Int J Public Health (2013) 58:485–488
DOI 10.1007/s00038-012-0401-x

HINTS & KINKS

Use of areas under the receiver operating curve (AROCs) and some caveats

B. Kowall · W. Rathmann · K. Strassburger

Received: 20 December 2011 / Revised: 25 July 2012 / Accepted: 2 August 2012 / Published online: 4 September 2012

AROCs in public health research

Assessment of diagnostic and prognostic scores is an important public health issue, and is needed to validate the performance of scores, to find out which of several scores does best, or to assess whether a given score improves after inclusion of additional variables. To make this decision, areas under the receiver operating curves (AROCs), sometimes called c-values, are widely used.

As indicated by its name, AROC is the area under the ROC curve, that is the curve obtained by plotting the true positive rate (Sensitivity) against the false positive rate (1-Specificity) for all possible cut-offs of a given disease score. In logistic regression the AROCs can be calculated from all discordant pairs of subjects each consisting of one case and one non-case. The AROC is identical to the proportion of pairs for which the estimated probability of getting the disease is larger for the case than for the non-case (Table 1). This has been extended to survival time analyses. The principal idea is to compare all pairs of individuals with either two events or with one event and one non-event, and the so called overall C accords to the proportion of comparisons where the predicted probability of survival is larger for the individual with the larger actual survival time (Pencina and D′Agostino 2004).

Theoretically, AROCs range from 0 to 1. The worst value possible is 0.5, i.e., the model is no better than random guess. Values below 0.5 indicate that the model is informative in predicting the reverse event.

As an example used for demonstration throughout this paper, in the KORA (Cooperative Health Research in the Region of Augsburg) cohort study three models for the prediction of incident diabetes from baseline variables were compared (Rathmann et al. 2010):

Model 1: age, sex, hypertension, BMI, parental diabetes, smoking
Model 2: model 1 + fasting glucose, HbA1c, and uric acid
Model 3: model 2 + 2-h plasma glucose.

ROC curves for these models are shown in Fig. 1. For the original data set, AROCs from logistic regression models increase from 0.763 in model 1 to 0.844 in model 2 and to 0.886 in model 3. This means that the simple model including only non-invasive variables performs fairly well,

Table 1 Calculation of AROC for a fictive study group of 5 subjects (AROC: area under the receiver operating curve)

Subject	Incident disease	Predicted probability (p)	Discordant pairs	p (case) > p (non-case)
1	Case	0.18	Subjects 1, 3	Yes
2	Case	0.12	Subjects 1, 4	Yes
3	Non-case	0.08	Subjects 1, 5	Yes
4	Non-case	0.16	Subjects 2, 3	Yes
5	Non-case	0.06	Subjects 2, 4	No
			Subjects 2, 5	Yes

Number of discordant pairs consistent of one case and one non-case = 6

Number of discordant pairs with p(case) > p(non-case) = 5

AROC = 5/6 = 0.83

B. Kowall (✉) · W. Rathmann · K. Strassburger
Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Auf'm Hennekamp 65, 40225, Düsseldorf, Germany
e-mail: bernd.kowall@ddz.uni-duesseldorf.de

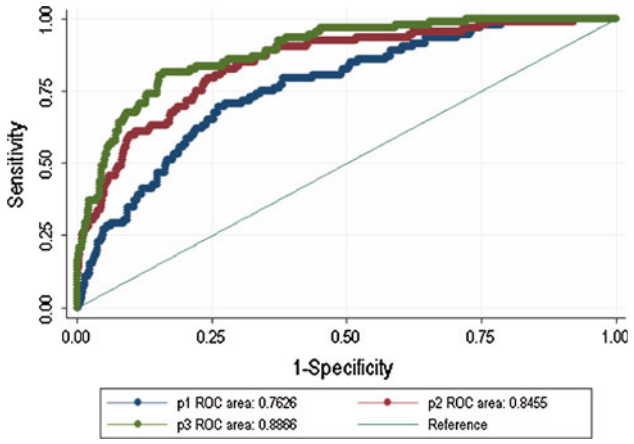

Fig. 1 ROC curves for KORA models 1–3 (ROC: receiver operating curve, KORA: Cooperative Health Research in the Region of Augsburg)

but is considerably improved by addition of blood parameters. Normally, AROCs do not decrease when a simpler model is nested within a more complex model in the same data set—although in rare cases of regression analyses, the more complex model might have a larger maximum likelihood, but a lower AROC.

CAVEATs

Using AROCs care should be taken in several regards.

1. *Model overfit* In logistic and Cox regression analyses, there is a risk of overfit, which has for consequence that AROCs might overestimate the diagnostic or prognostic ability of models. If we create a new KORA data set by randomly permuting the diabetes states (91 cases, 782 non-cases) among the KORA participants but keeping the original predictor variables fixed, we expect that, if we generate 1,000 such data sets, the AROCs of these 1,000 data sets should be 0.5 on average. This is because the models should not be able to predict randomly assigned diabetes states. However, what we observe is a mean AROC which is much larger than 0.5 (Table 2; Fig. 2). Mean AROCs ranged from 0.581 for model 1 to 0.602 for model 3. Generally, model overfit is less of a problem in large data sets. Overfit can be large in smaller data sets, in models with too many predictor degrees of freedom, and, in case of logistic regressions, when the number of events per predictor is very low (Babyak 2004). Thus, AROCs obtained from the original data tend to be too large. Therefore, any score has to be evaluated in a second sample which is different from the one which served to set up the model. If such an external validation is not possible, several ways of internal validation including cross-validation and bootstrapping have been suggested which are beyond the scope of this paper (Harrell et al. 1996; Muche et al. 2012).
2. *AROCs depend on the range of predictor variables in the study group* Comparing AROCs of scores developed in different study populations can be misleading if the study populations differ in the range of one or more predictors. To give an example, validating a diabetes prediction model in two external data sets, one with older subjects, and one including middle-aged and older subjects, the AROC is likely to be larger in the latter. Among the younger, there are many subjects with low age, low BMI, and normal blood pressure who do not develop diabetes and who are easily identifiable as non-cases—and this is reflected by a larger AROC.
3. *Little increase of AROC upon addition of strong predictors to a given model* Addition of strong predictors of a disease will often not lead to an increase in AROCs. For example, fasting insulin alone strongly predicts incidence of diabetes, and this is also true for KORA data (AROC = 0.697). But upon addition of insulin to model 2, there was only a small, insignificant increase in AROC from 0.844 to 0.845. This can partly be explained by the fact that predictors already in the model are often highly correlated to the additional predictors so that there is little additional information in the new variables. Moreover, only rank information is used for the calculation of AROCs, and, thus, improvement in disease prediction is only partly reflected by AROCs. If in Table 1, e.g., the predicted probability increased from 0.18 to 0.24 for subject 1, who is a case, and decreased from 0.16 to 0.13 for subject 4, who is a non-case, this would not change the AROC because the discordant pairs fulfilling the condition of $p(\text{case}) > p(\text{non-case})$ are still the same (cf. Table 3). In other words, AROC does not improve although the predictive ability of the model has increased as indicated by a larger predicted probability for a case, and a lower one for a non-case.

Table 2 Mean, minimum and maximum of the AROCs obtained when KORA models 1–3 were used to predict a randomly reassigned diabetes status (1,000 repetitions of the random reassignment for each model; $N = 873$; AROC: area under the receiver operating curve, KORA: Cooperative Health Research in the Region of Augsburg)

	Number of predictor variables	Mean of AROC values	Minimum of AROC values	Maximum of AROC values
KORA model 1	6	0.581	0.513	0.668
KORA model 2	9	0.597	0.527	0.679
KORA model 3	10	0.602	0.527	0.681

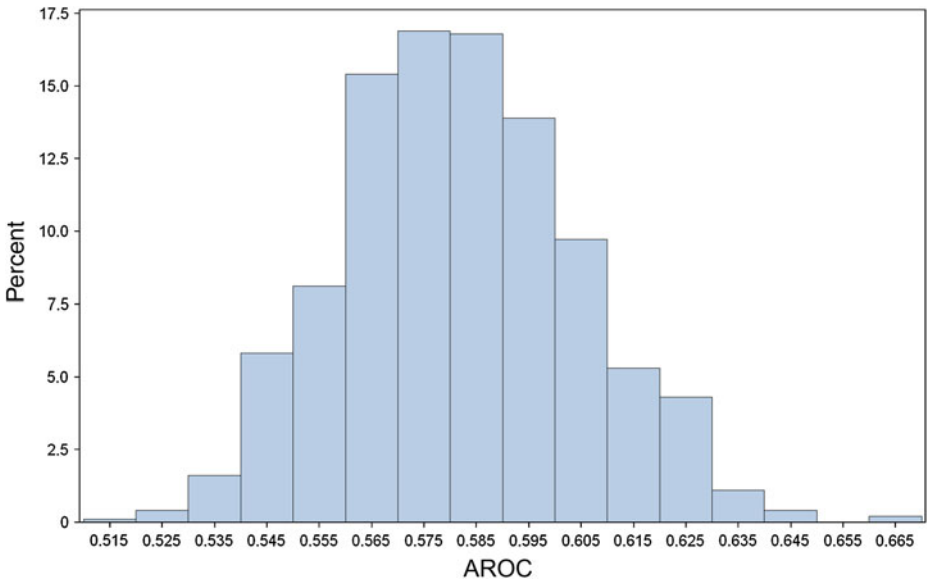

Fig. 2 Distribution of AROC values when KORA model 1 was used to predict a randomly reassigned diabetes status (1,000 repetitions of the random reassignment; AROC: area under the receiver operating curve, KORA: Cooperative Health Research in the Region of Augsburg)

Pencina et al. suggested the "integrated discrimination improvement" (IDI) which should be additionally used for comparison of models (Pencina et al. 2008).

From Tables 1 ("old model") and 3 ("new model"), the mean predicted probabilities for cases and non-cases are the following in the old and in the new model:

$p_{\text{cases,new}}$ = mean predicted probability for cases in the new model: 18 %

$p_{\text{cases,old}}$ = mean predicted probability for cases in the old model: 15 %

$p_{\text{non-cases,new}}$ = mean predicted probability for non-cases in the new model: 9 %

$p_{\text{non-cases,old}}$ = mean predicted probability for non-cases in the old model: 10 %

Obviously, the new model does better because estimated probabilities increase in cases and decrease in non-cases. IDI is calculated as follows:

Table 3 Calculation of AROC for a fictive study group of 5 subjects after changing predicted probabilities given for subjects 1 and 4 in Table 1 (AROC: area under the receiver operating curve)

Subject	Incident disease	Predicted probability (p)	Discordant pairs	p (case) > p (non-case)
1	Case	0.24	Subjects 1, 3	Yes
2	Case	0.12	Subjects 1, 4	Yes
3	Non-case	0.08	Subjects 1, 5	Yes
4	Non-case	0.13	Subjects 2, 3	Yes
5	Non-case	0.06	Subjects 2, 4	No
			Subjects 2, 5	Yes

Number of discordant pairs consistent of one case and one non-case = 6

Number of discordant pairs with p (case) > p (non-case) = 5

AROC = 5/6 = 0.83

$$\begin{aligned}\text{IDI} &= \left(p_{\text{cases,new}} - p_{\text{cases,old}}\right) - \left(p_{\text{non-cases,new}} - p_{\text{non-cases,old}}\right)\\ &= (18 - 15\%) - (9 - 10\%) = 4\%\end{aligned}$$

There is an improvement of 4 % in IDI. Tests of significance can also be done (Pencina et al. 2008). The "net reclassification improvement" (NRI) is analogous to IDI but NRI relies on self-selected categories of probabilities whereas IDI is a continuous measure.

To conclude, AROCs are a popular tool to discriminate subjects with a disease (or future disease) from subjects without the disease. But AROCs have drawbacks some of which were described here. From a clinical perspective, Greenland has brought forward further critical issues pointing out, e.g., that costs and benefits of additional predictors are not considered in using AROCs and IDIs, and that both statistics are cut-point free, and, thus, give weight to cut-points that will never be of relevance in clinical practice (Greenland 2008).

References

Babyak MA (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 66:411–421

Greenland S (2008) The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M.J. Pencina et al., statistics in medicine. Stat Med 27:199–206. doi:10.1002/sim.2929

Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361–387

Muche R, Ring C, Ziegler C Development and validation of logistic prognostic models by predefined SAS-macros. Available from

http://www.uni-ulm.de/med/med-biometrie/forschung/sas-makros-fuer-prognosemodellierung-mit-logistischer-regression.html. Assessed 20 June 2012

Pencina MJ, D′Agostino RB (2004) Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 23:2109–2123

Pencina MJ, D′Agostino RB Sr, D′Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172

Rathmann W, Kowall B, Heier M et al (2010) Prediction models for incident Type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. Diabet Med 27:1116–1123