Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Goldfarb-Rumyantzev and Dong, J Biom Biostat 2016, 7:1
DOI: 10.4172/2155-6180.1000275
Research Article
Open Access

Combining Prediction Models in a Linear Way: Results of Numeric Simulation

Goldfarb-Rumyantzev AS[1]* and Ning Dong[2]

[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

Abstract

Background: Using standard expressions for logistic regression and proportional hazard models and data from published outcome studies might allow generating prediction models and risk stratification tools in a more streamline fashion. However it might require combining the models, adding or removing predictors. The feasibility of this approach has been examined here.

Methods: The outcome of this simulation study is mortality. The simulation exercise was based on the imaginary population of 20,000 subjects whose mortality was completely determined by five variables in the specified logistic regression model. In the first simulation exercise using "full model", we evaluated the option of combining the results of two separate studies (studies A and B) each based on subset of the population. In the second simulation exercise studies A and B were based on limited number of predictors. Each simulation was repeated 50 times.

Results: Both simulation exercises demonstrated the robustness of the model and feasibility of adding or removing predictors to/from the model. We also compared the results of linear model to the more complex exponential model using all five predictors. In subjects with lower risk indicator the outcome of linear model is similar to the outcome of the logistic regression model and to the true outcome rate, however it underestimates the risk in the high-risk groups. On the other hand, logistic regression model is accurate compared to actual outcomes. This confirms our hypothesis that dropping or adding variables should not distort the prediction in any noticeable way.

Conclusions: Simple linear combination of prediction models, adding or removing predictors do not cause distortion of the model and predictions remain robust. Prediction of linear model is similar to exponential model, except the former underestimate the outcome in the high risk groups.

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

Introduction

Prediction models in medicine are able to generate quantified measure of individual outcome and are an important tool in medical decision making, especially in the context of personalized medicine [1]. While using specific risk factors in decision making is common [2,3], actual quantified predictions are still lacking. That has to do with the fact that developing prediction models is time and labor consuming. However using standard expressions for logistic regression and proportional hazard models and data from published outcome studies might allow to generate prediction models and risk stratification tools in a more streamline fashion. We previously validated this approach in the numeric simulation exercise and also using actual data from the subjects participating in the NHANES study. These results are reported in this issue of the journal. However, few issues with this technique need more focused attention as described below.

It is noted that while statistical tools might be the same, there is a difference between explanatory modeling (etiological modeling, where the study design aimed at causal explanation) and prediction modeling [4]. For example, in explanatory approach, i.e., hypothesis-driven research, the choice of independent variables would be driven by the selection of the primary variable of interest and confounding factors. These models sometimes do not include the variables that would be important for prediction, but are not pertinent to the hypothesis being tested. On the other hand, some of the potential confounders included in the model would not necessarily make a good predictors of outcome. That creates two issues: (1) some of the variables will need to be removed from the original model; and (2) other variables, deemed important for prediction have to be added to the model. Below we illustrated this issue in more details and evaluated it by performing numeric simulation. Authors are aware that the entire field of meta-analysis and meta-regression exists [5,6] and addresses similar issue of combining of results of separate outcome studies; however, here we evaluated a very simple approach of removing or adding variables from/to the model without doing any other adjustments to the prediction expression.

First, we hypothesized that simple linear expression might perform very similar to the more complicated exponential expressions in low-risk subjects. Second, we propose that multivariate models coming from different sources but based on the same or similar populations might be combined in a very simple way without the risk of distorting the outcome. Similarly, adding to or subtracting variables from the model should not affect the performance of the model in a major way other than changing the accuracy of prediction to a reasonable degree. Theoretically, since our prediction expressions include the risk indicator and the term for intercept, then no additional adjustment is necessary if variables are being added or removed. As we describe below, the

***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 (2016) Combining Prediction Models in a Linear Way: Results of Numeric Simulation. J Biom Biostat 7: 275. doi:10.4172/2155-6180.1000275

intercept is based upon average risk indicator in the population which theoretically should absorb the differences after variables were added to or removed from the model.

Methods

Brief description of the Woodpecker™ technique

In developing this approach we used general expression for regression models to derive prediction formulae defining the probability of the outcome and relative risk indicator. Risk indicator (R) is calculated as $R=b_1 x_1 + b_2 x_2 + \cdots + b_i x_i$ and then placed on the scale of 0 to 10 for interpretability. Prediction expression for the probability (P) of the outcome using logistic regression $P(R)=\frac{1}{1+\left(e^{a+R}\right)^{-1}}$, where intercept $a=ln\frac{r}{1-r}-\hat{R};$ r is the outcome rate in the population and $\hat{R}$ is the value of risk indicator for the "average" person in the population: $\hat{R}=b_1 z_1 + b_2 z_2 + \ldots + b_i z_i$; where z_i - values of the predictors equal to their mean (for continuous) or fraction of total (for categorical). In the population For proportional hazard model the probability of outcome is calculated as follows: $P(R)=1-e^{-q_T\cdot(R)}$, where $q_T=-\frac{\ln(1-r)}{e^{\hat{R}}}$. Finally, we also evaluated simplified linear expression $P(R)=\frac{R\cdot r}{\hat{R}}$. Except for linear expression, the value of R used in calculation is the one prior to scaling.

These models were previously validated in a numeric simulation and also using the data from NHANES study.

Theoretical background

Intuitively, based on the expression for probability derived from regression $P(R)=f(a+b_1x_1+b_2x_2+\ldots+b_ix_i)=f(R+a)$; it seems that if the model is changed by removing one or more of the predictors or adding additional ones from a separate model it should invariably distort the value of the probability. However, if some mechanism of correction is introduced in the formula, the issue might be resolved and make the method more practical and robust. For example, if $R/\hat{R}$ is used in the calculation, adding or subtracting the variables as well as scaling will not matter since both R and $\hat{R}$ change proportionally with adding or subtracting variables and with scaling. Alternatively one might rely on the fact that the prediction formula (e.g., for logistic regression $P(R)=\frac{1}{1+\left(e^{a+R}\right)^{-1}}$ includes the intercept $(a=ln\frac{r}{1-r}-\hat{R})$, which is based upon $\hat{R}$ and this fact might provide necessary adjustment to the formula (the change in probability will be correctly accounted for by calculating the intercept, as long as the predictors are independent and there is no interaction). In other words, the question becomes if $a+R=ln\frac{r}{1-r}-\hat{R}+R$ (or specifically($R-\hat{R}$)) will remain reasonably stable with adding/subtracting predictors from the model. It is not a straight forward answer since while regression coefficients are the same for R and $\hat{R}$ calculation, the values of the predictors (x) are different (i.e., in the former case these are values for individual subjects, in the latter case – average values in the population). Therefore with adding or removing predictors ($R-\hat{R}$) will not remain exactly the same, but it might remain stable enough to preserve the robustness of the probability calculation.

Numeric simulation

The supposed outcome of this simulation study is mortality. The simulation exercise was based on the imaginary population of 20,000 subjects whose mortality was completely determined by five variables in the logistic regression model $logit(P)=a+b_1x_1+b_2x_2+b3x3+b4x4+b5x5$. Suppose two studies were performed (study A and study B), each randomly sampled 4000 subjects from the population and evaluated different predictors using logistic regression model.

We performed two separate simulation exercises

(1) In the first simulation exercise using "full model", we evaluated the option of combining the results of two separate studies and also compared the results of prediction generated by exponential and linear models. Study A using 4000 subjects from the population evaluated prediction models based on variables x1, x2 and x3 while study B performed the same analysis based upon variables x4 and x5 in a separate subset of 4000 subjects. Then we combined these two sets of predictors and used our exponential and the linear formulae to predict the outcome. We calculated the mortality for the subjects and compared the prediction using the two different formulas to the true mortality. It is important to mention that when the models are combined, the intercept is calculated after the comprehensive model has been generated.

(2) In our second simulation, called here "the reduced model" we tested the hypothesis that some of the predictors can be removed from the model without major distortion to the prediction. Here study A was based only upon predictors x1 and x2 while study B was based only on variable x3. As in the previous exercise, we combined predictors and derived the probability of the outcome using exponential and linear formulae.

The exponential formula is based upon logistic regression: $P(R)=\frac{1}{1+\left(e^{a+R}\right)^{-1}}$, where $a=ln\frac{r}{1-r}-\hat{R}$; while linear formula is as follows: $P(R)=\frac{R\cdot r}{\hat{R}}$.

The independent variables are described in Table 1. First predictor is a binary variable (e.g., presence of diabetes) with OR of 2.0 (beta=0.69), the frequency of it in the population is 0.2. Second variable is also a binary variable (e.g., presence of hypertension) with OR of 2.5 (beta=0.92) and frequency in the population of 0.3. Third predictor is a numeric variable based on 5 levels (e.g., degree of proteinuria) with OR of 1.6 (beta=0.47), mean in the population is 2. Variable number four is a binary variable, which is protective (e.g., regular physical exercise), OR is 0.7 (beta=-0.35), the frequency in the population is 0.3. Finally, the fifth variable is a continuous one (e.g., age, ranging from 20 to 100), OR is 2.7 (beta=0.02) per 50 years of life (comparing 80 year old to 30 year old). Average age in the population is 43 years. As expected, odd ratios derived by study A and study B are different from initially assigned, as the analyses were performed with a fraction of population and with incomplete set of predictors. The outcome rate (mortality) is 16%. Each simulation was repeated 50 times. The simulation was done using SAS 9.1 and the figures were plotted using R.

	X1	X2	X3	X4	X5
Practical example	Diabetes mellitus	Hypertension	Proteinuria	Regular exercise	Age (years)
Levels	0, 1	0, 1	0, 1, 2, 3, 4	0, 1	20-100
Mean	0.2	0.3	2	0.3	43
Odds Ratio	2	2.5	1.6	0.7	2.7 with 50 units of increase

Table 1: Characteristics of the simulation population described by five independent variables (x1 through x5), which are predictors used in simulation exercises. The overall outcome rate (mortality) was 16%.

Results

First simulation exercise tested the hypothesis of feasibility of simple combination of predictors from two different models into a single model. Study A was based on three variables and study B was based on 2 separate variables. Combined model results are presented in Figure 1. Predictions based on model that combined the results of study A and study B demonstrated good performance. We also compared the results of linear model to the more complex logistic regression model using all five predictors (Figure 1). As demonstrated on the graphs, in subjects with lower risk indicator the outcome of linear model (open square) is very similar to the outcome of the logistic regression model (filled circle) and to the true outcome rate (grey open circle), however it underestimates the risk in the high-risk groups. On the other hand, logistic regression model is accurate compared to actual outcomes.

The second simulation exercise deals with the case where the information regarding predictors is limited, here we constructed models with limited number of predictors: using only x1 and x2 in study A and x3 in study B and later combined those three predictors in a single model. As shown in Figure 2 with each predicted risk score calculated based on x1, x2 and x3, there is a range of true mortalities. This variation is accounted for by the other two predictors x4 and x5 which were not included in the model of study A or study B. However, the logistic model still predicted the mean mortality at each predicted risk score (grey line) with good accuracy (Figure 2).

This confirms our hypothesis that dropping or adding variables should not distort the prediction in any noticeable way. Including more significant variables to the model increases the precision of prediction. Both exponential model and linear model were good approximation of the observed outcome, but as before, in high risk population linear model did underestimate the risk.

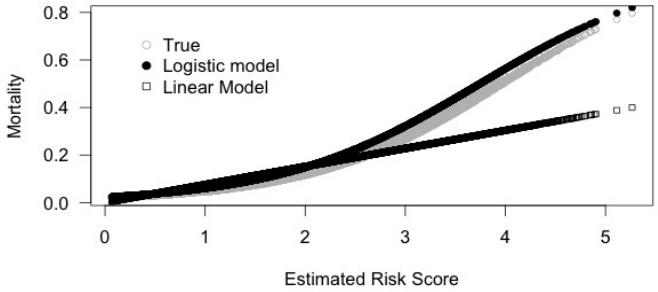

Figure 1: Results of prediction models using all five predictors derived from two separate studies.

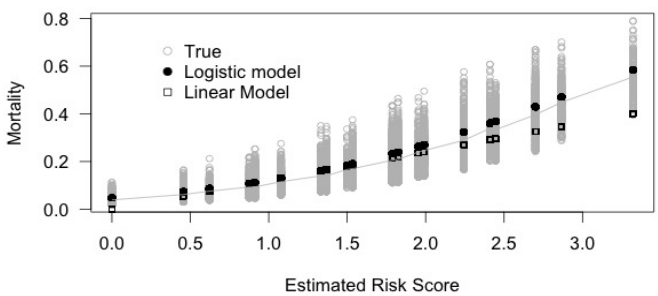

Figure 2: Results of prediction models using only three predictors derived from two separate studies. Grey line is the mean of true mortality at each predicted risk score.

Discussion

Woodpecker™ technique is used to develop prediction models using published reports of outcome studies. For the purpose of developing prediction algorithm the most adequate report would be based upon an exploratory study where multiple variables are included in the model and are reported in the paper. The selection of the variables in this "ideal" model is based on trying to choose the best predictors of the outcome [4]. However it is not always the case and sometimes to save space authors present only the HR/OR associated with primary variable of interest, while indicating that the model was adjusted for other variables [5], and the information regarding other covariates is omitted. Furthermore, it is noted that the design of exploratory models is different from that of prediction model as discussed elsewhere [6]. The study design which is based on a particular hypothesis focused on a primary variable of interest, therefore the selection of the covariates is based on the adequate adjustment of the model for confounders rather than selecting the comprehensive list of likely predictors [7]. These models translated into prediction algorithm will likely have a high degree of uncertainty.

To address this issue one has either to look for the purely exploratory papers where authors specifically use the list of best predictors rather than the association of a particular primary variable of interest with the outcome or to combine several models from different publications. Combining the models is an approximation, and several assumptions are made: (1) Homogeneity [8]: The populations are very similar between the studies (that can be demonstrated by comparing baseline statistics); (2) The independent variables are indeed independent and there is no colinearity; (3) There is no interaction between the variables, especially between those that are being moved from one model to another. We note that meta-analysis and meta-regression methods [8,9] present more elaborate way to combine results rather than simply combining the regression coefficient in a linear way. The latter is simplified method and it is based on approximations, but for practical purpose of developing prediction algorithm it has many advantages.

In this numeric simulation study we demonstrated the feasibility of simple linear combination of the models coming from different studies in the same population. Adding or removing predictors does not seem to distort the model and the predictions remain accurate. Even model with limited number of predictors (using only three out of five variables) can still generate reasonably accurate predictions. As in our previous analyses, linear model tends to underestimate the probability in the subgroups of subjects with higher level of risk.

Conclusion

In conclusion, simple linear combination of prediction models, adding or removing predictors do not cause distortion of the model and predictions remain robust. Prediction of linear model is similar to exponential model, except the former underestimate the predicted probability in the high risk groups.

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. 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: 485-497.

2. Cantarovich F, Martinez F, Heguilen R, Thervet E, Mamzer-Bruneel MF, et al. (2010) Proteinuria > 0.5 g/d, a prevalent prognostic factor for patient and graft survival in kidney transplantation. Clin Transplant 24: 175-180.

3. McGregor JC, Kim PW, Perencevich EN, Bradham DD, Furuno JP, et al. (2005) Utility of the Chronic Disease Score and Charlson Comorbidity Index as comorbidity measures for use in epidemiologic studies of antibiotic-resistant organisms. Am J Epidemiol 16: 483-493.

4. 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.

5. Eddington H, Hoefield R, Sinha S, Chrysochou C, Lane B, et al. (2010) Serum phosphate and mortality in patients with chronic kidney disease. Clin J Am Soc Nephrol 5: 2251-2257.

6. Shmueli G (2010) To Explain or to Predict? Statistical Science :289-310.

7. Kestenbaum B, Sampson JN, Rudser KD, Patterson DJ, Seliger SL, et al. (2005) Serum phosphate levels and mortality risk among people with chronic kidney disease. J Am Soc Nephrol 16: 520-528.

8. Normand SL (1999) Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 18: 321-359.

9. van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 21: 589-624.