

PredictABEL: an R package for the assessment of risk prediction models

Suman Kundu, Yurii S. Aulchenko, Cornelia M. van Duijn, A. Cecile J.W. Janssens^{*}

Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands

^{*}Corresponding author

A. Cecile J.W. Janssens, Department of Epidemiology, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands

e-mail: a.janssens@erasmusmc.nl, Telephone : +31-10-7044214, Fax : +31-10-7044657

Abstract

The rapid identification of genetic markers for multifactorial diseases from genome-wide association studies is fuelling interest in investigating the predictive ability and health care utility of genetic risk models. Various measures are available for the assessment of risk prediction models, each addressing a different aspect of performance and utility. We developed PredictABEL, a package in R that covers descriptive tables, measures and figures that are used in the analysis of risk prediction studies such as measures of model fit, predictive ability and clinical utility , and risk distributions, calibration plot and the receiver operating characteristic (ROC) plot. Tables and figures are saved as separate files in a user-specified format, which include publication-quality EPS and TIFF formats. All figures are available in a ready-made layout, but they can be customized to the preferences of the user. The package has been developed for the analysis of genetic risk prediction studies, but can also be used for studies that only include non-genetic risk factors. PredictABEL is freely available at the websites of GenABEL (<http://www.genabel.org>) and CRAN (<http://cran.r-project.org/>).

Keywords: risk prediction, genetic, assessment, measures, software

Abbreviations

AUC Area Under the ROC Curve

IDI Integrated Discrimination Improvement

NRI Net Reclassification Improvement

ROC Receiver Operating Characteristic

Introduction

The rapid identification of genetic markers for multifactorial diseases from genome-wide association studies is fuelling interest in investigating the predictive ability and health care utility of genetic risk models. Genetic risk models are investigated for their potential to target diagnostic, preventive and therapeutic interventions for multifactorial diseases. Implementation of these models in health care requires a series of studies that encompass all phases of translational research [1-2], starting with a comprehensive evaluation of genetic risk prediction.

Various measures are available for the assessment of risk prediction models, each addressing a different aspect of performance and utility [3-4]. The GRIPS Statement recommends that transparent and complete reporting should provide a description of the risk factors and the risk model by reporting univariate and multivariate odds ratios for the predictors, present risk distributions for individuals with and without the outcome of interest, and report measures of model fit, predictive ability and others, if pertinent [5-6]. Examples of measures include the Hosmer-Lemeshow statistic [7] and Nagelkerke's R^2 [8] for model fit, the area under the receiver operating characteristic (ROC) curve (AUC) [9] and integrated discrimination improvement (IDI) [10] for predictive ability, and percentages of total reclassification [11] and net reclassification improvement (NRI) [10] for clinical utility.

Even though the assessment of risk prediction models is relatively standard, there is no single statistical package that would allow for the computation and production of all these measures and plots. Therefore, we developed PredictABEL, a freely available R package, which contains functions to obtain all descriptive tables, measures and plots that are used in genetic risk prediction studies.

Description of PredictABEL

The core part of PredictABEL comprises functions for the assessment of risk prediction models. The measures and plots covered in PredictABEL are listed in Table 1. Most functions can be applied to predicted risks, risk scores or any other continuous predictor variable, but some to predicted risks (probabilities) only. Predicted risks and genetic risk scores can be obtained using functions in the package, but they can be imported from other programs as well. The functions to obtain predicted risks using logistic regression analysis are specifically written for models that include genetic variables, eventually in addition to non-genetic factors, but they can also be applied to construct models based on non-genetic risk factors only. Genetic risk scores can be computed as unweighted and weighted risk scores, where weights are obtained from uploaded data or imported from meta-analyses, e.g., as beta coefficients.

The tables and plots generated using PredictABEL are saved as separate files in the working directory. Tables can be saved as Excel or tab-delimited text files and figures can be saved as publication-quality EPS or TIFF files or as JPEG files for insertion in manuscripts. All figures are available in a ready-made layout, but they can be customized to the journal style or preferences of the user. A hypothetical dataset and examples of use are included in the package to demonstrate all functions.

Example

The hypothetical dataset included in the package was reconstructed from an empirical study on age-related macular degeneration (AMD) [12], using a simulation method that has been described in detail elsewhere [13]. Based on published frequencies and

odds ratios of the genetic variants and non-genetic risk factors implicated in AMD and on published population disease risks, we created a dataset that contains genotype data and disease status for 10,000 individuals. Predicted risks were obtained using logistic regression analysis, for which the codes are provided in the package. Two risk models were constructed: a model based on non-genetic risk factors only and a model based on genetic and non-genetic predictors.

Figure 1 presents three examples of plots that are produced by PredictABEL. Figure 1a shows distributions of predicted risks based on genetic and non-genetic factors for individuals with and without AMD. The degree of overlap between the two histograms is indicative for the discriminative accuracy of the risk model. This discriminative accuracy is assessed by the AUC and visualized in a ROC plot. Figure 1b presents the ROC curves for the two risk models. The figure shows that the model with genetic factors had a higher AUC than the model without. Using the same function, the AUC values were quantified as 0.80 and 0.74. Finally, Figure 1c presents the calibration plot for the risk model based on the genetic and non-genetic variables as predictors, which shows how well predicted risks match observed risks. The calibration plot suggests that the model was well calibrated, which was supported by the non-significance of the Hosmer-Lemeshow test ($p = 0.65$).

Finally, Table 2 presents an example of the reclassification table and statistics that are produced by PredictABEL. The reclassification table presents the categorization into risk groups according to the initial and updated risk models. The table provides information about the total number of individuals that change between risk categories and about correct and incorrect reclassification. The percentage of total reclassification and NRI are calculated from the reclassification table. The table indicates that net 8.8% of the individuals without AMD and 9.6% of those with AMD

would be correctly reclassified when the clinical model was updated by the addition of genetic factors.

Conclusions

PredictABEL is a comprehensive software package, designed for the development and assessment of genetic risk prediction models. PredictABEL is a part of the GenABEL software suite for statistical genomics [14-15] and for that reason written in R to enable easy transfer of data from gene discovery to genetic prediction studies. A detailed manual is available that demonstrates and explains all the functions in the package. The manual is accessible for researchers who do not regularly use R software. The manual and the package are freely available from the GenABEL project website (<http://www.genabel.org>) and from CRAN (<http://cran.r-project.org/>).

The current version of PredictABEL (version 1.1) includes all basic descriptive tables, measures and plots that are used in the assessment of risk prediction models. Planned extensions of the package include other strategies to construct risk models, e.g., using Cox Proportional Hazards analysis for prospective data, and functions to construct simulated data for the evaluation of genetic risk models [13]. Furthermore, we will optimize the interconnectivity between PredictABEL and other packages in the GenABEL suite.

Where the GRIPS Statement aims to improve the transparency, quality and completeness of reporting [5-6], PredictABEL has similar goals for the assessment of genetic risk prediction studies. The collection of all measures and plots in a single, software package gives a comprehensive overview of the various measures that are available for the assessment of risk prediction studies. This overview emphasizes that different measures are available to answer different questions in the assessment of risk

models and facilitates the selection of the most appropriate measure for the question under study.

Acknowledgements

This work was supported by the Vidi grant from the Netherlands Organization for Scientific Research (NWO), the Young Investigator grant from the Erasmus University Medical Center Rotterdam and by the Center for Medical Systems Biology within the framework of the Netherlands Genomics Initiative.

References

1. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med*. 2007; 9: 665-74.
2. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009; 119: 2408-16.
3. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Arch Intern Med*. 2008; 168: 2304-10.
4. 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.
5. Janssens ACJW, Ioannidis JPA, van Duijn CM, Little J, Khoury MJ. Strengthening the reporting of genetic risk prediction studies: The GRIPS Statement. *Eur J Epidemiol*. 2011; 26: 000-000.
6. Janssens ACJW, Ioannidis JPA, Bedrosian S, Boffetta P, Dolan SM, Dowling N, Fortier I, Freedman AN, Grimshaw JM, Gulcher J, Gwinn M, Hlatky MA, Janes H, Kraft P, Melillo S, O'Donnell CJ, Pencina MJ, Ransohoff D, Schully SD, Seminara D, Winn DM, Wright CF, van Duijn CM, Little J, Khoury MJ. Strengthening the reporting of genetic risk prediction studies (GRIPS): elaboration and explanation. *Eur J Epidemiol*. 2011; 26: 000-000.
7. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997; 16: 965-80.
8. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika*. 1991; 78: 691-692.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143: 29-36.
10. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27: 157-72.
11. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115: 928-35.
12. Seddon JM, Reynolds R, Maller J, Fagerness J A, Daly MJ, Rosner B. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest Ophthalmol Vis Sci*. 2009; 50: 2044-53.
13. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med*. 2006; 8: 395-400.
14. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. 2007; 23: 1294-6.
15. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*. 2010; 11: 134.

Table 1 Measures and plots covered in PredictABEL (version 1.1)

	Measures and plots	Description
Description of the data	Frequencies	Allele and genotype frequencies by disease status
	Univariate odds ratios	Odds ratios per allele and per genotype
Description of the model	Multivariate odds ratios	Odds ratios adjusted for all predictors in the logistic regression model*
	Risk distribution	Histogram of predicted risks by disease status
	Predictiveness curve	Cumulative percentage of individuals against predicted risks
Overall model performance	Nagelkerke's R^2	Percentage of variance in the outcome explained by predictors in the logistic regression model*
	Brier score	Average squared difference between predicted risks and observed disease status
Calibration	Hosmer-Lemeshow statistic	Average difference between observed and predicted risks across subgroups.
	Calibration plot	Observed and predicted risks across subgroups
Discrimination	Receiver operating characteristic (ROC) curve	Sensitivity and specificity for all possible cut-off values of predicted risks
	Area under the ROC curve (AUC)	Measure of discriminative accuracy
	Discrimination box plot	Box plot of predicted risks by disease status
	Integrated discrimination improvement (IDI)	Comparison of mean difference in predicted risks of individuals with and without the disease between initial and updated model
Reclassification	Reclassification table	Number of individuals per risk category of the initial against the updated model by disease status
	Net reclassification improvement (NRI)	Net improvement in risk classification in individuals with and without the disease.

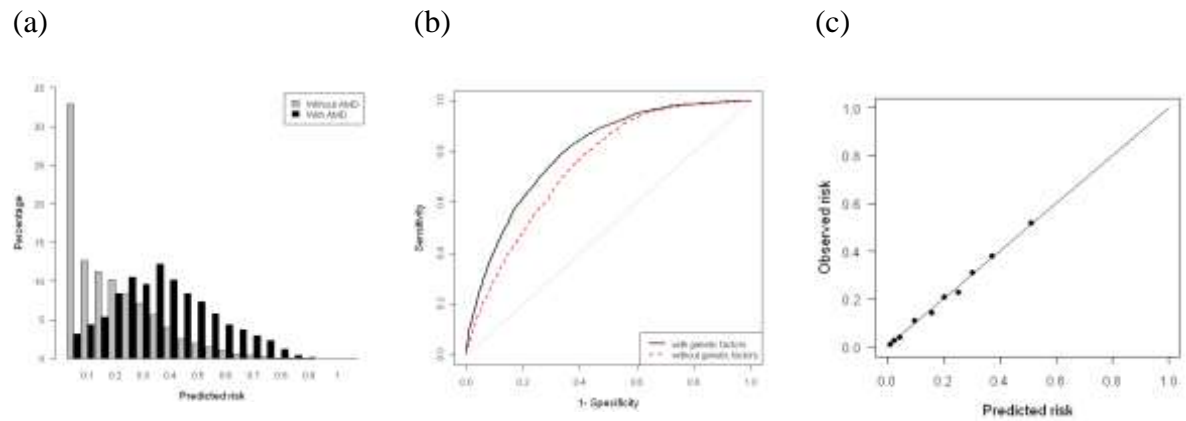
* These functions can only be used when the logistic regression model is constructed using the functions in PredictABEL.

Table 2 Reclassification table comparing clinical risk models without and with genetic factors.

Without genetic predictors	With genetic predictors			Reclassified		Net correctly reclassified
	<10%	10-35%	>35%	Increased risk	Decreased risk	
Individuals without AMD						
<5%	2187	459	0	816	1520	8.8%
10-35%	1225	2913	357			
>35%	15	280	577			
Individuals with AMD						
<5%	53	34	0	360	170	9.6%
10-35%	93	919	326			
>35%	1	76	485			
Net reclassification improvement 18.4% (95% CI 15.8 to 20.9); p<0.001						

Legend: AMD = age-related macular degeneration, CI = confidence interval. Values are numbers unless otherwise indicated. The cut-off risk thresholds chosen are for illustration purposes only and do not reflect clinically significant categories.

Figure 1 Example graphs produced by PredictABEL.



Legend: (a) Distributions of predicted risks in individuals with and without age-related macular degeneration (AMD); (b) ROC plot presenting risk models without and with genetic variants; and (c) Calibration plot comparing predicted risks with observed risks. Figure 1a and 1c present the risk model based on genetic and non-genetic risk factors.