



Statistical methods to assess the prognostic value of risk prediction rules in clinical research

Graziella D'Arrigo¹ · Mercedes Gori² · Annalisa Pitino² · Claudia Torino¹ · Stefanos Roumeliotis^{1,3} · Giovanni Tripepi¹

Received: 7 February 2020 / Accepted: 24 March 2020 / Published online: 2 April 2020
© Springer Nature Switzerland AG 2020

Abstract

Prognosis aims at estimating the future course of a given disease in probabilistic terms. As in diagnosis, where clinicians are interested in knowing the accuracy of a new test to identify patients affected by a given disease, in prognosis they wish to accurately identify patients at risk of a future event conditional to one or more prognostic factors. Thus, accurate risk predictions play a primary role in all fields of clinical medicine and in geriatrics as well because they can help clinicians to tailor the intensity of a treatment and to schedule clinical surveillance according to the risk of the concerned patient. Statistical methods able to evaluate the prognostic accuracy of a risk score demand the assessment of discrimination (the Harrell's C-index), calibration (Hosmer–May test) and risk reclassification abilities (IDI, an index of risk reclassification) of the same risk prediction rule whereas, in spite of the popular belief that traditional statistical techniques providing relative measures of effect (such as the hazard ratio derived by Cox regression analysis or the odds ratio obtained by logistic regression analysis) could be per se enough to assess the prognostic value of a biomarker or of a risk score. In this paper we provide a brief theoretical background of each statistical test and a practical approach to the issue. For didactic purposes, in the paper we also provide a dataset ($n = 40$) to allow the reader to train in the application of the proposed statistical methods.

Keywords Prognostic research · Discrimination · Calibration · Risk reclassification analysis

Introduction

Prognosis, together with diagnosis and treatment, is one of the three decisional processes of clinical medicine, and a fundamental element of public health as well. Prognosis of a given patient over a pre-defined time period is generally done by prognostic biomarkers and/or risk prediction rules, the latter being mathematical combinations of multiple

prognostic factors (i.e. biomarkers and/or other quantitative and qualitative variables such as age and gender), to be able to calculate the probability of a specific outcome on individual basis. Furthermore, biomarkers and risk prediction rules should be intended as a support in clinical medicine and not be used alone.

Before being adopted in daily clinical practice a candidate risk prediction rule needs to be carefully developed in a representative patients' cohort and externally validated in an independent series of patients affected by the same disease. Finally, a randomized controlled clinical trial would be ideally needed to demonstrate that the allocation of patients to specific treatments according to a given risk stratification tool leads to better outcomes as compared to those of patients allocated to a different risk stratification rule.

The evaluation of the prognostic accuracy of a risk score demands the assessment of discrimination, calibration and risk reclassification abilities of the same prediction rule. In this paper, using a simulation study, we describe how to formally calculate the Harrell's C-index (to assess discrimination) [1], the Hosmer–May Test (an index of calibration) [2] and the integrated discrimination improvements (IDI, an

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s40520-020-01542-y>) contains supplementary material, which is available to authorized users.

✉ Giovanni Tripepi
gtripepi@ifc.cnr.it

¹ Institute of Clinical Physiology (IFC-CNR), Clinical Epidemiology and Physiopathology of Renal Diseases and Hypertension of Reggio Calabria, Ospedali Riuniti, Via Vallone Petrarà snc, Reggio Calabria, Italy

² Institute of Clinical Physiology (IFC-CNR), Rome, Italy

³ Division of Nephrology and Hypertension, 1st Department of Internal Medicine, AHEPA Hospital, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

index of risk reclassification) [3] to evaluate the accuracy of a new prognostic scoring system (the PREDICT model) for disease-specific mortality in a cohort of patients aged > 60 years with localized prostate cancer [4]. For didactic purposes we use a small sample of patients ($n = 40$) but the reader should be aware that the validation of a risk prediction rule requires a sample size adequately predefined according to a specific power analysis [5]. In the same example, we also describe how to compare the accuracy of PREDICT with that provided by the score proposed by Ash et al. [6]. In the Online Appendix, we also provide the clinical dataset used to explain the above mentioned concepts and the STATA commands that interested readers can be easily trained in performing data analysis.

Indexes of prognostic accuracy

Discrimination measures how much a predictive model distinguishes patients with a end point from those without a given endpoint (for example, mortality). Discrimination, as assessed by the Harrell's C-index [1], ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination). The higher the Harrell's C-index, the higher the accuracy of the model in predicting the study outcome. The interpretation of the Harrell's C-index is conceptually similar to that of a ROC curve analysis [7]. Briefly, the Harrell's C-index represents the proportion of all possible pairs of patients in which the risk of death as estimated by the model agrees with the observed outcome. The concept underlying this index is that, under the assumption of random sampling, the predicted probabilities in patients who experience a given outcome should be systematically higher than those in patients who did not (tied observations are excluded). Calibration measures how much the outcome probability estimated by a predictive model matches the "real" probability of the same outcome. In calibration analysis, predicted and observed outcome probabilities are compared by the Hosmer–May test. When not significant, this test provides statistical evidence that predicted and observed outcome probabilities do not differ between them implying that the model is calibrated. To understand the difference between discrimination and calibration, we consider a hypothetical set of patients followed up for 20 years with an observed frequency of mortality of 40%. A prognostic model including clinical and laboratory variables which provides a 40% probability of death at 20 years in this patients' cohort is perfectly calibrated (estimated and observed death probabilities coincide). Now let us suppose that another prognostic model (i.e. a new risk prediction rule) provides a 25% chance of death in all survivors and a 26% chance in all non-survivors. It is clear that the threshold of 25% (as derived by the new prognostic model) accurately discriminates patients who die from those who

survive, but it is not calibrated because the estimated probability of mortality (25%) is much lower than the observed death probability (40%). Thus, the model is "discriminating" but underestimates the true risk of mortality. IDI is an index of risk reclassification which quantifies whether a new variable provides a clinically relevant improvement in prediction beyond and above that provided by a model based on a previous risk prediction rule and not including the same variable. The methodological background of IDI is that an adequate prognostic variable will increase the estimated risk for patients with the outcome of interest and will decrease the estimated risk for patients not experiencing the same event.

The prognostic accuracy of the PREDICT in the study cohort

PREDICT (pre treatment estimation of the risk of death in cancer of the prostate) is a risk prediction rule for disease-specific mortality developed by Kerkmeijer et al. [4] in 3383 patients with localized prostate cancer enrolled between 1989 and 2008 in four university hospitals in Netherlands and Belgium. Such a risk score was specifically developed to help clinicians in therapeutic decision-making. Kerkmeijer et al. identified, by multivariate Cox regression analysis, clinical T stage, biopsy grade, PSA and age as correlates of mortality. The PREDICT score provides five prognostic strata: low risk category (score < 3), intermediate risk category (score ranging from 3 to 7), moderate risk category (score ranging from 8 to 11), high risk category (score ranging from 12 to 16) and very high risk category (> 16). We tested the prognostic accuracy of PREDICT as categorical variable in a series of 40 patients aged > 60 years and affected by localized prostate cancer (see Table S1 in Online Appendix) and followed up for a median time of 15 years (range 1–26 years). Over a total person-time of 552 years, 24 patients died. According to PREDICT, at baseline 12 patients (30%) were classified at low risk, 9 patients (22.5%) at intermediate risk, 11 patients (27.5%) at moderate risk, and the remaining 8 patients (20%) at high risk. No patient was classified at very high risk. A Kaplan–Meier survival analysis (Fig. 1) shows that the cumulative survival reduces from the low risk category onwards and the difference in survival among the four groups is of high statistical significance (Log rank test = 29.3; $P < 0.001$). Furthermore, in a Cox model 1 unit increase in PREDICT categories entailed a 3.8 times increase (95% CI 2.2–6.5, $P < 0.001$) in the hazard ratio of death (Table 1 basic model). Although both the Kaplan–Meier and Cox regression analyses consistently indicate a significant and direct association between

Fig. 1 Kaplan–Meier survival curves according to PREDICT categories

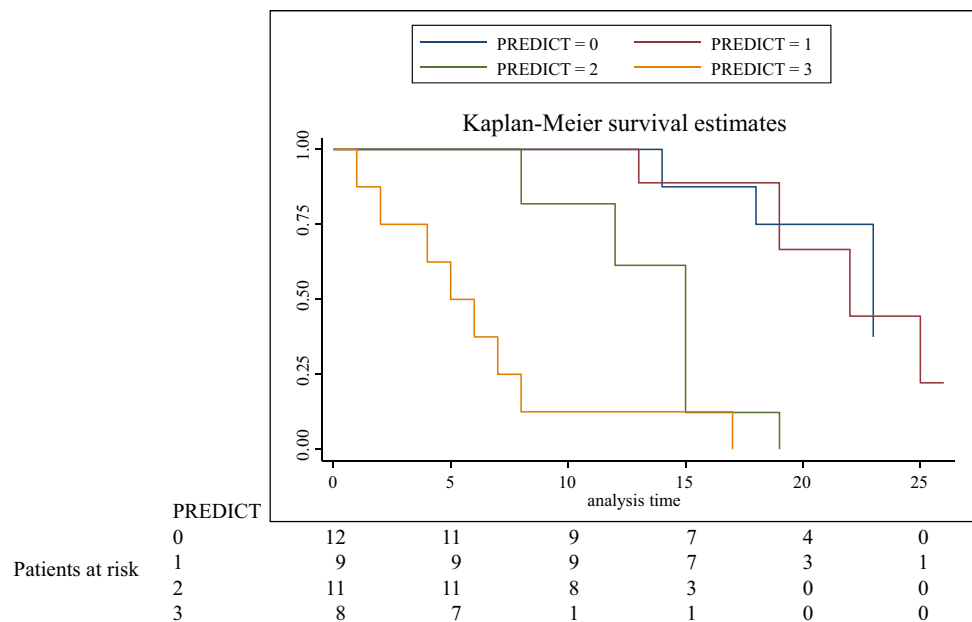


Table 1 Univariate and multiple Cox regression analyses for all-cause mortality of PREDICT either unadjusted (basic model) or adjusted (expanded model) for sedentary lifestyle

	Units of increase	Basic model (univariate Cox analysis)		Expanded model (multiple Cox analysis)	
		HR (95% CI)	<i>P</i>	HR (95% CI)	<i>P</i>
PREDICT ^a	1 unit	3.8 (2.2–6.5)	<0.001	4.0 (2.2–7.2)	<0.001
Sedentary lifestyle	0=no; 1=yes			2.6 (1.1–6.6)	0.04
		– 2 log likelihood (– 2LL)= – 2 × 56.8 = 113.6		– 2 log likelihood (– 2LL)= – 2 × 54.5 = 109	

The difference in the – 2LL (113–109=4.6, *P*=0.032) between the two models which just differ for one variable has a Chi squared distribution with one degree of freedom (1 *df*) and can be used to assess whether the addition of a variable (sedentary lifestyle in our case) to an existing model significantly improves the goodness of fit. Data are hazard ratios (HRs), 95% CIs and *P* values

^aThe hazard ratio of the PREDICT score for death corresponds to 1 unit increase of the same covariate. The prognosis of patients in each PREDICT category is reported graphically in Fig. 1. Given the fact that the two functional forms of PREDICT (continuous versus categorical) have identical prognostic performances (data not shown), we introduced this score as a continuous variable into the model

PREDICT and the hazard ratio of mortality, they do not provide quantitative information about the discriminatory power of this risk prediction rule for death in the study cohort. Therefore, to assess the discriminatory power of PREDICT we calculated the Harrell’s C-index and we found a value of 84.3%, a figure of high statistical significance (*P* < 0.001). Furthermore, the Hosmer–May test indicates that the prognostic estimates derived by PREDICT are calibrated because observed and predicted outcomes do not differ between them (Hosmer–May Wald Chi Squared = 2.28, *P* = 0.32). Overall, in the example including 40 patients aged > 60 years with localized prostate cancer, PREDICT accurately discriminates patients

who died from those who survived and provides calibrated risk estimates.

Does sedentary lifestyle increase the prognostic accuracy of PREDICT for mortality in the study cohort?

A fundamental issue in prognosis is whether a new variable increases the prognostic accuracy of a previous risk prediction rule in predicting the occurrence of a given outcome. In our study cohort, in a Cox regression model, sedentary lifestyle is significantly related to the risk of mortality (hazard ratio 2.8, 95% CI 1.2–6.5, *P* = 0.02). Given the fact that

PREDICT developed by Kerkmeijer et al. does not include sedentary lifestyle, now we want to calculate to what extent taking into consideration this variable could increase the prognostic power of PREDICT for mortality in the study cohort. From a practical point of view, we should compare two prognostic models: a model based on PREDICT (basic model) and a model including PREDICT and sedentary lifestyle (expanded model). The two models are “nested” because the basic model (based on PREDICT alone) is a simplified version of the expanded model (PREDICT+ sedentary lifestyle).

First, we build a Cox model including the two variables (see Table 1 expanded model) and we find that both PREDICT and sedentary lifestyle are significantly related to the hazard ratio of mortality. It implies that sedentary lifestyle is associated with the mortality risk in the study cohort independent of PREDICT and vice-versa. However, although statistically significant in a Cox model, we cannot take for granted that sedentary lifestyle adds meaningful prognostic information to PREDICT in predicting mortality in the study cohort.

As shown in Fig. 2, the inclusion of sedentary lifestyle to the model based on PREDICT marginally increased discrimination (Harrell’s C-index from 84.3 to 86.1%, + 1.8% gain) and the goodness of fit (Chi Square = 4.6, $P = 0.032$ with 1 degree of freedom) and provided a poor and not significant integrated discrimination improvement (IDI, + 4%, $P = 0.22$). The goodness of fit of a statistical model specifically describes how well it fits a set of observations. Measures of goodness of fit (such as the Chi Square) typically summarize the agreement between observed and expected values, i.e. values estimated by the model.

Calibration did not materially change (Hosmer–May Wald Chi Square = 1.86, $P = 0.39$, see Fig. 2) after the inclusion of sedentary lifestyle into the model. Overall, these results indicate that sedentary lifestyle does not materially increase the prognostic accuracy for mortality of PREDICT in the study cohort.

Comparison of the prognostic power of PREDICT and Ash classification for predicting mortality in the study cohort

In a Cox regression model, the Ash classification [6] resulted to be directly and significantly related to the incidence of mortality in the study cohort [HR (1 category increase): 3.1, 95% CI 1.5–6.4, $P = 0.002$]. The question now is whether the prognostic information provided by PREDICT are comparable to those provided by an existing risk prediction rule, the Ash classification [6].

When comparing the model based on PREDICT with that based on Ash classification we deal with two prognostic

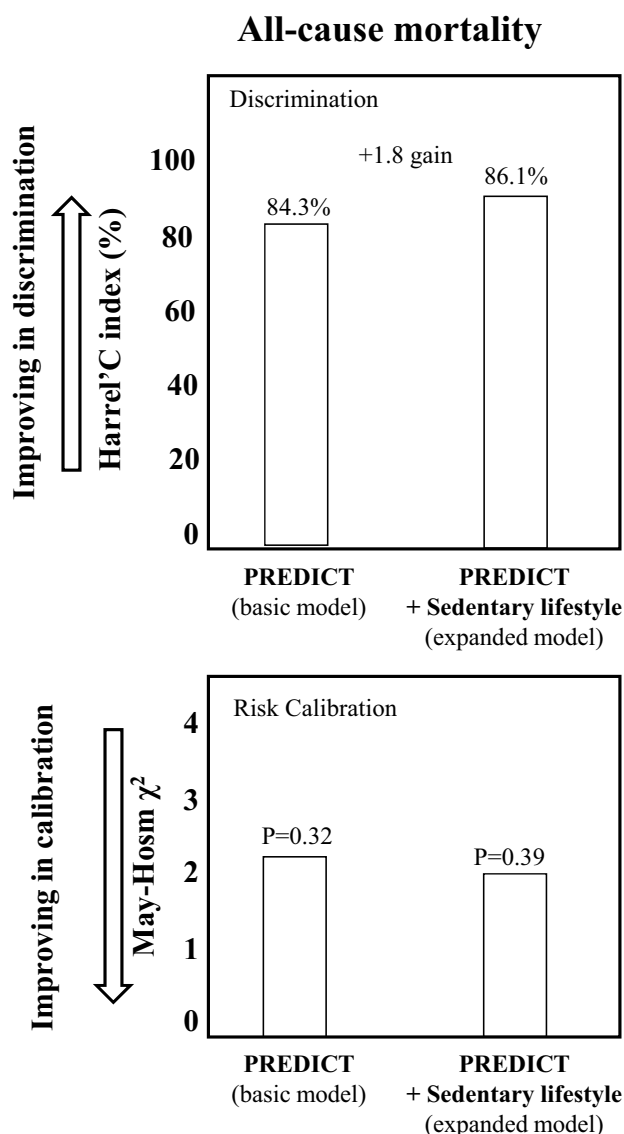


Fig. 2 Discrimination and calibration abilities of Cox models including PREDICT alone and PREDICT+ sedentary life style (see text for more details)

models which are not nested between them, i.e. the one cannot be considered a simplified version of the other, as in the previous example. The correct approach to be applied in this instance is the Akaike Information Criterion (AIC) which does not contemplate the nested assumption. Given the fact that our example is based on a small sample size, we need to correct the AIC by a correction factor (AICc) which takes into account the number of patients and the degrees of freedom. The lower the AIC or AICc, the higher the prognostic accuracy of a predictive model. Because AIC is a “criterion” and not a “test”, the AIC calculation does not provide a P value.

To assess which one among the set of the two candidate models (PREDICT and Ash classification based models)

is the best prognostic model, we analyze delta AICc and Akaike weights. To calculate delta AICc we preliminarily identify the model with the minimum AICc and then we subtract this value from the AICc values of the remaining models. The delta AICc is used to derive the Akaike weights, i.e. the probability that a given model is the best one among the full set of candidate models. In our example, the model based on PREDICT has >99% chances to be the best model. Accordingly, both discrimination (84.3%) and calibration (Hosmer–May Wald Chi Square = 2.28, $P = 0.32$) provided by PREDICT are better than those of the Ash classification (discrimination: 73%; calibration: Hosmer–May Wald Chi Square = 1.02, $P = 0.60$).

Conclusions

Risk prediction rules are tools to predict the future course of a disease and to help clinicians to provide a prognosis or to decide to start/intensify a given therapy on individual basis. To evaluate the implementation of risk prediction rules in clinical practice, they should be carefully developed and externally validated by appropriate statistical methods such as measures of discrimination, calibration and risk reclassification. Finally, researchers should provide robust evidence, in the framework of a randomized clinical trial, that the use of a new risk prediction rule precludes better clinical outcomes as compared to those achieved by the adoption of an existing risk prediction model or standard clinical care.

Author contributions GD, CT, SR, GT contributed to the study conception and design. The first draft of the manuscript was written by GD, CT, SR, GT. MG AP critically revised the manuscript and provided important intellectual contribution. All authors read and approved the final version of the manuscript.

Funding No Author has to disclose any conflict of interest that could have direct or potential influence or impart bias on the work.

Availability of data and material Data are provided in supplementary material.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Human and animal rights statement This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate For this type of study formal consent is not required.

Consent for publication Not applicable.

Code availability Not applicable.

References

1. Harrell FE Jr, 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
2. Crowson CS, Atkinson EJ, Therneau TM (2016) Assessing calibration of prognostic risk scores. *Stat Methods Med Res* 25:1692–1706
3. Pencina MJ, D’Agostino RB Jr, D’Agostino RB Jr et al (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
4. Kerkmeijer LG, Monninkhof EM, van Oort IM et al (2016) PREDICT: model for prediction of survival in localized prostate cancer. *World J Urol* 34:789–795
5. Schmoor C, Sauerbrei W, Schumacher M (2000) Sample size considerations for the evaluation of prognostic factors in survival analysis. *Stat Med* 19:441–452
6. Ash D, Flynn A, Battermann J et al (2000) ESTRO/EAU/EORTC recommendations on permanent seed implantation for localized prostate cancer. *Radiother Oncol* 57:315–321
7. Tripepi G, Heinze G, Jager KJ et al (2013) Risk prediction models. *Nephrol Dial Transplant* 28:1975–1980

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.