Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Goldfarb-Rumyantzev et al., J Biom Biostat 2016, 7:1
DOI: 10.4172/2155-6180.1000276

Research Article Open Access

Developing Prediction Models from Results of Regression Analysis: Woodpecker™ Technique

Goldfarb-Rumyantzev AS[1]*, Ning Dong[2], Sergei Krikov[3], Olga Efimova[4], Lev Barenbaum[5] and Shiva Gautam[6]

[1]Division of Nephrology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA
[2]Department of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Division of Medical Genetics, Department of Pediatrics, University of Utah, Salt Lake City, UT, USA
[4]VA salt Lake City,Department of Epidemiology, University of Utah, Salt Lake City, UT, USA
[5]Intelligent Decision, Inc., Izhevsk, Russia
[6]Department of Internal Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA

Abstract

Background: Developing medical prediction models remains time and labor consuming. We propose the approach where information collected from published epidemiological outcome studies is quickly converted into prediction models.

Methods: We used general expressions for regression models to derive prediction formulae defining the probability of the outcome and relative risk indicator. Risk indicator (R) is calculated as a linear combination of predictors multiplied by regression coefficients and then is placed on the scale of 0 to 10 for interpretability. Prediction expression for the probability (P) of the outcome is derived from general expression for logistic regression and proportional hazard models. The intercept is calculated based upon the outcome rate in the population and the risk indicator assigned to a subject representing mean characteristics of the population ($\hat{R}$). We also consider linear expression where probability of outcome is the product of risk indicator and the ratio of observed outcome rate and $\hat{R}$.

Results: These models were explored and compared in a numeric simulation exercise and also using real data obtained from NHANES dataset. All three expressions generate very similar predictions in the lower categories of risk indicator. In the groups with the higher value of risk indicator linear expression tends to predict lower probability than exponential expressions and also lower than observed.

Conclusions: We demonstrated simple technique (named Woodpecker™) that might allow deriving functional prediction model and risk stratification tool from the report of clinical outcome study using multivariate regression model.

Keywords: Prediction; Regression; Outcome; Woodpecker™; Epidemiology; Mathematical modeling.

Introduction

Large part of medical practice today is a set of wholesale treatments and diagnostics driven by guidelines and protocols. It is clear however, that given the individual differences, the approach to particular patient should be more personalized. Personalized medicine has to do with identifying the differences and tailoring treatment strategy to a particular individual. One way to address it is to use decision support systems based on prediction modeling of individual patient outcome. Generally, decision support systems, and in particular prediction models are inferior to a physician making a prediction [1]. As such, decision support systems should not compete with physicians making a decision, but instead contribute by providing quantified prediction data based on recent literature, especially in the context of uncertainty and limited time [2]. Demographics, comorbidities, genetic differences would make individuals behave and respond to the particular illness and medical interventions differently, so that the combined effect can be quantified by using prediction models with individual patient parameters. Techniques to quantify these differences in a form of generating prediction models of the individual outcome are underdeveloped. Part of the reason is that developing prediction model or risk stratification algorithm is a time and labor consuming process [3]. At the same time, the amount of quantified medical information being generated in the form of outcome clinical studies is massive and makes it difficult for professionals to follow the most recent trends. Furthermore, reports demonstrating statistical associations between particular variables and the outcome, effect of medications on the outcome do not become integrated part of the electronic tools available to practitioner, and these research data become largely underused. While the amount of data is increasing, the implementation of new knowledge remains slow and the time from "bench to bedside" is quite long. As a result scientific papers generally do not have much of an immediate practical impact. Moreover, clinical studies are limited to the population that was studied, inclusion and exclusion criteria, so extrapolating it to other, even similar populations might present a challenge [4]. Similar generalizability problem has to do with aging data, in other words, data sources used for prediction model development are at least several year old, so prospective use of these prediction models is a potential problem. Prediction model that can be easily adjusted to the new population would be of value as it will not be a subject to these limitations.

We propose that rather than developing prediction algorithms from raw data one might tap into published reports for information that has already been processed. Number of approaches was used to develop the practical risk score or risk-stratification tool. The most intuitive approach is a risk factor counting, where each risk factor

***Corresponding author:** Goldfarb-Rumyantzev AS, MD, PhD, Division of Nephrology, Beth Israel Deaconess Medical Center and Harvard Medical School, 185 Pilgrim Rd, FA-832, Boston, MA 02215, USA, Tel: 617-632-9880; Fax: 617-667-5276; E-mail: agoldfar@bidmc.harvard.edu

Received January 25, 2016; **Accepted** January 30, 2016; **Published** February 08, 2016

Citation: Goldfarb-Rumyantzev AS, Dong N, Krikov S, Efimova O, Barenbaum L, et al. (2016) Developing Prediction Models from Results of Regression Analysis: Woodpecker™ Technique. J Biom Biostat 7: 276. doi:10.4172/2155-6180.1000276

present in a subject adds another point to the risk score [5]. Somewhat more sophisticated approach is weighted addition of the risk factors. In that case points proportional to the regression coefficients are assigned to every risk factors and then added together [6,7] to comprise a final prediction score. The risk scores in the study population then examined and the range of scores is divided into several groups (e.g., low, intermediate, and high risk) [8].

Our vision is to use the information collected from the published epidemiological outcome studies and using mathematical expressions described below and named Woodpecker™ convert them into a line of prediction models and further into electronic tools that can be used in everyday practice. These tools can be integrated into electronic medical record systems, using existing patient information to populate input points and generate automated quantified predictions. In this report we describe the technique to derive prediction models and risk stratification tools from the results of multivariate analysis including those published in literature.

Methods

Assumptions

The performance of predictions drawn from statistical inference using Woodpecker™ technique is based on the validity of several assumptions. Some of the assumptions are the reflections of relationships between variables in the multivariate models, and in an ideal case scenario should be tested by the authors of the reports used to generate prediction models. Specifically, multivariate models are based on the assumptions of linearity, independence (lack of colinearity), and lack of interaction between the predictors. Additional distribution assumption regarding normality in the data is important for certain steps of the technique. These assumptions should be kept in mind by the reader interpreting results of the prediction modeling. Extrapolation of the model should also be mentioned. The need for extrapolation of the model arises from inevitable differences between the study population (defined as the population used to develop prediction model) and target population (the population used to validate the model or the population from which cases for prospective predictions are drawn). While the assumption being made that the populations are very similar, it might not always be the case. Therefore the model should ether be applied to the cases or the population that is closely resembles the study cohort (e.g., age, racial composition, comorbidities, etc.), or it should be able to be easily adjusted to a new population - the approach discussed below.

Calculating risk indicator

The risk for individual subject (R) for any of the models described below is calculated as a sum of the products of regression coefficients and the values of the predictors: $R = b_1x_1 + b_2x_2 + \ldots + b_ix_i$. For easier interpretation by the user the result will be related to a scale of 0 to 10 (R_{scaled}). We calculate lower end of the scale (L) using the values of predictors (x_i) corresponding to lowest risk (e.g., hypothetical patient 18 years old without comorbidities). Similarly to calculate the upper end of the scale (U) we use the data (x_i) of a hypothetical "high risk" patient (e.g., hypothetical patient 100 years old with multiple comorbidities). The calculated risk for individual patient on the scale of 0-10: $R_{scaled} = \frac{(10-0)\cdot(R-L)}{U-L} = \frac{10\cdot(R-L)}{U-L}$. It is important to note that this risk indicator is based on the relative scale (e.g., 0 to 10), in other words, the risk of 10 does not mean that the probability of the outcome is 100%, but rather higher than average risk of reaching the outcome.

Predicting absolute value of the outcome or the probability of the event

While calculating R is fairly intuitive, it is more complicated to convert it into actual value of probability of outcome. Conceptually, R is based purely on individual characteristics and determines individual risk relative to other members of the group. To convert this relative indicator into absolute probability of the outcome one has to use population characteristics (i.e., descriptive statistics of the predictors and overall outcome rate) and project individual risk on the population scale.

Continuous outcome: Continuous outcome is predicted by linear regression model: P(R) = a + R, where *a* is the intercept. If the intercept is not reported it can be derived from descriptive statistics: $a = y-(b_1 z_1 + b_2 z_2 + \cdots + b_i z_i)$ assuming that *y* is equal to the reported outcome rate and z - values of the predictors equal to their population mean (for continuous) or fraction of total (for categorical variables).

Categorical outcome: We considered three approaches to derive the probability of categorical outcome from R and population characteristics. Categorical outcome is predicted by either logistic regression or Cox regression model, in addition there is a simplified calculation based on assumed linear relationship between R and P.

Simplified linear calculation: To derive estimated probability of outcome the assumption is made of linear relationship between risk indicator (R) and the probability of the outcome (P) as follows: $\frac{R}{\hat{R}} = \frac{P}{r}$. The estimated probability $P(R) = \frac{R \bullet r}{\hat{R}}$, where r is the outcome rate in the target population and $\hat{R}$ is the target population mean R. The intercept in this linear expression is ignored. $\hat{R}$ is calculated as a risk indicator for a person with "average" characteristics: $= b_1z_1 + b_2z_2 + \ldots + b_iz_i$, where z_i - values of the predictors equal to their mean (for continuous) or fraction of total (for categorical).

Logistic regression: We start with general expression for logistic regression: $\ln\frac{P}{1-P} = a + R$ and therefore $P(R) = \frac{1}{1+\left(e^{(a+R)}\right)^{-1}} = \frac{e^{a+R}}{1+e^{a+R}}$ or $P(R) = \frac{1}{1+\left(e^{a}\cdot OR_1^{x1}\cdot OR_2^{x2}\cdot \ldots OR_i^{xi}\right)^{-1}}$ if odds ratios are used instead of regression coefficients. While R is based only on individual characteristics, the conversion of R into actual probability of outcome is defined by the population characteristics ($\hat{R}$ and r).

Therefore $a = ln\frac{r}{1-r} - \left(b_1z_1 + b_2z_2 + \ldots + b_iz_i\right) = ln\frac{r}{1-r} - \hat{R}$.

Proportional hazard model (Cox regression): Proportional hazard model is based on survival function S(t) = 1-F(t) (probability of surviving at a given point in time); f(t), the probability of death at a particular time point; $f(t) = \frac{dF(t)}{dt}$; and also hazard function h(t), which is a conditional probability of death $h(t) = h_0(t)\cdot\left(e^{\beta_1x_1+\beta_2x_2\ldots+\beta_ix_i}\right)$ where $h_0(t)$· is the baseline hazard (does not depend on independent variables, but only on time – it is equal to a hazard when all regression coefficients are equal to zero). Starting with $\frac{-d\left(logS(t)\right)}{dt} = h_0(t)\cdot\left(e^{\beta_1x_1+\beta_2x_2\ldots+\beta_ix_i}\right)$ we derive the expression for predicted probability of outcome $F(t) = P(R) = 1 - e^{-q_T\cdot\left(e^R\right)}$ or $F(t) = P(R) = 1 - e^{-q_T\cdot HR_1^{x1}\cdot HR_2^{x2}\cdot\ldots\cdot HR_i^{xi}}$ if hazard ratios were used instead of regression coefficients. To derive baseline hazard function we use the same technique as described above. Therefore, using the baseline statistics of the study and assuming that the outcome rate in the population (r) corresponds to mean risk ($\hat{R}$)

one can calculate baseline hazard as $q_T = -\frac{\ln(1-r)}{e^{(\beta_1 z_1+\beta_2 z 2\ldots+\beta_i z_i)}} = -\frac{\ln(1-r)}{e^{\hat{R}}}$.

Real data validation

To generate prediction model we used previously published report by McCullough et al. [9], with data collected from January 1, 2000, through December 31, 2003 of subjects from 42 National Kidney Foundation affiliates in 49 states. The outcome of the study was mortality, the rate of which in the study population was 0.0051.

For the target (validation) dataset we used the data of adult (≥18 years old) subjects from the National Health and Nutrition Examination Survey (NHANES) cohort which initially included 29,402 subjects enrolled between 1999 and 2004 with mortality information available through December 31, 2006. It is recognized that data collection for NHANES was based on a oversampling of children, females, older persons, black persons, and Mexican Americans. Files covering 1999-2000 (n = 9,965), 2001-2002 (n =11,039), and 2003-2004 (n = 10,122) were merged and variable name inconsistencies were corrected in the merged dataset. We deleted records missing estimated glomerular filtration rate (eGFR) value, mortality information or any of the five predictor values used in the model. Furthermore, we deleted records of those subjects who were followed for less than 2 years, so that the final dataset consisted of 12,515 subject records from 1999-2004.

Predictors and outcome definition

Variables used in prediction were defined in NHANES records as follows: presence of diabetes mellitus was based on questionnaire variable DIQ010 ("Doctor told you have diabetes") or DIQ050 ("Taking insulin now"); urine albumin-to-creatinine ratio was calculated from urine albumin concentration (variable URXUMASI) and urine creatinine concentration (URXUCR). Presence of CKD stage 3 or above was based on estimated glomerular filtration rate (eGFR) by MDRD expression [10]: eGFR = 175*(Serum creatinine$^{(-1.154)}$)*(Age in years$^{(-0.203)}$) * 0.742 if female * 1.212 if black. The stage of CKD has been assigned using NKF-KDOQI classification based on eGFR [11]. The presence of CVD was defined based upon at least one of the following four variables being positive: MCQ160C ("Ever told you had coronary heart disease") or MCQ160D ("Ever told you had angina/angina pectoris") or MCQ160E ("Ever told you had heart attack") or MCQ160F ("Ever told you had a stroke"). Age was not used as a predictor in this model.

Mortality data (from the date of survey participation through December 31, 2006) linked to the NHANES files were obtained from the CDC website.

Risk stratification tool

The calculation of risk indicator (R) is based on following predictors: male sex (HR 2.07), diabetes mellitus (HR 1.67), albuminuria>30 mg/dl (HR 1.77), prevalent chronic kidney disease of stage 3 and above (CKD; HR 1.98), prevalent cardiovascular disease (CVD; HR 3.02), and the combination of CKD and CVD (HR 3.8) [9].

To calculate the risk indicator, we first derived regression coefficients (as a natural logarithm of hazard ratios): 0.73, 0.51, 0.57, 0.68, 1.11, 1.34, respectively, for male sex, diabetes, albuminuria, CKD alone, CVD alone, and CKD with CVD the predictors listed above.

To establish a scale for the relative risk indicator we calculated upper (U) and lower (L) ends of the risk scale as follows: L,U = $b_1 x_1+b_2 x_2+\cdots+b_i$ xi using imaginable examples of the subject with the lowest possible and highest possible risk of the outcome. The former is a subject with no comorbidities (and therefore L = 0), and the latter is the one with all comorbidities used in the model (U = 3.15). To scale the risk indicator from 0 to 10 all coefficients were multiplied by 3.175 and as a result the risk indicator was calculated as a sum of following factors: for male gender - add 1.8, for diabetes add 1.3, for albuminuria>30 add 1.4, for CKD add 1.7, for CVD add 2.8, for the combination of CKD and CVD add 3.3 (last three categories are mutually exclusive).

Calculating the probability of event

The probability of the outcome (P) can be derived from the risk indicator using standard regression model expressions. The probability is the function of the risk indicator *R* (prior to scaling), the outcome rate in the target population (*r*; which is 0.0174 in this particular example) and a descriptive statistics of the predictors used in the model. We used descriptive statistics to estimate the mean *R* in the population ($\hat{R}$) as follows. We assumed that $\hat{R}$ is equal to the risk indicator calculated for imaginable individual with characteristics equal to the means of the target population (age = 48.9, male sex = 0.48, presence of diabetes = 0.095, etc. as shown on Table 1. Using this approach for our target population $\hat{R}$ = 0.6175.

We used three different expressions to calculate the predicted probability of the outcome derived from general expressions for regression models as described above.

- Linear expression $P(R) = \frac{R \cdot r}{\hat{R}} = R \cdot 0.028179$
- Exponential (logistic) $P(R) = \frac{1}{1+\left(e^{(a+R)}\right)^{-1}}$, where a is the intercept: $a = ln\frac{r}{1-r} - \hat{R} = -4.65121$
- Exponential (Cox) $P(R) = 1 - e^{-q_T \cdot (e^R)}$, where q_T is baseline hazard: $q_T = -\frac{\ln(1-r)}{e^{\hat{R}}} = 0.009466$

Validation and statistical analysis

Variables were summarized using means and standard deviations for continuous and percent of total for categorical variables. To quantify goodness of fit of our prediction models we used area under the ROC

	Mean (SD) or % of total for categorical variables	Range: minimum - maximum	95% CL for mean
Age (years)	48.9 (18.1)	20.0 - 84.9	48.5 - 49.2
Sex			
Male	48%		
Female	52%		
Race			
Non-Hispanic White	50.2%		
Non-Hispanic Black	18.9%		
Mexican American	23.4%		
Other Hispanic	4.6%		
Other	2.9%		
Presence of diabetes	9.5%		
Presence of CVD	6.8%		
Presence of CKD	5.4%		
Presence of CKD + CVD	2.8%		
Urine albumin-to-creatinine ratio (mg/g)	46.9 (400.0)	0.1 – 16636.4	39.9 – 53.9

Table 1: Baseline characteristics of the target population of 12,544 patients derived from NHANES [20].

curve. We divided the entire population based upon calculated risk into four categories: Low Risk (R=0-1.0), Low Intermediate Risk (*R*>1.0-3.0), High Intermediate Risk (*R*>3.0-6.0), and High Risk (*R*>6.0). The predicted outcome probability in each group was compared to calculated mortality rate.

Statistical analysis was performed using the SAS software version 9.3 (SAS Institute, Cary, North Carolina).

Results

We compared the prediction results of three models in a numeric simulation exercise. We used an imaginable population with the following characteristics similar to the group described in reference [9]: subjects are 48% males, 9.5% had diabetes mellitus, 11.7% with albuminuria of >30 mg/g, 5.4% with CKD, 7% with CVD, and 2.8% with a combination of CKD and CVD. Mortality rate (*r*) in this population is 0.0174 over two years of follow-up.

We used one linear and two exponential models as described above. Linear model is based upon P(R) = R·0.028179. The exponential expression based on logistic regression model is as follows: $(R)=\frac{1}{1+\left(e^{(a+R)}\right)^{-1}}$, where $a=ln\frac{r}{1-r}-\hat{R}=-4.65121$. Finally, the exponential expression based on Cox model is $P(R)=1-e^{-q_T\cdot\left(e^{R}\right)}$, where $q_T=-\frac{\ln(1-r)}{e^{\hat{R}}}=0.009466$.

Numeric simulation

We ran numeric simulation exercise using these formulae to calculate probability of 2 year mortality according to the risk indicator R. The results of this simulation are presented in Figure 1 (Panels A, B, and C). As demonstrated in Figure 1 Panel A, all three expressions generate very similar predictions in the lower 7 categories (R from 0 to 7).

In the groups with the higher value of R linear expression tends to predict lower probability than exponential expressions, while logistic regression and Cox regressions are very close to each other (Figure 1, Panel A). In a population with higher mortality (r = 0.075 in this example) but same R̂ exponential curves become flat and close to linear predictions (Figure 1, Panel B). The opposite trend is observed in a population with lower mortality rate (r = 0.005 in this example), where exponential curves are almost indistinguishable from each other, but separate farther from the linear predictions (Figure 1, Panel C).

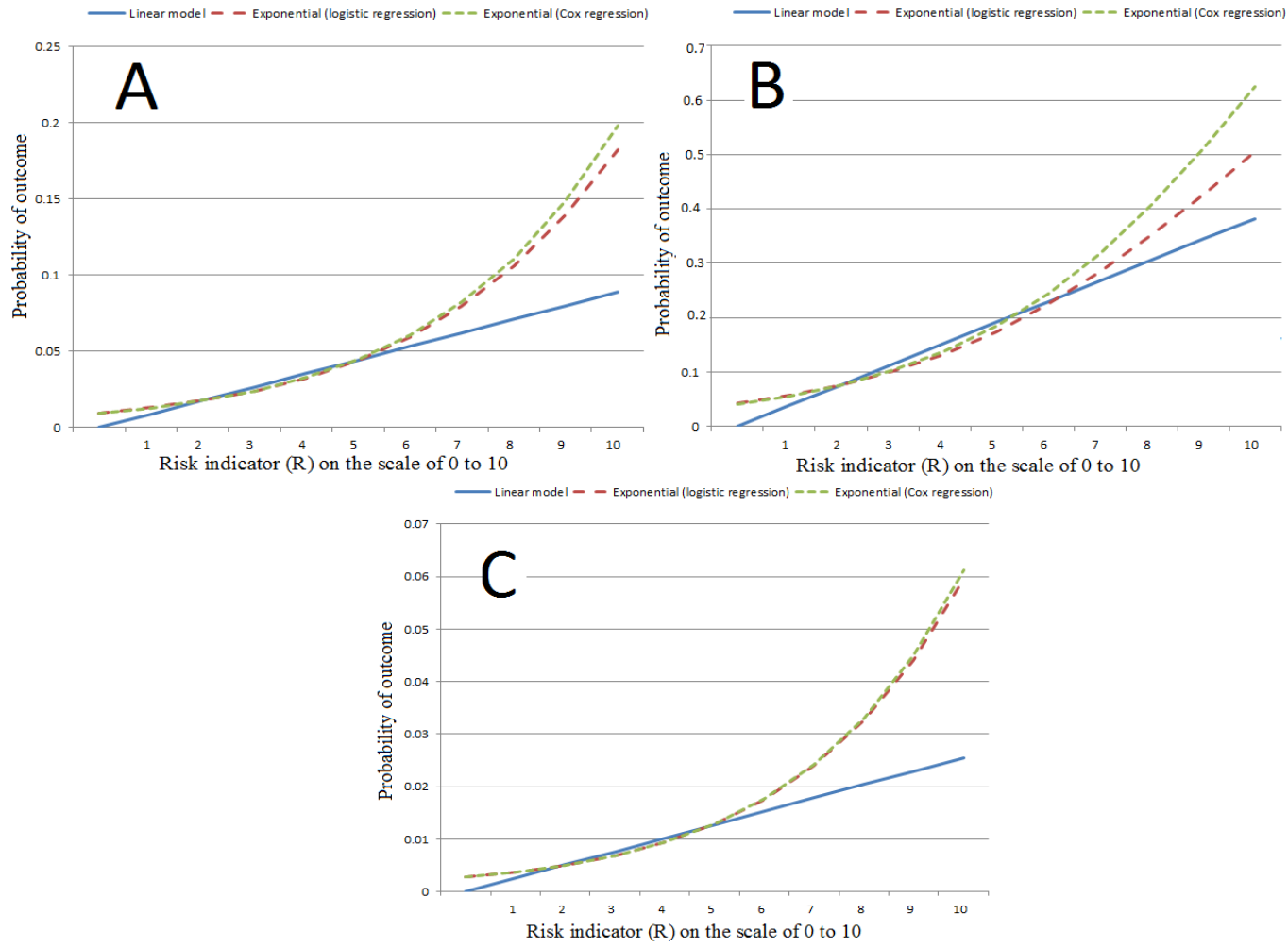

Figure 1: Results of simulation exercise for prediction of probability of outcome based on risk indicator values. Note that scaled risk indicator values are indicated on the figure, but unscaled values were used for calculation. Panel A, B, and C represent predictions based on different values of the mortality rate in the population. Panel A demonstrates the pattern associated with mortality rate of 0.0174, close to general US population. Panel B and Panel C demonstrate patterns associated with high (0.075) and low (0.005) mortality rates.

However, if *r* and $\hat{R}$ change in the same direction proportionally the pattern of predictions does not change.

Validation using NHANES dataset

After applying inclusion and exclusion criteria, the final target population of the study consisted of 12,515 subjects with mean age of 48.9 ± 18.1 years, 48% males, 50.2% non-Hispanic White, 18.9% non-Hispanic Black, and 23.4% Mexican American. Of these, 9.5% had diabetes mellitus, 11.7% had albuminuria of >30 mg/g, 6.8% had CVD, 5.4% had CKD, and 2.8% had a combination of CKD and a CVD. Patients were followed for an average of 57 ± 20 months. Of the target population, 1.7% died within the two years of follow-up. Other baseline characteristics of the target population are presented in Table 1.

After the risk indicator (*R*) was generated for each patient, we divided the patients into the four risk categories as described above. First, the model based on the above variables was used to predict the risk indicator *R* for two years mortality in the target dataset. The predictions were then compared to actual outcome and yielded the area under ROC curve of 0.74.

We stratified the patients into 4 categories based on the value of *R*: Low Risk (0-1.0), Low Intermediate Risk (>1.0-3.0), High Intermediate Risk (>3.0-6.0), and High Risk (>6.0). These categories consisted of 65%, 20%, 12% and 3% of the total population. The actual 2-year mortality rates in these categories are presented in Table 2. The predicted mortality rates using three prediction models are presented in Figure 2. Two exponential models generated very similar prediction results. Linear model was very similar to exponential models in the lower risk categories, but predicted lower probabilities in the risk categories with R>6. All three models were reasonably close to observe mortality rates in the groups (Figure 2).

Discussion

This report describes a method to derive prediction expressions from the results of multivariate analysis and its validation. Before the model can be used in practice thorough validation procedure is usually done (i.e., discrimination, calibration, and reclassification) in the population that was used for original study (internal validity) as well as in the separate group of subjects (external validity) [12,13]. Furthermore, clinical usefulness of the model can be evaluated in the implementation study after the model has been proven to be valid [14,15]. Internal validity of the models in terms of discrimination, calibration and reclassification as well as goodness of fit evaluation, which would normally be performed for prediction model, would require original raw data, and therefor is not available using this technique. However, we tested the technique using external validation approach. Similar approach has been successfully used in the past [16]. If the concept presented here is valid the prediction models may be generated and successfully used in the decision support systems in clinical practice. In that scenario these tools will be able to determine the individual risk of a particular event. The vision of the authors is that these tools will become an integral part of electronic medical records. The risk for individual subject will be determined as follows. After evaluating the applicability of the study to a particular individual using inclusion and exclusion criteria, the information pertinent to the model will be obtained (values of the predictors for a particular patient), and then the probability of the outcome and risk indicator will be determined.

Few practical points relevant to using this method should be emphasized. One of the issues is handling missing data. In case of calculating the probability of the outcome for a particular subject, the values for the missing data could be imputed with the population means. Alternatively, for predictor that is missing the value for a particular subject theoretically can be removed from the formula. Other important issue has to do with using R in calculation of probability of outcome. As indicated above the original calculated R is scaled to make it easier to interpret. It should be noted that either original R or R_{scaled} can be used for the expression assuming linear relationship between R and P ("simplified expression" above) as both R and $\hat{R}$ in the formula are scaled in the same linear way. However, for exponential expression the original R (as opposed to R_{scaled}) should be used.

As almost always the case in using prediction models in practice, there might be a substantial difference between the study population, where the model has been developed and the target population, where it is applied. The assumption that the model can be successfully extrapolated to a different population might not always be true as we demonstrated in the past [4,17]. To make some adjustment for these potential differences we propose to use the outcome rate (*r*) and baseline statistics of the target population in the calculation of the predicted probability of outcome as described above. That helps to avoid problems with over-fitting, aging data, and other challenges based of the difference between study and target population.

It should be noted that the Woodpecker™ approach has some important limitations. First, while we rely on the reports by other authors, the quality of the data and statistical analysis is always in question. However, provided our target reports are published in a quality peer-reviewed journals one can make an assumption that the results are sound. Another potential issue is that assumptions regarding the data might be violated. Indeed, different steps of Woodpecker™ technique are based on particular assumptions that cannot really be tested in this context. Specifically, in this approach we assume that the variables lack collinearity and have no interaction, which might

Risk category	n	Number of deaths	Mortality rate
Low (R = 0-1)	8143	24	0.29%
Low Intermediate (R>1-3)	2541	63	2.48%
High Intermediate (R>3-6)	1443	74	5.13%
High (R>6)	388	52	13.40%
Total	12515	213	1.70%

Table 2: Description of the 4 patient groups divided by risk indicator.

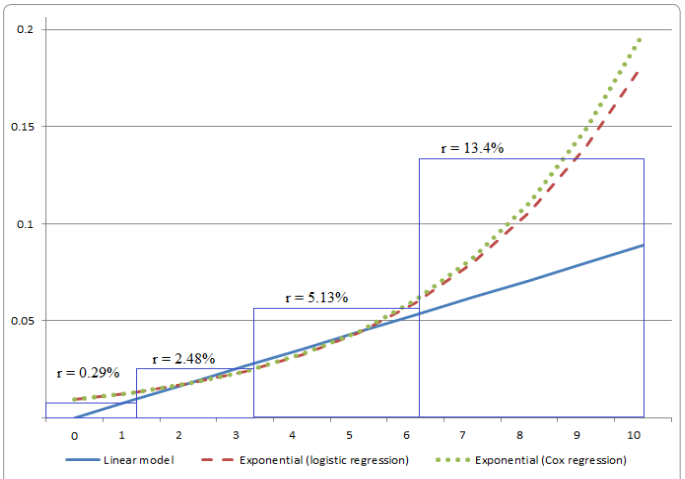

Figure 2: Comparison between three prediction models: linear (straight line), logistic regression (broken line) and Cox model (dotted line). Actual mortality rates in NHANES validation dataset are indicated by the blocks with the number on the top indicating percent mortality in each of the four groups.

not necessarily be true, or simply might not be tested by the authors of the original report. For some steps of this exercise we also assume normality in the data, which might also be violated.

One of the serious methodological challenges of this method is the difference between explanatory modeling (etiological modeling) and prediction modeling. While statistical tools might be the same, the study design aimed at causal explanation and theory building might be different from that aimed at prediction. This distinction in the goal of analysis may have impact on different steps of modelling process. Shmueli [18] elaborately described this difference on every step of statistical modeling process, i.e., goal definition, study design, data collection, data preparation and analysis, choice of variables and methods, model selection and validation, and reporting results. For example, in explanatory modeling the choice of independent variables would be driven by the selection of the primary variable of interest and confounding factors, while in prediction modeling every potential predictor might be included in the model. This conceptual difficulty might be minimized or avoided by careful selection of the studies to be used for prediction modeling. Specifically, hypothesis-driven studies that claim causality (rather than association) be less suitable. On the other hand, data-driven exploratory analyses even though not design for prediction might be more useful to be converted into prediction model. Therefore for our purpose of developing prediction algorithm the most adequate report is based on the study where multiple variables are included in the model and are reported in the paper. The selection of the variables in this "optimal" model is based on trying to choose the best predictors of the outcome [19,20].

Conclusion

In conclusion, we demonstrated simple technique (named Woodpecker™) that might allow deriving functional prediction model and risk stratification tool from the report of clinical outcome study using multivariate regression model.

Acknowledgements

None of the authors of the manuscript have any conflict of interest to declare. The study did not have any outside sponsor or funding agency. All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Farion KJ, Wilk S, Michalowski W, O'Sullivan D, Sayyad-Shirabad J (2013) Comparing Predictions Made by a Prediction Model, Clinical Score, and Physicians: Pediatric asthma exacerbations in the emergency department. Appl Clin Inform 4: 376-391.
2. Nagenthiraja K, Walcott BP, Hansen MB, Ostergaard L, Mouridsen K (2013) Automated decision-support system for prediction of treatment responders in acute ischemic stroke. Front Neurol 4: 140.
3. Goldfarb-Rumyantzev AS, Scandling JD, Pappas L, Smout RJ, Horn S (2003) Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset. Clin Transplant 17: 485-97.
4. Tang H, Goldfarb-Rumyantzev AS, Hunter C, Poynton MR, Tu M, et al. (2007) Validating prediction models of kidney transplant outcome using local data. AMIA Annu Symp Proc :1128.
5. Inouye SK, Zhang Y, Jones RN, Shi P, Cupples LA, et al. (2008) Risk factors for hospitalization among community-dwelling primary care older patients: development and validation of a predictive model. Med Care 46: 726-731.
6. Sullivan LM, Massaro JM, D'Agostino RB Sr (2004) Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med 23: 1631-1660.
7. Brueckmann B, Villa-Uribe JL, Bateman BT, Grosse-Sundrup M, Hess DR, et al. (2013) Development and validation of a score for prediction of postoperative respiratory complications. Anesthesiology 118: 1276-1285.
8. Rassi AJ, Rassi A, Little WC, Xavier SS, Rassi SG, et al. (2006) Development and validation of a risk score for predicting death in Chagas' heart disease. N Engl J Med 355: 799-808.
9. McCullough PA, Jurkovitz CT, Pergola PE, McGill JB, Brown WW, et al. (2007) Independent components of chronic kidney disease as a cardiovascular risk state: results from the Kidney Early Evaluation Program (KEEP). Arch Intern Med 167: 1122-1129.
10. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group," Ann Intern Med 130: 461-70.
11. Li C, Wen TF, Yan LN, Li B (2014) Risk factors for abdominal bleeding after living-donor liver transplant. Exp Clin Transplant 12: 424-428.
12. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128-138.
13. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, et al. (2012) Reporting and methods in clinical prediction research: a systematic review. PLoS Med 9: 1-12.
14. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, et al. (2013) Risk prediction models for patients with chronic kidney disease: a systematic review. Ann Intern Med 158: 596-603.
15. 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.
16. Keane WF, Zhang Z, Lyle PA, Cooper ME, de Zeeuw D, et al. (2006) Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. Clin J Am Soc Nephrol 1: 761-767.
17. Tang H, Hurdle JF, Poynton M, Hunter C, Tu M, et al. (2011) Validating prediction models of kidney transplant outcome using single center data. ASAIO J 57: 206-212.
18. Shmueli G (2010) To Explain or to Predict?. Statistical Science :289-310.
19. Foley RN, Murray AM, Li S, Herzog CA, McBean AM, et al. (2005) Chronic kidney disease and the risk for cardiovascular disease, renal replacement, and death in the United States Medicare population, 1998 to 1999. J Am Soc Nephrol 16: 489-495.
20. National Health and Nutrition Examination Survey. http://www.cdc.gov/nchs/nhanes.htm; last accessed on February 1, 2016.