Journal of
Health Education Research & Development

Nguyen, et al., J Health Edu Res Dev 2015, 3:3
DOI: 10.4172/2380-5439.1000141

Research Article Open Access

Abilities of Statistical Models to Identify Subjects with Ghost Prognosis Factors

Nguyen JM[123*], **Gaultier A**[1] **and Antonioli D**[3]

[1]*SEB, CHU NANTES, 85, Rue Saint Jacques 44093 Nantes Cedex 01, France*

[2]*INSERM, UMR892, 8 quai Moncousu - BP 70721, 44007 Nantes Cedex 01, France*

[3]*HWRS, Atlanpôle, Route de Gachet, 44300 Nantes, France*

[*]**Corresponding author:** Nguyen JM, SEB, CHU NANTES, 85, Rue Saint Jacques 44093 Nantes Cedex 01, France, Tel: +33-2-40-08-33-33; E-mail: jeanmichel.nguyen@chu-nantes.fr

Rec date: Oct 22, 2015; **Acc date:** Nov 04, 2015; **Pub date:** Nov 06, 2015

Abstract

Background

Many tools are available to estimate prediction quality, but none are available to assess the ability, of a predictive model to identify completely missing or unknown prognostic factors, designated as ghost factors (GFs). However, it may be possible to predict whether a subject carries a GF.

Methods

To simulate the presence of a GF, a significant prognostic factor and all variables correlated with it were removed prior to model analysis. Public datasets and simulated data were used. A predictive statistical model was developed to assess the relationship between the presence of a GF and the predictive capacity of a given model based on the correlation between predicted outcome and GF presence. Five statistical models were compared using this procedure.

Results

After evaluating 6 real databases, the only statistical method consistently able to identify subjects with GFs was the use of optimized regression models. Using simulated, linearly correlated data, optimized regression models exhibited up to a 92% success rate, whereas conventional linear models had less than 53% success. Random forest and classification tree models had the highest success rates compared to the other evaluated models.

Conclusions

Model-based outcome prediction was assessed with respect to the presence of GFs. As GFs are unknown, only subjects who are carriers of significant unknown prognostic factors can be identified. As complex models outperformed linear models in identifying GF presence, we assume that the associations between GFs and outcome-predictive factors are also complex and not linear.

Keywords: Completely missed prognosis factor; Blind man's bluff test; Random forest; Linear models; Optimized regression; Goodness-of-fit; Confounding effect

Abbreviations

GF: Ghost Factor; ROP: Regression Optimized; CART: Classification and Regression Tree; LR: Logistic Regression; RF: Random Forest; DA: Discriminant Analysis; BMB: Blind Man's Bluff

Background

Numerous studies have not been submitted for publication or have been rejected for publication because they were inconclusive. There are many causes for unexpected or inconclusive results. However, real situations may occur where negative results could be attributed to an important but unidentified factor in the original data. Randomization can be used to balance risk factors between randomized groups but cannot control their effects. Thus, if a risk factor has a specific negative effect on an experimental treatment (interaction), inconclusive results may be produced. In this case, stratification is required, but it can only be used if the prognostic factor is known. For example, the results of an inconclusive phase 3 study comparing the probability of event-free survival in stage III melanoma are analyzed. The control group was treated with chemotherapy, and the experimental group was treated with a promising targeted therapy that had demonstrated a very high level of success in phase 2. The data collected do not include information to indicate the presence of a genetic mutation that specifically blocks the mechanism of action of the targeted therapy. This genetic mutation does not modify the response to chemotherapy. Thus, a significant interaction between the genetic mutation and the treatment exists. By randomization, an equal number of patients in each of the 2 study groups have this specific genetic mutation. As the genetic mutation is not yet known, information regarding its presence

is not collected in the trial, and no interaction test is performed. Therefore, the inconclusive results of the trial are caused by a factor that was never identified or analyzed. We call this completely missing factor a Ghost Factor (GF). Another example is a case in which the diagnostic performance of a model is unsatisfactory. It cannot be discerned whether the diagnostic insufficiency is a problem of goodness-of-fit (GOF) due to model choice or due to a lack of analysis of an important predictor in the data. For example, an analysis of data related to survival from the sinking of the Titanic using 3 well-known factors (sex, age, ticket class level) produces a maximum sensitivity of 84.3% and a maximum specificity of 72.13%, regardless of the conventional prediction model used (logistic regression (LR) or random forest (RF)).

Thus, the existence of unknown information within the data that could explain all survival and all mortality is questioned. It is shown that most misclassifications can be explained by the testimonies of surviving passengers [1] and that the use of this information extensively increases prediction quality.

Conversely, numerous studies have published hypotheses that were actually not conclusive because their data did not account for primary etiologic factors but rather for confounding factors. Identifying new prognostic factors represents a real challenge that many researchers aim to solve. We assume that all factors are correlated in real life (i.e., in biological data), particularly factors obtained from the same subject. Such relationships can be rather complex, and they result from multiple events. We hypothesize that this complex information can be used to identify additional missing information, which may or may not be confounding, in analyzed datasets. The concept of a GF differs from the notion of latent variables. In statistics, latent variables are variables that are not directly observed and measurable [2] but rather are inferred from other variables that are directly measured. Thus, a latent variable is correlated with other factors. Latent variables are used to reduce the dimensionality of data for enhanced interpretation. This concept also differs from the problem of missing data and imputation methods [3,4]. Indeed, imputation methods are based on the hypothesis that a predictor has already been collected, but not for all subjects.

This concept also differs from the problem of study-specific missing covariates that can be simulated or imputed (e.g., covariates observed in some studies but missing in other studies) [5]. Here, the covariate is unknown and was not identified elsewhere.

Here, a potential gap exists in the data, but this gap is unknown. Many tools, such as analyses of R^2, discrimination measures, Brier Score, and heterogeneity of random effects, are available to estimate prediction quality [6]. However, no tools are currently being used to assess the ability of a predictive model to identify a completely missing or unknown prognostic factor.

We hypothesize that the information included within a set of predictors can facilitate the retrieval of another predictor that is directly observable and measurable. Because we suppose that this important prognostic factor is unknown, we are able to identify whether subjects are carriers of this GF, but we are not able to exactly identify the true missing information.

To test this hypothesis, 3 steps are required. For simplicity, we assume that the GF is either a binary factor (Yes/No) or a continuous factor with a threshold effect that can be transformed into a dichotomous factor. First, we must prove that it is possible to identify subjects who are carriers of the GF [6]. Second, identifying the presence of the GF must increase the prediction quality of outcome Y^ for all types of predictive models. Third, the GF must be applicable to real data.

The objectives of the current paper were focused on answering the following 2 questions: i) Is it possible to identify subjects who are GF carriers? ii) If yes, what statistical model best achieves this objective?

Materials

Blind Man Buff Test (BMB test)

To answer the first question, we developed a statistical procedure called the Blind Man's Bluff (BMB) test [6,7]. To simulate a GF, a significant prognostic factor that was correlated to outcome Y was removed from a public dataset. We also removed all other predictors correlated with the GF from the initial dataset to minimize confounding effects. Using a simulated dataset, we controlled for confounding effects to assess the role of such confounding effect on the success of the BMB test. In the second step, we assessed the relationship between the predicted outcome Y^ and the GF. We assumed that in the case of a significant association, the tested model was able to take into account the GF when predicting the outcome. Fisher's exact test was performed for a binary GF, and a Wilcoxon signed-rank test was used for a continuous GF.

Datasets

Real public datasets (Table 1): The objective of public dataset selection was to facilitate the verification of the results by all interested parties. As such, the number of predictors had to be limited, and the data had to be easily accessible. All of the data can be loaded using the references provided. Moreover, all of the variables are described on the website. Six public databases with binary outcomes (2 prostate cancer databases [9,10] 1 pharyngeal cancer database [11], 1 prematurity database [12], 1 ICU database [13], and 1 benign breast disease database [14]) were selected. To simulate a GF, the "X-ray" factor was removed from the prostate cancer dataset *9], and the "Smoke" factor and 2 variables ("RACE" and "PTL") that were correlated to minimize the confounding effect were removed from the low birth weight dataset prior to analysis [12]. Similarly, the "DPROS", "COND", "WT" and "SER" factors were removed from the prostate cancer dataset (PCS) [10], the pharyngeal cancer dataset [11], the benign breast disease matched casecontrol dataset (BBDM13)[14] and the ICU dataset [13], respectively. Table 1 presents the GF and the remaining factors that were used to develop each dataset.

Simulated datasets: We simulated datasets by evaluating a binary outcome Y distributed as a logistic function. The prevalence was set at 30% for all of the simulations, creating approximately 2 controls for 1 case. This situation optimizes the power of statistical testing. For each dataset, 500 subjects were simulated with 5 predictors following a binomial distribution. To assess the impact of the statistical level of correlation between the GF (X1) and the outcome (Y), we used statistical correlation levels with p-values equal to 0.01, 0.02, 0.04, 0.06, 0.08 or 0.1. To assess the impact of confounding factors between the GF (X1) and the other predictors (X2, X3,X4, X5), we set a statistical correlation level with a p-value equal to 0.5 for X4 and X5. For X2 and X3, the statistical correlation levels with X1 were set to (0.01; 0.01), (0.04; 0.04), (0.1; 0.1), (0.01; 0.1), (0.01; 0.04), and (0.04; 0.1). To assess the false positive rate, we also simulated a situation in which X1 was

correlated to Y with a p-value of 0.5, and X2, X3, X4, and X5 were correlated to X1 with p-values of 0.5.

Dataset	Source	Ghost factor	Outcome	Other predictors included	Ghost factor ~ Other predictors included (error rate)	ROP	RL	RF	CART	DA
Benign Breast Disease 1-3 Matched Case-Control Study (BBDM13.DAT)	Hosmer and Lemeshow (2000) Applied Logistic. Regression: Second Edition, page 245. https://www.umass.edu/statdata/data/bbdm13.txt	WT	Final diagnosis	AGMT+HIGD+AGLP+DEG	Adjusted R^2 0.004108	0.043 (*)	0.96 (NS)	0.011 (*)	0.671 (NS)	0.882 (NS)
ICU.DAT	Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) - https://www.umass.edu/statdata/data/icu.txt	SER	Vital Status	GENDER+CRN+SYS+PRE+LOC	Misclassification rate : 0.35	0.020 (*)	0.049 (*)	0.028 (*)	0.058 (NS)	0.034 (*)
Prostate Cancer	http://www.agrocampus-ouest.fr/math/livreR/cancerprostate.txt	Xrays	Y	AGE+ACID+GRADE+SIZE	Misclassification rate : 0.3585	0.003 (**)	0.159 (NS)	0.002 (**)	0.384 (NS)	0.116 (NS)
Low birth weight	https://www.umass.edu/statdata/statdata/data/lowbwt.txt	Smoking	LOW	AGE+LWT+PTL+HT+UI+FTV	Misclassification rate : 0.3757	0.0318 (*)	0.035 (*)	0.020 (*)	0.002 (**)	0.059 (NS)
Prostate Cancer Study (PCS.DATA)	Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition. https://www.umass.edu/statdata/data/pros.txt	DPROS	CAPSULE	AGE+RACE+VOL	Adjusted R^2 0.004469	0.034 (*)	0.960 (NS)	0.094 (NS)	0.671 (NS)	0.881 (NS)
Pharynx (PHARYNX.DAT)	"The Statistical Analysis of Failure Time Data, by JD Kalbfleisch& RL Prentice, (1980), Published by John Wiley & Sons - https://www.umass.edu/statdata/statdata/data/pharynx.txt	COND	TX	SEX+TX+GRADE+T_STAGE+N_STAGE	Adjusted R^2 0.05577	0.043 (*)	0.844 (NS)	0.008 (**)	0.223 (NS)	0.008 (**)

Table 1: Real public datasets.

Three hundred datasets were simulated for each of these situations, which led to 108,300 (37*300) simulated samples. The simulations were performed using R statistical software. All simulation programs are available from the authors upon request.

Comparison of statistical models

The performance of 5 statistical models, including parametric and non-parametric models, was assessed. Two parametric and linear model (LR, discriminant analysis (DA)) and 3 nonparametric and nonlinear models (classification and regression tree (CART), RF [15] and the regression optimized (ROP) model) were compared. The ROP [6,7] model is based on a tree on which each node is a set of risk scores including linear combinations of predictors. It is nonparametric because coefficients are not estimated but rather systematically screened from a range of all possible values. It is nonlinear because it acts as a type of decision tree. Coefficient selection depends on an algorithm that tests all possibilities of combinations, for which a patent has been filed. ROP model results are presented as a decision tree that is easy to understand, interpret and use. A detailed example able to verify the results and performance levels presented in this paper is presented in Figure 1. Additional examples are available from the authors upon request.

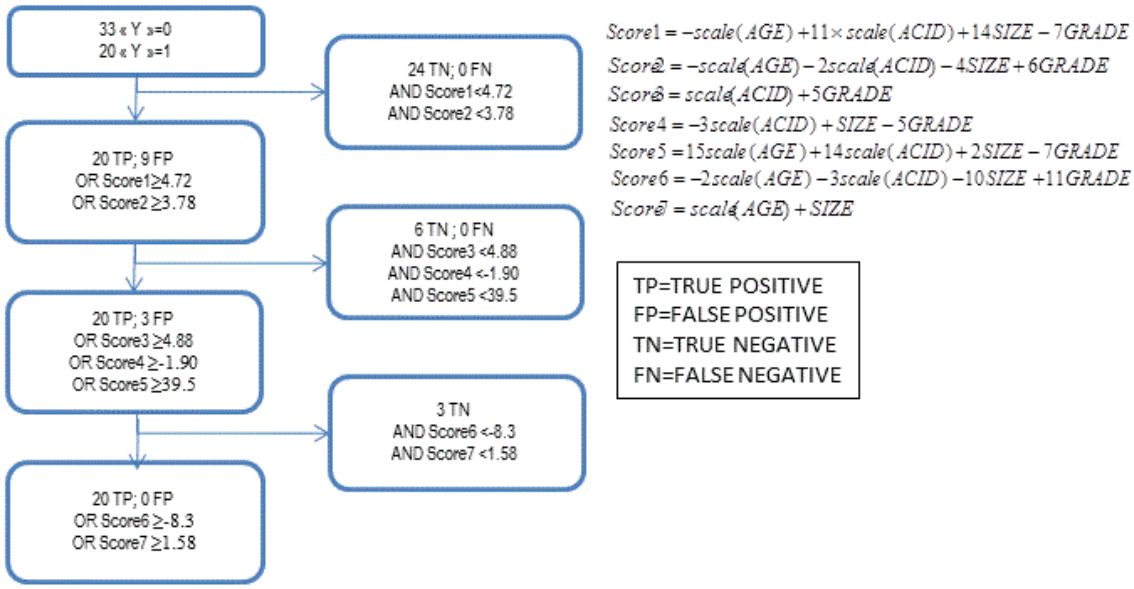


Figure 1: Prostrate cancer dataset. Outcome='Y'. Predictors='AGE'; "ACID"; "SIZE"; 'GRADE'.

Results

Performance with real databases (Table 1)

Conventional statistical methods (LR, CART, DA and RF) allowed us to identify subjects with GFs in 2/6, 1/6, 2/6 and 5/6 datasets, respectively. The ROP model demonstrated success for all of the datasets (Table 1). Thus, a significant relationship was noted between the outcome that was predicted by the model and the GF. An example of the ROP model, including risk scores and thresholds, is presented in Figure 1.

Performance with simulated databases

The RF and ROP models exhibited the best performance in the BMB test. The proportion of success for the BMB test increased with the level of correlation between the GF and the outcome from 19% to 92% (Figures 2A-2C). The ROP model consistently demonstrated the best performance compared to the other models.

For high levels of correlation between the GF and the outcome (p=0.01; p=0.02; Figure 2A), the rate of success (positive BMB tests) varied from 76% to 92% for the ROP model, from 66% to 89.3% for the RF model, from 40% to 55.3% for the CART model, from 27.7% to 43.7% for the LR model and from 26% to 52.3% for the DA model.

For intermediate levels of correlation (p=0.04 and p=0.06, Figure 2B), the rate of success (positive BMB tests) varied from 46% to 71.3% for the ROP model, from 43% to 62.3% for the RF model, from 26.7% to 41% for the CART model, from 19.7% to 41% for the LR model and from 21.7% to 48% for the DA model.

For insignificant levels of correlation (p=0.08 and p=0.1, Figure 2C), the rate of success (positive BMB tests) varied from 31.3% to 69.7% for the ROP model, from 31.3% to 39.7% for the RF model, from 23% to 32% for the CART model, from 22% to 38.3% for the LR model and from 20.7% to 41% for the DA model.

For the scenarios wherein no statistical correlations were noted between X1 and Y or between X1 and the other predictors (all p-values=0.5), the rate of success was equal to 6.3% with a 95% confidence interval (CI) of [3.86% -9.71%] for the ROP model, 3.3% with a 95% CI of [1.61% -6.05%] for the RF model, 12% with a 95% CI of [8.55% -16.22%] for the LR model, 9% with a 95% CI of [6% -12.82%] for the CART model and 9.3% with a 95% CI of [6.29% -13.21%] for the DA model.

Discussion

The BMB test results showed that outcome Y^ in the ROP and RF models was linked to a GF. Consequently, the model made it possible to identify patients for whom a prognosis factor was not collected. As the GF is an unknown factor, the only way to prove this concept is to remove a real and known predictor from a dataset and allow the removed predictor to play the role of a GF. The challenge is then to retrieve subjects who are carriers of the GF.

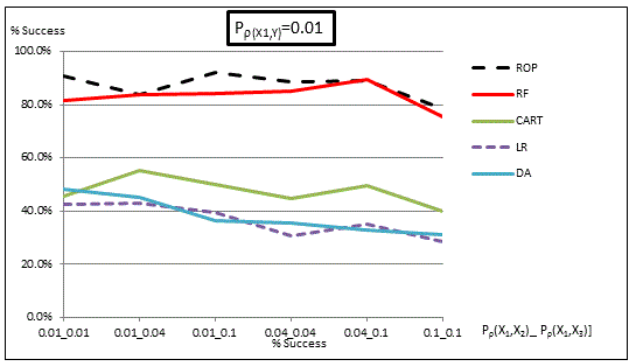
Model performance

In real databases, no significant linear correlations were observed between the GF and the other evaluated variables; all other predictors correlated with the GF were previously removed from the analyzed data. Thus, there is no linear confounding effect. Our hypothesis was that all data are linked in a biological system. Therefore, some information related to the GF persisted within the remaining factors. A given biological value is the result of millions of other values, but the nature of these associations is mostly unknown. We hypothesize that a non-linear relationship may be involved and therefore linear models are not suitable to assess how one predictor is associated with others.

This hypothesis is consistent with the finding that nonlinear models (i.e., the RF and ROP models) outperformed linear models in the BMB test.

To verify this hypothesis, we simulated data with linear correlations between the GF and outcome as well as between the GF and other predictors (confounding effect). A logistic function was used to simulate data. We hypothesized that linear models would exhibit comparable performance in the BMB test to nonlinear models, such as the ROP and RF models, in cases of linear correlation. The results obtained using simulated data did not confirm our hypothesis. Surprisingly, linear model performance, such as that of the LR model, was not comparable to the performance of the ROP and RF models. The nonparametric models tended to perform better due to the degree of complexity of the model itself.

In the ideal situation, wherein 2 confounding factors correlated with the GF with a p-value of 0.01 (R^2=0.02), the rates of success in the BMB test for the linear models did not exceed 50%, whereas the success rates were 90.7% and 81.7% for the ROP and RF models, respectively. The differences in performance between the linear (LR, DA) and nonlinear (ROP, RF, CART) models decreased when the correlation between the GF and the outcome decreased. For correlation level p-values equal to 0.1 (R^2=0.006) between the GF and the outcome as well as between the GF and the other predictors, the differences in performance were less than 10% (33.3% for ROP, 23.3% for LR; Figure 2C).



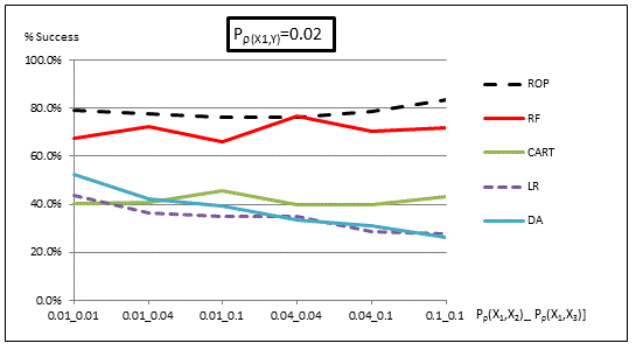
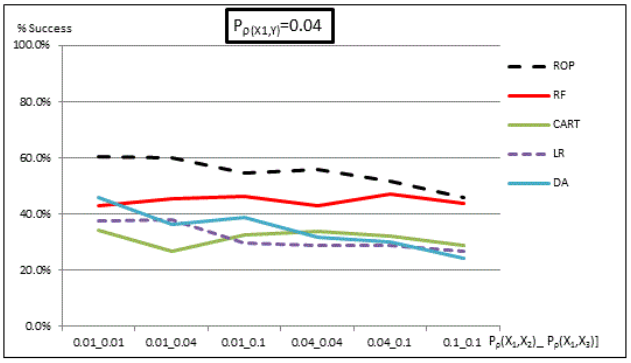


Figure 2A: Performances to the BMB test according to different levels of correlation (p=0.01 and p=0.02) between the ghost factor and the outcome [$P_{p(X1,Y)}$], between the ghost factor and the other predictors [$P_{p(X1,X2)}$-$P_{p(X1,X3)}$].



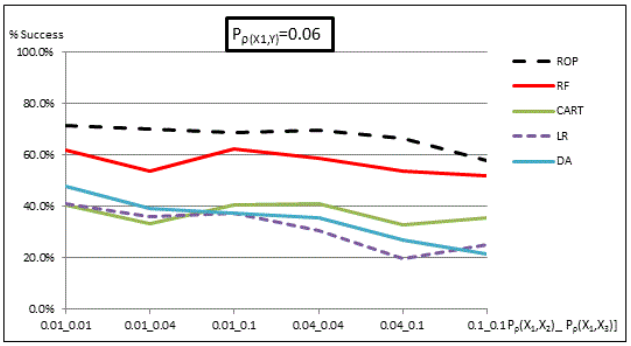


Figure 2B: Performances to the BMB test according to different levels of correlation (p=0.04 and p=0.06) between the ghost factor and the outcome [$P_{p(X1,Y)}$], between the ghost factor and the other predictors [$P_{p(X1,X2)}$.$P_{p(X1,X3)}$].

The results from the case in which all correlation level p-values were set to 0.5 are interesting. The ROP and RF models provided lower rates of success compared with the other models. This result indicates that the proportion of false positive tests was greater for the linear models. We noticed that the 95% CIs for the ROP and RF models included a 5% threshold (considered the threshold of randomness), whereas the other models excluded this threshold. These results demonstrated that obtaining a p-value of less than 0.5 for a linear correlation test is an interesting finding that could lead to the identification of new prognosis factors for RF and ROP models. Of all the tested models, both linear and nonlinear, the ROP model consistently demonstrated the best ability to identify patients with GFs and appeared to fit the data. The solutions that were proposed by the ROP models revealed a portion of the complexity of the relationship between predictors. The risk coefficients for a given factor changed between branches, thus indicating a different role according to each subgroup of subjects.

Perspective Conclusion

The next objective of our research is to demonstrate that adding outcome Yˆ from the ROP model as a covariate in the LR or RF models significantly increases prediction quality up to the attainment of a perfect prevision with sensitivity and specificity equal to 100%. These results could demonstrate that outcome Yˆ from the ROP model represents the missing information needed for a perfect prevision and therefore contains a GF.

The development of an algorithm that does not use predicted outcome Yˆ is ongoing. This algorithm uses several subgroups defined by the ROP model. The first results we obtained are impressive and prove that information from a complex model can actually be used to increase the prediction quality of conventional models.

As a consequence, it will be possible to identify the nature of this new information by investigating its relationship with different molecular mechanisms of pathology.

Competing Interests

Financial competing interests: The development of the ROP model was supported by Atlantpôle, a public organization supporting the development of technological innovations in Nantes (France). The possibility of developing commercial software is being evaluated.

Non-financial competing interests: None

Contributing Authors

JMN wrote all sections of the paper, analyzed and interpreted all data, and was involved in the conception and design of the paper.

AG developed programs for the simulation and acquisition of data, wrote the simulation results, and was involved in the drafting of the results.

DA wrote the ROP programs in R and discussed all results and conclusions.

All contributors have approved of the final published version of the manuscript.

References

1. Nguyen JM, Gaultier A, Antonioli D (2014) The Titanic reviewed by ROP, a new method of nonparametric regression combined with classification, RESP 62: 45-46
2. Bollen KA (1989) Structural equations with latent variables. Wiley: New York pp: 11.
3. Enders CK (2010) Applied Missing Data Analysis, Guilford Press: New York, 2010.
4. Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 35: 1925-1931.
5. Wang F, Song PX, Wang L (2015) Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach. Biometrics .
6. Nguyen JM, Gaultier A, Antonioli D (2013) A Combined NonParametric Regression and Classification Model for Subgroup Selection. ISBC, Munich pp: 25-29.
7. Nguyen Jean-Michel (2013) Blind man's bluff test using ROP- WIN symposium, Paris.
8. Nguyen JM, Gaultier A, Antonioli D (2012) How to identify an unknow factor in targeted therapies. P04-074-2nd International Symposium of the Cancer Research Center of Lyon.
9. Cornillon PA, Guyader A, Husson F, Jégou N, Josse J, et al (2012) R for Statistic, Chapman & Hall/CRC press, New York.
10. Hosmer DW, Lemeshow S (2000) Applied Logistic Regression: (2nd Edn) Wiley: New York.
11. Kalbfleisch JD, Prentice RL (1982) The Statistical Analysis of Failure Time Data, Wiley: New York.
12. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression: (3rd Edn) Wiley: New York.
13. Hosmer DW and Lemeshow S (2013) Applied Logistic Regression: (2nd Edn)Wiley: New York
14. Breiman L(2001) Random Forest. Machine Learning 45: 5-32.