects random performance.

The VUS and PDI use a different definition of concordance. For the VUS, a geometric criterion is often used [6, 41]. A patient can be represented in a $k$-dimensional plot using the estimated risks for each category. Perfect

**Table 3** Summary of existing $c$-indexes for nominal polytomous outcomes [11]

| Measure | Set/pair | Random | Reduction to standard $c$-index | Weighting | Interpretation (if unweighted): probability to … |
|---|---|---|---|---|---|
| M-index | P | 0.5 | Yes | Possible | Correctly discriminate between two patients from two randomly selected categories |
| Obuchowski | P | 0.5 | Yes | Possible | Correctly discriminate between two patients from two randomly selected categories |
| 1-vs-Rest | P | 0.5 | Yes | Possible* | Correctly discriminate between a patient from a randomly selected category and a patient from another category |
| VUS | S | $1/k!$ | Yes | No | Correctly discriminate between a set of patients |
| PDI | S | $1/k$ | Yes | Possible | Correctly identify a patient from a randomly selected category within a set of patients |

* This measure implicitly weights by prevalence because of its set-up, additional explicit weighting is possible

prediction means that the estimated risk is 1 for the correct category and 0 for any other category. When representing a set of patients in a $k$-dimensional plot, the aggregate distance from each patient to its perfect position is computed. The set is concordant if this distance is smaller than any other aggregate distance obtained by linking the $k$ patients to the $k$ perfect positions. For the PDI, a set of $k$ patients is fully concordant if it holds for each category that the estimated risk is highest for the patient that belongs to this category. A set can be partially concordant if this rule holds for some but not all categories. The score given to a set equals the number of categories for which the rule holds divided by $k$.

A similar discussion of existing measures for ordinal outcomes [4, 10, 12, 14] can be found elsewhere [14]. The VUS-type $c$-index for the umbrella situation is found in work by Nakas and Alonzo [13].

## Application on the case studies

Results for the different polytomous $c$-indexes for the two applications are given in Table 4, followed by all 1-vs-rest and pairwise $c$-indexes to allow for a detailed investigation of the models' performance. The polytomous $c$-indexes based on pairwise category comparisons ($M$-index and unweighted Obuchowski) gave similar results: 0.82 and 0.84 for the ovarian tumor model, 0.73 and 0.74 for the testicular cancer model. For weighted alternatives the performance increased because the most prevalent categories were better distinguished from other categories than less prevalent categories (see the pairwise $c$-indexes). This was more pronounced for the ovarian tumor model. The unweighted polytomous $c$-index based on 1-versus-rest $c$-indexes also resulted in higher performance for the ovarian tumor model (0.90) due to the implicit role of prevalence in this measure: the near random discrimination between primary invasive and metastatic invasive tumors is obscured due to the very low prevalence of metastatic invasive tumors (3.4 %). The 1-versus-rest $c$-index for metastatic tumors was 0.88 because of the excellent discrimination from the highly prevalent benign tumors. For the testicular cancer model, the 1-versus-rest measure (0.75) was only mildly higher than the unweighted pair approaches as the link between prevalence and discrimination was less strong for this application. Explicit weighting of this measure further increased the resulting value.

Because they are based on setwise comparisons, VUS and PDI have clearly lower values. For the testicular cancer model, the VUS indicates that there is a 41 % chance that a set of three patients is correctly separated by the model. The PDI estimates that there is a 58 % chance that a patient from a randomly selected category is correctly identified within a set of patients. For the ovarian tumor diagnosis model, the VUS and PDI equal 0.36 and 0.61, respectively.

The one-vs-rest $c$-indexes per category obscure interesting findings revealed by the pairwise $c$-indexes. For example, in the testicular cancer model the ability to discriminate mature teratoma from benign tissue or viable cancer was markedly different (0.8 vs 0.6). Yet the one-vs-rest $c$-index for mature teratome approached 0.8 due to the low prevalence of viable cancer. For the ovarian tumor model, a similar observation was made as described earlier in this section. The pairwise $c$-indexes give a superior summary of the discrimination plot of a model, but Table 4 shows that the method used to compute these pairwise indexes can have a strong impact on the obtained results. The 'avg' method gave the lowest values whereas the 'conditional-risk' method gave the highest values, the difference in values between these methods was sometimes large (up to 0.09).

**Table 4** Discrimination of the prediction models for ovarian tumor diagnosis and testicular cancer diagnosis: polytomous *c*-indexes and dichotomous *c*-indexes comparing one event with all other events (1-vs-rest) or with one other event pairwise

| Discrimination measure | | Ovarian tumor diagnosis | | Residual testicular cancer diagnosis |
|---|---|---|---|---|
| *Polytomous* c-*indexes* | | | | |
| *M*-index | | 0.82 | | 0.73 |
| Obuchowski, unweighted | | 0.84 | | 0.74 |
| Obuchowski, weighted | | 0.93 | | 0.78 |
| 1-versus-rest, unweighted | | 0.90 | | 0.75 |
| 1-versus-rest, weighted | | 0.93 | | 0.78 |
| VUS | | 0.36 | | 0.41 |
| PDI | | 0.61 | | 0.58 |
| *Dichotomous* c-*indexes* | | | | |
| 1-versus-rest *c*-indexes | Benign: | 0.94 | Benign tissue: | 0.82 |
| | Borderline: | 0.85 | Mature teratoma: | 0.77 |
| | Pr. Invasive: | 0.93 | Viable cancer: | 0.65 |
| | Metastatic: | 0.88 | | |
| Pairwise *c*-indexes* | Ben versus Bor: | 0.89/0.89/0.90 | Ben versus Ter: | 0.83/0.83/0.83 |
| | Ben versus PrI: | 0.96/0.96/0.96 | Ben versus Can: | 0.77/0.79/0.79 |
| | Ben versus Met: | 0.96/0.96/0.96 | Ter versus Can: | 0.57/0.61/0.64 |
| | Bor versus PrI: | 0.76/0.83/0.85 | | |
| | Bor versus Met: | 0.79/0.85/0.88 | | |
| | PrI versus Met: | 0.56/0.57/0.63 | | |

* Pairwise *c*-indexes for category A versus B are computed using three approaches: the first is the 'avg' method used in the *M*-index, the second is the 'diff' method used in Obuchowski's measure, the third is the 'conditional-risk' method

## Discussion and recommendations

This paper discussed the evaluation of the discrimination performance of polytomous prediction models. Many issues come into play, and because of this several possible extensions of the *c*-index can be conceived as evidenced by the literature.

### Lessons from the case studies

The case studies show that different approaches to construct a polytomous *c*-index result in highly different probabilities. For example, the results varied between 0.36 and 0.93 for the ovarian tumor model. It is therefore crucial to understand the interpretation and setup of the measures when using them.

The results confirm that dichotomizing the outcome categories by comparing one category with a rest group containing all other categories is suboptimal. The 1-vs-rest overall *c*-index can be artificially high as the creation of a rest group with categories of varying prevalence can obscure relevant information. It is also clear that the method used to compute pairwise *c*-indexes can have strong influence. The unweighted Obuchowski measure uses the 'diff' method and yielded slightly higher values than the *M*-index that uses the 'avg' method. When the pairwise *c*-indexes were considered individually, the differences between both approaches were sometimes substantial. Yet, if the 'conditional-risk' method was used, which no existing measure does, the pairwise results were even higher. The 'avg' method probably underestimates pairwise discrimination because the risks for the two categories at hand are evaluated separately. The 'conditional-risk' method captures pairwise discrimination most efficiently. It corrects for the risks of the remaining categories such that the risks of the two categories at hand sum to one. Importantly, a strong argument in favor of this method is that it corresponds to the setup of the standard multinomial logistic regression model.

### Recommendations for practice

We recommend the use of a discrimination plot to visualize the achieved risk differentiation, followed by an overall polytomous *c*-index. In case the overall *c*-index suggests discriminatory capacity, pairwise *c*-indexes can be computed with the 'conditional-risk' method to get a more detailed picture. We acknowledge that, in some situations, a stepwise or tree approach can be more desirable than a pairwise approach.

As for the overall polytomous *c*-index, we believe that the PDI is the measure whose set-up corresponds most closely to what is needed [11]. A set approach focuses on an assessment of simultaneous discrimination between all categories. A pair approach does not evaluate all outcome categories jointly, thus making it a less coherent polytomous measure. PDI gives a graded score to sets such that it captures more information than the VUS. It also results in a value of $1/k$ for randomly performing models, which is more convenient than $1/k!$. It is natural that the value for random prediction decreases with increasing $k$ as correct prediction becomes harder. A polytomous $c$-index with range 0.5–1 irrespective of $k$ seems convenient, but the meaning of a specific result still varies depends on $k$.

We further recommend not to weight for prevalence when evaluating polytomous discrimination, in order to focus on category-based discrimination. Thereafter the model should ideally be evaluated in a utility framework that takes prevalence and misclassification costs into account. Unfortunately, research on the latter is scarce for polytomous models [42, 43]. Yet we advise that conclusions regarding the usefulness of a model are not merely drawn from an evaluation of discriminatory ability. In addition to utility, calibration is another important aspect of prediction models that should be assessed [1].

### Ordinal outcomes

This paper mainly focused on nominal outcomes. Most discussions in this paper apply to ordinal outcomes as well, with two notable exceptions. Firstly, for ordinal outcomes where prediction models give predictions on a single scale, for example using proportional odds regression, the ordinality reduces the difference between pair- and set approaches: an ordinal variant of the PDI (ordinal $c$-index, ORC) was constructed as a set approach but can be rewritten as a simple average of pairwise $c$-indexes [14]. Secondly, a random performance value of 0.5 is more attractive for ordinal than for nominal outcomes [14]. If a model ranks a set of patients randomly, on average half of the pairs of patients within the set will be incorrectly ordered irrespective of $k$. For ordinal outcomes, the ORC is our recommended polytomous $c$-index.

### *c*-index/AUC and the ROC curve

We have considered polytomous discrimination by extending the $c$-index to polytomous outcomes. Researchers have also investigated the extension of ROC curves to polytomous outcomes. The $c$-index and ROC frameworks overlap, obviously. For example, the VUS refers to a volume under a multidimensional ROC surface [6]. We did not use links to ROC curves as their extension to outcomes with more than three categories becomes unpractical.

### Concluding comments

There are many opportunities for the use of polytomous risk prediction models in clinical medicine and epidemiology. However, polytomous models are more complicated than dichotomous models. The evaluation of their performance, for example, is not straightforward. In this perspective paper, we tried to present an overview of issues involved in assessing polytomous discrimination performance and summarized these by presenting recommendations for practice. Further work on other performance aspects for polytomous models is needed, but we hope that the present work may instigate researchers to consider polytomous prediction models more often.

**Acknowledgments** This work was supported by the Research Foundation—Flanders (FWO) (grants 1.2516.09 N, 1.2516.12 N, G.0493.12 N), Agency for Innovation by Science and Technology (IWT Vlaanderen) (grant TBM070706-IOTA3), the Research Council of the KU Leuven, and the Netherlands Organization for Scientific Research (grant 9120.8004). Ben Van Calster is a postdoctoral fellow of the Research Foundation—Flanders (FWO). Vanya Van Belle is supported by a postdoctoral fellowship from KU Leuven's Special Research Fund (BOF) and a postdoctoral fellowship from FWO.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
2. Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KGM. Polytomous logistic regression analysis could be applied more often in diagnostic research. J Clin Epidemiol. 2008;61:125–34.
3. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
4. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
5. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
6. Mossman D. Three-way ROCs. Med Decis Making. 1999;19:78–89.
7. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–86.
8. Obuchowski NA, Goske MJ, Applegate KE. Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. Stat Med. 2001;20:3261–78.
9. Provost F, Domingos P. Tree induction for probability-based ranking. Mach Learn. 2003;52:199–215.
10. Obuchowski NA. Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. Acad Radiol. 2005;12:1198–204.
11. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the $c$-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. Stat Med. 2012;31:2610–26.

12. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Stat Med. 2004;23:3437–49.
13. Nakas CT, Alonzo TA. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. Biometrics. 2007;63:603–9.
14. Van Calster B, Van Belle V, Vergouwe Y, Steyerberg EW. Discrimination ability of prediction models for ordinal outcomes: relationship between existing measures and a new measure. Biom J. 2012;54:674–85.
15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
16. Panici PB, Muzii L, Palaia I, Manci N, Bellati F, Plotti F, et al. Minilaparotomy versus laparoscopy in the treatment of benign adnexal cysts: a randomized clinical study. Eur J Obstet Gynecol Reprod Biol. 2007;133:218–22.
17. Tinelli R, Tinelli A, Tinelli FG, Cicinelli E, Malvasi A. Conservative surgery for borderline ovarian tumors: a review. Gynecol Oncol. 2006;100:185–91.
18. Hennessy BT, Coleman RL, Markman M. Ovarian cancer. Lancet. 2009;374:1371–82.
19. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. A logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis (IOTA) group. J Clin Oncol. 2005;23:8794–801.
20. Van Holsbeke C, Van Calster B, Testa AC, Domali E, Lu C, Van Huffel S, et al. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the International Ovarian Tumor Analysis Study. Clin Cancer Res. 2009;15:684–91.
21. Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. Ultrasound Obstet Gynecol. 2010;36:226–34.
22. Van Holsbeke C, Van Calster B, Bourne T, Ajossa S, Testa AC, Guerriero S, et al. External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. Clin Cancer Res. 2012;18:815–25.
23. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstet Gynecol. 2000;16:500–5.
24. Van Calster B, Valentin L, Van Holsbeke C, Zhang J, Jurkovic D, Lissoni AA, et al. A novel approach to predict the likelihood of specific ovarian tumor pathology based on serum CA-125: a multicenter observational study. Cancer Epidemiol Biomarkers Prev. 2011;20:2420–8.
25. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
26. Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol. 2010;10:96.
27. Steyerberg EW, Keizer HJ, Fosså SD, Sleijfer DT, Toner GC, Schraffordt Koops H, et al. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol. 1995; 13:1177–87.
28. Steyerberg EW, Gerl A, Fosså SD, Sleijfer DT, de Wit R, Kirkels WJ, et al. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. J Clin Oncol. 1998;16:269–74.
29. Vergouwe Y, Steyerberg EW, de Wit R, Roberts JT, Keizer HJ, Collette L, et al. External validity of a prediction rule for residual mass histology in testicular cancer: an evaluation for good prognosis patients. Br J Cancer. 2003;88:843–7.
30. Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fosså SD, Gerl A, et al. Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. Eur Urol. 2007;51:424–32.
31. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5:e165.
32. Van Calster B, Van Belle V, Condous G, Bourne T, Timmerman D, Van Huffel S. Multi-class AUC metrics and weighted alternatives. In: Liu D, Kozma R, editors. Proceedings of the 21st international joint conference on neural networks. Los Alamitos: IEEE Computer Society; 2008. p. 1391–7.
33. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13.
34. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26: 565–74.
35. Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62:5–12.
36. Webb GI, Ting KM. On the application of ROC analysis to predict classification performance under varying class distributions. Mach Learn. 2005;58:25–32.
37. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140:189–202.
38. Moons KGM, van Es GA, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12–7.
39. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer (editorial). J Natl Cancer Inst. 2008;100:978–9.
40. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions using risk stratification tables. Ann Intern Med. 2008;149:751–60.
41. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making. 2000;20:323–31.
42. Skaltsa K, Jover L, Fuster D, Carrasco JL. Optimum threshold estimation based on cost function in a multistate diagnostic setting. Stat Med. 2012;31:1098–109.
43. O'Brien DB, Gupta MR, Gray RM. Cost-sensitive multi-class classification from probability estimates. In: Cohen WW, McCallum A, Roweis ST, editors. Proceedings of the 25th international conference on machine learning. New York: Association for Computing Machinery; 2008. p. 712–9.