The Methodology of Human Diseases Risk Prediction Tools

H. Mannan[*], R. Ahmed, M. Sanagou, S. Ivory and R. Wolfe

Department of Epidemiology & Preventive Medicine, Monash University, Melbourne, Australia

Abstract: Disease risk prediction tools are used for population screening and to guide clinical care. They identify which individuals have particularly elevated risk of disease. The development of a new risk prediction tool involves several methodological components including: selection of a general modelling framework and specific functional form for the new tool, making decisions about the inclusion of risk factors, dealing with missing data in those risk factors, and performing validation checks of a new tool's performance. There have been many methodological developments of relevance to these issues in recent years. Developments of importance for disease detection in humans were reviewed and their uptake in risk prediction tool development illustrated. This review leads to guidance on appropriate methodology for future risk prediction development activities.

Keywords: Disease risk prediction, missing data, model validation, model updating, model utility.

INTRODUCTION

The use of disease risk prediction tools is common, whether to identify high-risk individuals who should be treated or, in a population context, to whom screening or diagnostic tests should be applied. Risk prediction tools are also commonly used in clinical care settings for a variety of purposes including guiding clinical care of individual patients and performing risk-adjustment when comparing institutional performance.

The development of a risk prediction tool involves a considerable statistical challenge [1]. There are many facets to the statistical analysis and different methods exist for each facet. Usually no one method is correct and others incorrect, but statistical theory will often indicate that some methods are more desirable than others in particular circumstances or perhaps even in general terms. In this regard, a relatively recent account of methodology for risk prediction [1] provides excellent guidance, but there have been, even more recently, a number of important advances in understanding.

The key methodological components or steps of developing a new tool are as follows. First, an appropriate data set needs to be identified from which to develop a new risk prediction tool. Imperfections in the data, such as missing data, need to be identified and decisions made about how to proceed with analysis in light of the imperfections. Second, a general modelling framework and specific functional form for the new tool needs to be chosen. This should involve consideration of the eventual application of the new tool, e.g. is a simple point-scoring tool required for use in general practice? Third, variables that may contribute to the prediction of risk need to be selected from the data set and combined together in the chosen statistical model. Decisions need to be made as to whether inclusion of an extra variable improves in a meaningful way a simpler tool that excludes the variable, and these decisions have both a statistical aspect (fitting the observed data) and a utility of prediction aspect (weighing up the pros and cons of correct and incorrect identification of people with and without the disease of interest). Fourth, an independent data set needs to be identified to test the new tool's performance. Performance can be measured in different ways using different statistical summary measures, but two domains need to be considered: model discrimination, i.e. the ability to rank appropriately individuals at increasing risk, and model calibration, i.e. the ability to predict an accurate level of risk. The c statistic is a popular choice for measuring discrimination while the Hosmer-Lemeshow test is popular for measuring calibration. A wide variety of statistical methods have been employed in the development of risk prediction tools and there are examples of tools being developed with flaws in the application of statistical methodology; flaws that compromise the new tool's integrity.

This review aims to provide comprehensive coverage of the important areas of statistical methodology for the development of risk prediction tools with a focus on issues that have undergone important evolution in recent years. The review highlights how statistical methodology has been applied to risk prediction tools for disease in humans. Instances of inappropriate use of methodology are noted, together with guidance on more appropriate approaches.

[*]Address correspondence to this author at the Department of Epidemiology & Preventive Medicine, Monash University, Melbourne, Australia;
E-mail: hrmannan@gmail.com

E-ISSN: 1929-6029/13

CONSIDERATION OF THE CONTEXT

In practice it is often desirable to have a quick and simple point-score or chart-based risk prediction to facilitate efficient assessment of large numbers of people. For example a complex risk table was considered to have limited value in primary care for patient education and informed shared decision making [2]. Tool development methods generally have not reflected this desire, for example tools for disease presence have been developed using the relatively complex method of logistic regression and only afterwards the model coefficients simplified to point scores, sometimes erroneously [3]. If a simple prediction tool is desired then a development process that reflects this desire is entirely appropriate rather than coming as an afterthought.

Only recently have contextual considerations strongly entered into methodology for assessing the performance of a new tool. One such development is the incorporation of clinical utility in determining cut-offs for defining high risk from predicted risk scores [4]. Similarly, the use of a decision analytic framework to assess the suitability of a new tool for its intended screening purpose potentially offers new insights [5]. A review of risk scoring for cardiovascular disease was noteworthy for paying considerable attention to important practical issues such as the link between model form and tool acceptability [6].

MODELLING FRAMEWORK AND CHOICE OF FUNCTIONAL FORM

Many modelling frameworks have been proposed for risk prediction [1] and comparisons among different frameworks typically fail to find a preferred approach or disagree in their conclusions about which framework is to be preferred. Attention is limited here to commonly used regression-model based frameworks for disease. Discrimination and logistic regression models were original choices for developing risk prediction models for chronic disease onset and/or death using data from the Framingham cohort study, for example risk prediction models for cardiovascular (CVD) events [7, 8]. As an alternative, survival models have the advantage of incorporating time to event, rather than simply prevalence or incidence as in logistic and discriminant analysis models. Logistic regression is not recommended for long term predictions of chronic disease events such as coronary heart disease mortality since it has been found to have lower predictive ability than survival models for long term prediction of a chronic disease event [9]. It has been shown by a simulation study that ignoring time to event in studies with long follow-up might lead to biased estimates of measures of discrimination such as area under the receiver operating characteristic (ROC) curve, sensitivity and specificity [10]. Thus, for long term prediction of an event survival models are preferred over logistic regression since survival models analyze the time to occurrence of an event whereas logistic regression only analyzes the occurrence of the event in a given time frame and ignores the event’s exact timing. However, for short term prediction, in particular for predicting one year risk of CHD death, the performance of logistic regression has been found to be similar to survival models and is thus an acceptable alternative [9].

Possible inappropriate usage of logistic regression has been seen for screening of breast cancer patients through estimating a woman’s individual absolute risk of developing breast cancer over a 10-year period, as well as over her lifetime, based on environmental, reproductive, and hormonal factors [11]. Most other risk prediction models used for screening of breast cancer and other types of cancer have relied on the Cox proportional hazards model.

Among the survival models used for predicting CVD-related events, the non-proportional hazards Weibull model [12] is extremely flexible in terms of model assumptions. However, for convergence, this model requires the covariates to be mean centred. Proportional hazards models such as the Cox model and the standard Weibull model are preferable for applications requiring a relatively simple and interpretable survival model. A large number of chronic disease risk prediction equations are based on proportional hazards models [13-15] while a few have used Anderson’s non-proportional hazards Weibull model [18, 19]. The latter risk functions based on the Framingham heart study found that the proportional hazards assumption was strongly violated for long term (10 year) prediction of CHD events and for low risk individuals, indicating how sometimes compromise is necessary between model simplicity on the one hand and accurate representation of observed data on the other.

MISSING DATA

Missing data is a common problem in data sets used for the development of risk prediction models. Missing values of the outcome variable or of risk factors can occur for a variety of reasons such as non-response of the individual to some items on a questionnaire or failure to collect, or loss of, biological specimens before laboratory-based measurement. Since “missingness” in risk predictors is of great importance in prediction tool development, and offers scope for retrieval of information from study participants who have some missing and some observed data, hence the focus here is on this topic.

Complete Case Analysis

The simplest option to deal with missing data is to remove from analysis the individuals with missing risk factor values. This strategy is generally termed complete case analysis or listwise deletion. This approach continues to be used widely in developing risk prediction models. Recent examples include: development of a risk prediction model for chronic kidney disease screening by discarding individuals with missing serum creatinine measurements and other covariates [20, 21]; refinement of a risk prediction model for postpartum depression by incorporating season variation but discarding the many individuals who had missing data for either of the two key variables, education and income [22]; estimation of the diagnostic accuracy of stratus optical coherence tomography for glaucoma screening in high-risk populations using complete case analysis [23].

One general drawback of complete case analysis is possible over- or under-fit of the model (i.e. biased discrimination ability) arising from using a subset of data [24]. Also complete case analysis can lead to biased model coefficients when the individuals included in the analysis are systematically different from the individuals excluded from analysis because of their missing data [25]. Further, in complete case analysis standard errors of the model parameters will be unnecessarily large, i.e. the analysis makes inefficient use of the full sample by simply discarding anyone with missing values. A greater proportion of the sample with missing data corresponds to greater inefficiency of complete-case analysis. Comparison of different methods of dealing with missing data in developing a risk prediction model for a binary outcome has found that complete cases analysis “can lead to substantial bias and poor predictions which in practice could affect treatment strategies and decisions” [26].

Simple alternatives to complete case analysis are the missing indicator method [27] and single imputation. For the missing indicator method, an additional category is created for a categorical variable with missing data representing missingness in the variable while for a continuous variable it involves creating a new variable which recodes ***missing*** values to some common value. Although examples of use of the missing indicator method can be found [28, 29], and the method is conceptually efficient compared to complete case analysis , the method is susceptible to bias and generally it is not recommended [32]. Conceptually, single imputation is an attractive solution to the missing data problem as it fills in missing observations with plausible values, and examples of its use are not scarce [13, 33-35]. The simplest single imputation method is to substitute the missing value of a continuous predictor with the mean, or the most frequent category for a categorical predictor. However, such methods have serious limitations because they ignore potential correlation of the values of predictors among each other, and lead to an underestimation of variability in the predictor values among subjects. Regression imputation and conditional mean imputation are improved single imputation methods as they consider the correlation among predictors. In regression imputation a random draw from the distribution of predicted values is taken. In conditional mean imputation an imputation model is made to predict the missing values. Expected values are then imputed reflecting the correlations in the data. A simulation study [33] has shown that conditional mean imputation is preferable to complete case analysis. However, the method is only suitable if missingness depends on risk factors alone and not on the outcome of interest. The limitation of single imputation methods is that they only provide one of many possible imputed values to replace a missing value and, once imputed, there is no uncertainty associated with this single estimate. Thus single imputation will lead to over-estimation of the precision in parameter estimates, the opposite problem to complete case analysis. In this context the conservatism inherent in precision estimates from complete case analysis may seem an advantage.

Multiple Imputation

In multiple imputation missing data are replaced by imputing each missing value using some well defined imputation model. What differentiates the approach from single imputation is that the imputation process is repeated a number of times, say 10. Standard analysis is performed on each of the 10 data sets resulting in 10 sets of estimated model coefficients and associated standard errors. The 10 point estimates and their standard errors are then combined using Rubin's rules [37] to get a single estimated coefficient for each risk factor and its associated overall standard error. This overall standard error, unlike its counterpart in single imputation, takes into account the uncertainty involved in the process of imputation.

As elsewhere in medical research, multiple imputation has become increasingly popular in developing risk prediction models [1]. The increased uptake of multiple imputation is welcome but important issues have arisen which stem from the fact that a range of different approaches to multiple imputation are possible.

The simplest method of multiple imputations is known as “hotdecking” and there are a variety of hotdeck techniques available. The simplest hotdecking technique is to impute missing values for a predictor by using a sample drawn with replacement from the available values of that predictor [36]. A better approach matches, over all covariates, individuals with missing data to subsets of the remainder who have complete data; then imputation involves sampling with replacement from the relevant subset. Inclusion of the outcome of the risk prediction model in the imputation model is part of the hotdeck technique known as predictive mean matching [38].

Multiple imputation using normal-based methods [39] or a chained equation approach [40, 41] are popular alternatives to hotdeck techniques and a number of software options exist, e.g. the package MICE (Multiple Imputation by Chained Equations) in R [42], the mi system [43] and ice command in Stata [44-46].

To get unbiased estimates of model coefficients, the dependent variable of the risk prediction model must be included in the imputation model [39]. This somewhat counter-intuitive theory has been verified in a simulation study comparing complete cases analysis and multiple imputation [47]. More recently it has been found that using outcome as a predictor in the imputation model is preferable for all types of missing data mechanism [48]. Omitting outcome of the risk prediction model from the imputation model has led to erroneous omission of cholesterol from a tool for cardiovascular disease risk prediction [49] requiring a modified version to be published [16].

An important technical consideration in multiple imputation is the choice of the number of imputations (m). Rubin suggested that unless the rate of missingness is very high then m=5 is a reasonable choice. In developing risk prediction models, van Buuren [50] used m=3 for predictors with 20% missing values, Clark [51] used m=10 for predictors of ovarian cancer with about 40% missing values, and for predicting cardiovascular disease risk scores Hippisley-Cox [16, 49] used m=5. In general, the number of imputations required depends not only on the amount of missing data but also on the complexity of the risk prediction model and the data themselves. Royston *et al*. [46] discuss the impact of the number of imputations on the precision of estimates and suggest ways of determining the required number of imputations by evaluating the sampling error of the MI estimates. In a case study [52] 25 imputations were advocated to reduce the effect of sampling variability in the parameter estimates. However the conclusion is limited to the case study and was neither backed by any theoretical argument nor evaluated by extensive simulation.

Results from a multiple imputed data set are sensible only if the imputation model and the risk prediction model agree [53]. For example, if the risk prediction model is assessing interaction effect between two variables then the imputation model must also inclulde that interaction effect [54], or, if the risk prediction model is hierarchical then the imputation model has to share this structure [46]. Thus an important consideration in using multiple imputation effectively is to build a proper imputation model. A general guideline is: (i) structure the imputation model in a more general form so that imputed data can be used for a wider choice of risk prediction model [54], and (ii) use all variables including the outcome of the risk prediction model in the imputation model even if some variables have weak correlation with the variable to be imputed or have weak correlation with the missing data mechanism [52]. Following these guidelines makes the assumptions behind the missing data mechanism more plausible [46] and can increase the efficiency of the parameter estimates of the analyst’s model [55].

MODEL VALIDATION

For internal validation of new risk prediction models, the split-sample validation method has been used most frequently. Examples of its use can be found in models developed for predicting the risk of colorectal cancer [56] and CHD in primary care patients with chest pain [57]. However, there are several limitations of this method. First, if the samples are split fully at random as has been done in the aforementioned studies, then the distribution of the outcome and the predictors are likely to vary between the two sub-samples, and these distributions may be somewhat different in the original, combined, sample [1]. As sample size decreases the likelihood of these problems increases. As a result, inferences on model validation drawn from the testing (validation) data set may be misleading since the aim of using any internal validation method is to examine how valid a model is when applied to a particular sample. This limitation of the split-sample validation method can be overcome by stratifying the random sampling by outcome and relevant predictors [1].

There are other limitations of the split-sample validation method which cannot be avoided. For example, since only part of the data is used for model development this results in reduced precision in model parameter estimates compared with development based on the entire data. Also, since the validation sample is relatively small this can result in unreliable assessment of model performance [1]. The model may show, purely by chance, a poor performance in the validation random sample. The bias and instability of model results associated with the split-sample validation method reduce with increasingly large sample size [58].

An improvement on the split-sample validation method is offered by cross-validation which uses a larger part of the sample for model development compared with split-sample validation. In cross-validation the data is first randomly divided into deciles and the prediction model is developed based on nine of the deciles and tested in the remaining decile; when this process is repeated ten times with a different decile used for testing each time, it is known as ten-fold cross-validation. The performance of the model is estimated as the average of the ten testings. This method has been used, to a lesser extent than split-sample validation, for validating risk prediction models, for example as used in screening of breast cancer [28], infection in hospitalized patients with systemic lupus erythematosus [35] and pulmonary embolism in the emergency department [59]. Although the cross-validation method has an advantage over the split-sample method as it uses a larger part of the data for model development, to obtain truly stable results the method may need to be repeated many times, for example 400-times ten-fold cross-validation [1]. The most extreme cross-validation is to leave out each patient once, but with large numbers of patients this method is not efficient. The other problem with cross-validation is that it may not properly reflect all sources of model uncertainty as caused by using automated variable selection methods [1].

All limitations of the split-sample and cross-validation methods can, in principle, be overcome by the bootstrap validation method. For bootstrap validation the prediction model is evaluated both in the original sample and in the bootstrap samples. This method uses average validation across repeated bootstrap samples (all samples are of the same size as the original sample) drawn with replacement from the data set. The difference in performance between the original sample and the average performance across repeated bootstrap samples indicates the optimism, which is then subtracted from the apparent performance of the model in the original sample. This optimism-corrected performance estimate is rather stable, since samples of the same size as the original sample are used to develop the model as well as to test the model [1]. Despite the clear advantage of this method over the other commonly used validation methods, its application in development of a model for risk of having CHD in patients with chest pain [57] is a rare example of its use.

Discrimination and calibration are the two fundamental concepts associated with model validation. The discrimination of a model is its ability to correctly classify subjects into events and non-events. The area under the ROC curve or its equivalent, the c statistic, is the most commonly used measure of discrimination. Calibration, on the contrary, indicates how close the observed and predicted probabilities agree with each other. The closer the agreement, the better is the calibration. The most commonly used measure of calibration is the Hosmer-Lemeshow statistic [24]. The latter statistic has well-known limitations, for example, it assumes that the event is evenly distributed across the deciles yet this assumption is unlikely to hold when the event is uncommon. Alternatives to the Hosmer-Lemeshow statistic and ROC have been proposed and shown to have improved properties in specific situations [60] although it is yet to be ascertained whether other statistics are preferable in general.

Despite the importance of using both measures of discrimination and calibration for assessing the prediction ability of a model there has been a lack of consistency in their use. For example, in risk prediction models developed for predicting colorectal cancer [56] and CHD in primary care patients with chest pain [57], measures of both discrimination and calibration have been used while examples in breast cancer [28] and pulmonary embolism [61] have only used a measure of discrimination through the area under ROC curve.

MODEL UPDATING – INCLUSION OF EXTRA BIOMARKERS

For assessing the improvement in discrimination between two nested models, a test for difference in two correlated c statistics has been developed [62]. But, for models containing standard risk factors and possessing reasonably good discrimination, very large 'independent' associations of the new covariate with the outcome are required to result in a meaningfully larger c statistic [63-65]. Additionally, the c statistic has little or no direct relevance to clinical practice because it does not assist a doctor in making a treatment decision about an individual [5].

Reclassification tables are a way to express the results of prediction models in clinical terms [5] and together with the Net Reclassification Improvement, NRI, and Integrated Discrimination Improvement, IDI, provide valuable supplements to the c-statistic when comparing two nested models [66, 67]. The NRI and IDI attempt to quantify, in different ways, risk reclassification or the level of shift in the distribution of absolute risk after a new covariate is included in the model [68]. For calculating NRI, risk reclassification can be assessed by categorizing the predicted risk for an 'old' model and a 'new' model into clinically meaningful categories. For example, for evaluating improvement in CVD risk due to the inclusion of an additional covariate in the model the predicted risk may be categorised as <5%, 5% to <10%, 10% to <20%, and ≥20%, and used to examine how many individuals change from one category to another between the old and new models. For calculating IDI the mean absolute risk for the old model is subtracted from the mean absolute risk for the new model. For assessing the improvement in global fit between two nested models the conditional likelihood ratio test is the standard approach. For assessing calibration of a new model the Hosmer–Lemeshow statistic, and a modification for reclassification tables [66], are standard approaches.

These newer statistics have had rapid uptake for comparing models in risk prediction. For example, to assess whether the addition of breast density to a model that included current age, age at menarche, age at first live birth, family history of breast cancer, and number of breast biopsies, improved model discrimination for breast cancer risk among individuals without a prior history of breast cancer, the use of a test for difference in two correlated c statistics [69], was replaced with use of a reclassification table in later work by the same authors [28].

Of concern however, is that reclassification statistics were developed for comparing the predictive performances of two nested models, but despite this restriction, developments for predicting the risk of breast cancer [28], CVD [70] and survival among patients with coronary artery disease [71] have used these statistics for comparing two non-nested models. The major problem with non-nested comparisons is that the amount of reclassification does not represent differences in model performance [72]. Secondly, the proportion of events in the reclassification table's inner cells may be misleading as the cells may contain individuals selected on the basis of factors in both models, and not on the basis of an additional factor in the new model.

Another issue for model comparability has been the natural problem of having two separate concepts of model performance evaluation: discrimination and calibration. If the conclusion on model suitability differs between the two, then there can be some confusion. For example, in a comparison of the two prediction models, "Partin" and "Gallina," the Gallina model had better discrimination while the Partin model had better calibration [73]. The authors made the tentative conclusion that "limitations [of each model] need to be acknowledged and considered before their implementation into clinical practice." [5] It should be noted that discrimination is considered to be the primary metric in judging prediction accuracy since it cannot be improved by any adjustment [74, 75] unlike calibration which can be improved through recalibration without sacrificing discrimination [76]. While both will always be important to prediction, their relative importance can depend on the intended clinical application of the prediction rule. Calibration is the more important if the rule aims to estimate patient prognosis on the average, i.e. not necessarily at the level of the individual patient. On the contrary, discrimination is preferred if the use is to provide a prognostic classification for individuals [77].

The area under ROC curve, NRI and IDI have been extended to survival models as a function of the length of the period of risk to account for censoring [78]. Simulation studies have found the extended versions to have less bias, less variance and mean squared error than the traditional versions of these estimators [78]. Another limitation of summary model performance measures such as NRI, to be examined recently [79], is the obscuring of model features of importance to subsets of the population. New model performance measures have been developed for evaluating the precision gain within each of the risk groups of the reduced model.

COST OR UTILITY CONSIDERATIONS

There are some limitations of a reclassification table. First, although it aids in comparing the risk prediction performance of two nested models, it cannot be used to determine whether an individual model has better clinical value in terms of treating only those at high risk as identified by the model, compared to the strategy of treating all in the population. Also, reclassification tables do not directly help to determine which of two models has better clinical value, when one model reduces both true and false positives. For example, if a new model, having additional variable(s) compared to an old model, identifies fewer people subjected to intensive screening for a disease but detects fewer cases of disease early, then it is not immediately apparent whether a reduction in screening is worth the extra number of cases detected. To overcome this limitation of interpretability with reclassification tables, simple decision analytic approaches have been proposed for evaluating prediction models in terms of clinical benefits and costs. One such cost-benefit approach [5] uses different weighting schemes for true and false positives, to reflect whether delaying the diagnosis of a disease is more harmful than an unnecessary screening for the disease. Such approaches have great potential, recently summarised and illustrated [80], as they are capable of determining whether clinical implementation of prediction models is more beneficial than harmful in terms of costs.

DISCUSSION

This article has reviewed important statistical methodology issues for risk prediction and the main messages are now summarised.

Survival analysis regression models have an advantage over logistic regression in their incorporation of time to event resulting in better predictive ability for long term prediction of events.

In the presence of missing data on risk factors, multiple imputation should be used even though the theoretical and empirical basis to justify this position is not yet complete [1]. The alternatives have been shown to be worse in straightforward situations. Multiple imputation needs to be undertaken with care, for example inclusion of the outcome of the risk prediction model in the imputation model is important for valid prediction performance, and the number of imputations needs to be carefully chosen.

For internal validation of risk prediction models the bootstrapping validation method has certain advantages over split-sample validation and cross-validation methods and should be used more often for assessing internal validation of risk prediction models used for screening.

Measures of both discrimination and calibration should be used for assessing the prediction ability of a model but contextual considerations ultimately may be more relevant in deciding whether a model's performance is acceptable. Hence cost or utility considerations should be routinely assessed as part of the evaluation of the benefits of a new model.

Net reclassification improvement and integrated discrimination improvement are important statistics to be included in an evaluation of two competing and nested models. However they should not be used for non-nested models.

Methods for risk prediction have matured in the last twenty years. While advances continue to be made, for example recent extensions to measures of discrimination [81-84], there exists now a comprehensive methodology, well supported by statistical theory and simulation studies. Full adoption of these methods will lead to better quality risk prediction tools in the future.

COMPETING INTERESTS

None to declare.

REFERENCES

[1] Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer 2009.

[2] van Steenkiste B, van der Weijden T, Timmermans D, *et al.* Patients' ideas, fears and expectations of their coronary risk: barriers for primary prevention. Patient Educ Counsel 2004; 55(2): 301-307.
http://dx.doi.org/10.1016/j.pec.2003.11.005

[3] Moons KGM, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? [letter] J Clin Epidemiol 2002; 55(10): 1054-55. http://dx.doi.org/10.1016/S0895-4356(02)00453-5

[4] Baker SG, Cook NR, Vickers A, *et al.* Using relative utility curves to evaluate risk prediction. J Royal Stat Soc A 2009; 172(4): 729-48. http://dx.doi.org/10.1111/j.1467-985X.2009.00592.x

[5] Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: Towards a decision analytic framework. Sem Oncol 2010; 37(1): 31-38. http://dx.doi.org/10.1053/j.seminoncol.2009.12.004

[6] Beswick A, Brindle P. Risk scoring in the assessment of cardiovascular risk. Curr Opin Lipidol 2006; 17(4): 375-86. http://dx.doi.org/10.1097/01.mol.0000236362.56216.44

[7] Truett J, Kornfeld J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Diseases 1967; 20(7): 511-24. http://dx.doi.org/10.1016/0021-9681(67)90082-3

[8] Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967; 54(1): 167-79.

[9] Knuiman MW, Vu HTV, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. Eur J Cardio Prev Rehab 1997; 4(2): 127-34. http://dx.doi.org/10.1177/174182679700400209

[10] Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med 2006; 25(20): 3474-86. http://dx.doi.org/10.1002/sim.2299

[11] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004; 23(7): 1111-30. http://dx.doi.org/10.1002/sim.1668

[12] Anderson KM. A nonproportional hazards Weibull accelerated failure time regression model. Biometrics 1991; 47(1): 281-88. http://dx.doi.org/10.2307/2532512

[13] Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-Year follow-up of the Prospective Cardiovascular Münster (PROCAM) study. Circulation 2002; 105(3): 310-15. http://dx.doi.org/10.1161/hc0302.102575

[14] Conroy RM, Pyorala K, Fitzgerald AP. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003; 24(11): 987-1003. http://dx.doi.org/10.1016/S0195-668X(03)00114-3

[15] D'Agostino RB, Sr,Vasan RS, Pencina MJ, *et al.* General cardiovascular risk profile for use in primary care. Circulation 2008; 117(6): 743-53. http://dx.doi.org/10.1161/CIRCULATIONAHA.107.699579

[16] Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008; 336(7659): 1475-82. http://dx.doi.org/10.1136/bmj.39609.449676.25

[17] Hippisley-Cox J, Coupland C, Robson J, *et al.* Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. [electronic article] BMJ 2009; 338: b880. http://dx.doi.org/10.1136/bmj.b880

[18] Anderson KM, Odell, PM, Wilson PWF, *et al.* Cardiovascular disease risk profiles. Am Heart J 1990; 121(1) part 2: 293-98.

[19] Odell PM, Anderson KM, Kannel WB. New models for predicting cardiovascular events. J Clin Epidemiol 1994; 47(6): 583-92. http://dx.doi.org/10.1016/0895-4356(94)90206-2

[20] Bang H, Vupputuri S, Shoham DA, *et al.* Screening for Occult Renal Disease (SCORED) A Simple prediction model for chronic kidney disease. Arch Internal Med 2007; 167(4): 374-81. http://dx.doi.org/10.1001/archinte.167.4.374

[21] Bang H, Mazumdar M, Newman G, *et al.* Screening for kidney disease in vascular patients: SCreening for Occult REnal Disease (SCORED) experience. Nephrol Dial Transplant 2009; 24(8): 2452-57. http://dx.doi.org/10.1093/ndt/gfp124

[22] Panthangi V, West P, Savoy-Moore RT, *et al.* Is seasonal variation another risk factor for postpartum depression? J Am Board Family Med 2009; 22(5): 492-97. http://dx.doi.org/10.3122/jabfm.2009.05.080066

[23] Li G, Fansi AK, Boivin J-F, *et al.* Screening for glaucoma in high-risk populations using optical coherence tomography. Ophthalmology 2010; 117(3): 453-61. http://dx.doi.org/10.1016/j.ophtha.2009.07.033

[24] Harrell FE. Regression Modeling Strategies, New York: Springer-Verlag 2001.

[25] Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annual Rev Public Health 2004; 25: 99-117. http://dx.doi.org/10.1146/annurev.publhealth.25.102802.124410

[26] Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 2007; 16(3): 277-98. http://dx.doi.org/10.1177/0962280206074466

[27] Anderson AB, Basilevsky A, Hum DPJ. Missing Data. A Review of the Literature. In: Handbook of Survey Research, New York, NY: Academic Press 1983; pp. 415-492.

[28] Tice JA, Cummings SR, Smith-Bindman R, *et al.* Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med 2008; 148(5): 337-47. http://dx.doi.org/10.7326/0003-4819-148-5-200803040-00004

[29] Bach PB, Kattan MW, Thornquist MD, *et al.* Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003; 95(6): 470-78. http://dx.doi.org/10.1093/jnci/95.6.470

[30] Barlow WE, White E, Ballard-Barbash R, *et al.* Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 2006; 98(17): 1204-14. http://dx.doi.org/10.1093/jnci/djj331

[31] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142(12): 1255-64.

[32] Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991; 134(8): 895-907.

[33] van der Heijden GJ, Donders ART, Stijnene T, *et al.* Imputation of missing values is superior to complete case analysis and the missing indicator method in multivariable diagnostic research: A clinical example. J Clin Epidemiol 2006; 59(10): 1102-109. http://dx.doi.org/10.1016/j.jclinepi.2006.01.015

[34] Schemper M, Heinze G. Probability imputation revisited for prognostic factor studies. Stat Med 1997; 16(1): 73-80. http://dx.doi.org/10.1002/(SICI)1097-0258(19970115)16:1<73::AID-SIM472>3.0.CO;2-Z

[35] Yuhara T, Takemura H, Akama T, *et al.* Predicting infection in hospitalized patients with systemic lupus erythematosus. Intern Med 1996; 35(8): 629-36. http://dx.doi.org/10.2169/internalmedicine.35.629

[36] Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd ed. New York: Wiley 2002.

[37] Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. JASA 1986; 81(394): 366-74. http://dx.doi.org/10.1080/01621459.1986.10478280

[38] Little RJA. Missing-data adjustments in large surveys. J Bus Econ Stat 1988; 6(3): 287-96.

[39] Schafer JL. Analysis of Incomplete Multivariate Data, 1st ed. London: Chapman and Hall 1997. http://dx.doi.org/10.1201/9781439821862

[40] van Buuren S, Brand J, Groothuis-Oudshoorn C, *et al.* Fully conditional specification in multivariate imputation. J Stat Comp Simul 2006; 76(12): 1049-64. http://dx.doi.org/10.1080/10629360600810434

[41] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Meth Med Res 2007; 16(3): 219-42. http://dx.doi.org/10.1177/0962280206074463

[42] van Buuren S, Groothuis-Oudshoorn K. MICE 2.0: Multivariate Imputation by Chained Equations in R. J Stat Soft 2011; 45(3).

[43] Stata Corporation. Stata statistical software, release 11. Multiple Imputation Reference Manual. College Station, TX: StataCorp LP 2005.

[44] Royston P. Multiple imputation of missing values. The Stata J 2004; 4(3): 227-41.

[45] Royston P. Multiple imputation of missing values: Update of ice. The Stata J 2005; 5(4): 527-36.

[46] Royston P, Carlin JB, White IR. Multiple imputation of missing values: New features for mim. The Stata J 2009; 9(2): 252-64.

[47] Allison PD. Multiple imputation of missing data: A Cautionary Tale. Sociol Meth Res 2000; 28(3): 301-309. http://dx.doi.org/10.1177/0049124100028003003

[48] Moons KG, Donders RA, Stijnen T, *et al.* Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10): 1092-101. http://dx.doi.org/10.1016/j.jclinepi.2006.01.009

[49] Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007; 335(7611): 136-47. http://dx.doi.org/10.1136/bmj.39261.471806.55

[50] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999; 18(6): 681-94. http://dx.doi.org/10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

[51] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003; 56(1): 28-37. http://dx.doi.org/10.1016/S0895-4356(02)00539-5

[52] Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J, Tilling K. Strategies for Multiple Imputation in Longitudinal Studies. Am J Epidemiol 2010; 172: 478-87. http://dx.doi.org/10.1093/aje/kwq137

[53] Meng X-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. Stat Sci 1994; 9: 538-73.

[54] Schafer JL, Graham JW. Missing Data: Our View of the State of the Art. Psycho Meth 2002; 7: 147-77. http://dx.doi.org/10.1037/1082-989X.7.2.147

[55] Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stats Meth Med Res 2007; 16: 199-18. http://dx.doi.org/10.1177/0962280206075304

[56] Fazio VW, Paris PT, Remzi F, *et al.* Assessment of operative risk in colorectal cancer surgery: The Cleveland Clinic Foundation colorectal cancer model. Dis Colon Rectum 2004; 47(12): 2015-24. http://dx.doi.org/10.1007/s10350-004-0704-y

[57] Gencer B, Vaucher P, Herzig L, *et al.* Ruling out coronary heart disease in primary care patients with chest pain: a clinical prediction score [electronic article]. BMC Med 2010; 8: 9. http://dx.doi.org/10.1186/1741-7015-8-9

[58] Steyerberg EW, Harrell FE Jr., Borsboom GJ, *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54(8): 774-81. http://dx.doi.org/10.1016/S0895-4356(01)00341-9

[59] Gal GL, Righini M, Roy P-M, *et al.* Prediction of pulmonary embolism in the Emergency Department: The revised Geneva score. Ann Intern Med 2006; 144(3): 165-71. http://dx.doi.org/10.7326/0003-4819-144-3-200602070-00004

[60] Stallard N. Simple tests for the external validation of mortality prediction scores. Stat Med 2009; 28: 377-88. http://dx.doi.org/10.1002/sim.3393

[61] Aujesky D, Obrosky DS, Stone RA, *et al.* A prediction rule to identify low-risk patients with pulmonary embolism. Arch Intern Med 2006; 166(2): 169-75. http://dx.doi.org/10.1001/archinte.166.2.169

[62] Antolini L, Nam BH, Agostino RB. Inference on correlated discrimination measures in survival analysis: a nonparametric approach. Commun Stat Theory Methods 2004; 33(9): 2117-35. http://dx.doi.org/10.1081/STA-200026579

[63] Pepe MS, Janes H, Longton G, *et al.* Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004; 159(9): 882-90. http://dx.doi.org/10.1093/aje/kwh101

[64] Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. Arch Intern Med 2005; 165(21): 2454-56. http://dx.doi.org/10.1001/archinte.165.21.2454

[65] Ware JH. The limitations of risk factors as prognostic tools. NEJM 2006; 355(25): 2615-17. http://dx.doi.org/10.1056/NEJMp068249

[66] Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Annals Intern Med 2008; 149: 751-60. http://dx.doi.org/10.1002/sim.2929

[67] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, *et al.* Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27(2): 157-72.

[68] Cui J. Overview of risk prediction models in cardiovascular disease research. Annals Epidemiol 2009; 19(10): 711-17. http://dx.doi.org/10.1016/j.annepidem.2009.05.005

[69] Tice JA, Cummings SR, Ziv E, *et al.* Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. Breast Cancer Res Treat 2005; 94(2): 115-22. http://dx.doi.org/10.1007/s10549-005-5152-4

[70] van der Steeg WA, Boekholdt SM, Stein EA, *et al.* Role of the apolipoprotein B-apolipoprotein A-I ratio in cardiovascular risk assessment: a case-control analysis in EPIC-Norfolk. Ann Intern Med 2007; 146(9); 640-48. http://dx.doi.org/10.7326/0003-4819-146-9-200705010-00007

[71] Lauer MS, Pothier CE, Magid DJ, *et al.* An externally validated model for predicting long-term survival after exercise treadmill testing in patients with suspected coronary

artery disease and a normal electrocardiogram. Ann Intern Med 2007; 147(12): 821-28.
http://dx.doi.org/10.7326/0003-4819-147-12-200712180-00001

[72] Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med 2008; 149(10): 751-60.
http://dx.doi.org/10.7326/0003-4819-149-10-200811180-00009

[73] Zorn KC, Capitanio U, Jeldres C, *et al.* Multi-institutional external validation of seminal vesicle invasion nomograms: head-to-head comparison of Gallina nomogram versus 2007 Partin tables. Int J Radiat Oncol Biol Phys 2009; 73(5): 1461-67.
http://dx.doi.org/10.1016/j.ijrobp.2008.06.1913

[74] Harrell FE Jr, Lee, KL, Califf RM, *et al.* Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3(2): 143-52.
http://dx.doi.org/10.1002/sim.4780030207

[75] Lee KL, Pryor DB, Harrell FE Jr., *et al.* Predicting outcome in coronary disease: Statistical models versus expert clinicians. Am J Med 1986; 80(4): 553-60.
http://dx.doi.org/10.1016/0002-9343(86)90807-7

[76] D 'Agostino RB Sr, Grundy S, Sullivan LM, *et al.* Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA 2001; 286(2): 180-7.
http://dx.doi.org/10.1001/jama.286.2.180

[77] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130(6): 515-24.
http://dx.doi.org/10.7326/0003-4819-130-6-199903160-00016

[78] Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: Extension to survival Analysis. Stat Med 2011; 30(1): 22-38.
http://dx.doi.org/10.1002/sim.4026

[79] Whittemore AS. Evaluating health risk models. Stat Med 2011; 29(23): 2438-52.
http://dx.doi.org/10.1002/sim.3991

[80] van Calster B, Vickers AJ, Pencina MJ, *et al*. Evaluation of markers and risk prediction models: Overview of relationships between NRI and decision-analytic measures. Med Decis Making 2013; 33(4): 490-501.
http://dx.doi.org/10.1177/0272989X12470757

[81] Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004; 23: 723-48.
http://dx.doi.org/10.1002/sim.1621

[82] Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. Stat Med 2010; 29: 2508-20.
http://dx.doi.org/10.1002/sim.3994

[83] Pencina MJ, Steyerberg EW, D'Agostino RB. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011; 30(1): 11-21.
http://dx.doi.org/10.1002/sim.4085

[84] Pfeiffer RM. Extensions of criteria for evaluating risk prediction models for public health applications. Biostatistics 2013; 14(2): 366-381.
http://dx.doi.org/10.1093/biostatistics/kxs037

Received on 12-06-2013 Accepted on 22-07-2013 Published on 31-07-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.03.9