Role of Predictive Modeling in Healthcare Research: A Scoping Review

Nihar Ranjan Panda[1,2], Jitendra Kumar Pati[2] and Ruchi Bhuyan[3,*]

[1]*Department of Medical Research, IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha, India*

[2]*Department of Mathematics, CV Raman Global University, Bhubaneswar, Odisha, India*

[3]*Department of Medical Research, IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha, India*

Abstract: The huge preponderance of inferences drawn in empirical medical research follows from model-based relations (e.g. regression). Here, we described the role of predictive modeling as a complement to this approach. Predictive models are usually probabilistic model which gives a good quality fit to our data. In medical research, it's very common to use regression models for predictive purposes. Here in this article, we described the types of predictive modeling (Linear and Non-linear) used in medical research and how effectively the researchers take decisions based on predictive modeling, and what precautions, we have to take while building a predictive model. Finally, we consider a working example to illustrate the effectiveness of the predictive model in healthcare.

Keywords: Predictive model, Healthcare, Roc curve, Model validation.

INTRODUCTION

The healthcare division has witnessed an unbelievable progression following the development of new computer innovations and that pushed this region to deliver increasingly restorative information, that which brought forth special fields of research many activities are done to acclimatize to the blast of therapeutic information on one hand and to obtain precious knowledge from it. To assist in building decisions and to take out valuable information these encouraged specialists to relate all the specialized developments like predictive analytics, learning algorithms, and machine learning. In health science to establish the risk of the construction of a disease, the prediction models are used so that they can allow before-time treatment or deterrence of that particular disease. Prediction exploration, which motivates to predict future actions or outcomes based on patterns inside a set of variables, has to turn out to be gradually more admired in health research [1].

With the help of exact prognostic models, doctors can know the future course of a disease or how many risk factors are involved it helps the doctors to take appropriate decisions and manage. We can take an example here many predictive models are invented in gastroenterology to forecast the danger of disease [2,3]. In the current century diabetes is a very severe and common disease among people. It is a chronic disease, which is characterized by hyperglycemia. Many diseases are associated with diabetes directly or indirectly. According to the prediction, in 2040, globally diabetes patients will reach 640 million. It means every one patient out of ten will suffer from this chronic disease. With the rapid growth of technology and machine learning techniques, it is easy to predict diabetes mellitus. Many predictive models have been invented to predict the diabetes of a patient without any blood test and also the models give us an idea about the risk factors involved with diabetes so that we can take necessary actions accordingly. There are some differences between explanatory research and predictive modeling-based research [7]. Usually, explanatory research deals with statistical methods to check hypotheses on a prior theoretical concern e.g., hepatocellular carcinoma surveillance underutilization is related to provider-level factors) [4]. Numerous models have been invented to predict the readmission risk of patients [5,6]. Similarly many predictive models have been introduced in the literature to predict the birth weight of a baby and many healthcare models have been introduced in the literature, using maternal determinants to predict the birth weight of the baby in 2008 Biswal *et al.* described a community-based epidemiological study regarding the prediction of birth weight. Dharmalingam *et al*. (2010): described another model using a mother's nutrition status and sociological factors to predict birth weight. The objective of this article is to illustrate essential methods for doing prediction-based research, which can be separated into

*Address correspondence to this author at the Department of Medical Research, IMS & SUM Hospital, SOA deemed to be University, Bhubaneswar, Odisha, India; E-mail: ruchibhuyan@soa.ac.in

3 main steps: rising a predictive model, separately validating its performance, and prospectively studying its scientific impact.

TYPES OF PREDICTIVE MODELS IN HEALTHCARE

In many cases, prediction research in healthcare has conventionally used a Bayesian structure method [8,9]. Statistical techniques, data mining artificial intelligence and machine learning are in trend nowadays [8]. In medical research implementing a predictive model depends upon the data type in many scenarios the collected data are categorical and in other cases we may find some continuous variables as we know the measurement scale is four types i.e., nominal Ordinal, interval, and ratio our choice of predictive model will depend upon the nature of the outcome variable for a continuous response variable researchers usually use a linear regression approach whereas in a classification problem researchers may use binary logistic regression, multinomial regression, ordinal regression, Decision trees, Random forest, etc. it depends on whether our data set is having a linear relationship between them or a nonlinear relationship. Various software is available to do such kind of predictive modeling in current scenarios python programming language R programming language, spss, stata, sas are in trend to do predictive modeling [11,12]. If we will consider machine learning algorithms, it has various advantages over conventional explanatory statistical modeling. Usually, machine learning approaches do not follow a predefined hypothesis. Forthcoming a predictive problem with no precise fundamental assumption can be quite efficient when various possible predictors exist and when there are connections between predictors, which are ordinary in biological processes. Here we represent the

Table 1: Distribution of Regression Methods while doing Predictive Modeling

Predictive model	Outcome (Dependent Variable)	Predictors (Independent variables)	Example
Simple linear regression	Continuous	One independent variable.	To predict the birth weight of a newborn baby using one independent variable.
Multiple Linear regression	Continuous	More than 2 (factor/continuous)	To predict the length of stay in hospital for a certain disease using more than 2 predictors.
Binary logistic regression	Factor (2 levels)	More than 2 (factor/continuous)	Prediction of stillbirth (yes\No)
Ordinal logistic regression	Ordinal	More than 2 (factor/continuous)	Predicting patient satisfaction for hospital service i.e. satisfied, partially satisfied, not satisfied.
Multinomial logistic regression	Factor (more than 2 levels)	More than 2 (factor/continuous)	Predicting choice of drinking e.g. Coffee, Soft Drink, Tea, and Water

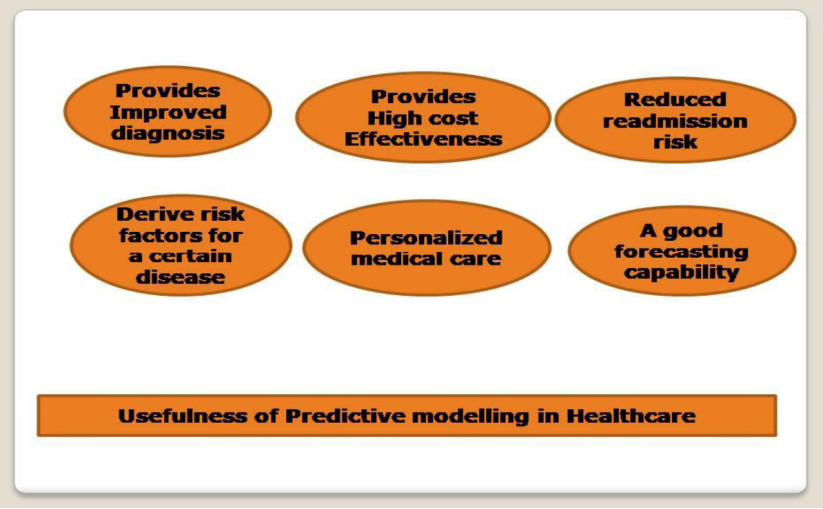

Figure 1: Application of predictive modeling in healthcare.

regression family to create the predictive modeling. The preferred regression models are chosen not only according to the variable scale but also by considering some assumptions. While creating a linear regression model there must be a linear relationship between response and predictor variables. There must not be multicollinearity present in the data set and we should follow the rule of homoscedasticity while building a linear model. Logistic regression does not require many assumptions as linear regression it is a general linear model and is also considered a classification model, where the response variable is dichotomous. It need not requires the assumption of normality, homoscedasticity, and multicollinearity. Ordinal regression assumes the response must be ordered and requires proportional odds.

MOUNTING A PREDICTIVE MODEL (PM)

The 1st footstep in rising a pm, while using the normal regression method, is choosing appropriate applicant-independent variables for probable insertion in the model. But there is no such best strategy to do so [13]. Researchers use the backward-elimination method to decide which variable will be included in the model or not. Basically the insignificant variables were removed from the final model. Whereas over fitting and selection bias are two measures concerned in the final model. When fitting a predictive model, we have to verify these things such as whether there is a problem of over fitting exists in our final model. Whether the RMSE (ROOT MEAN SQUARE ERROR) is coming too high in our model or not RMSE is defined as an error term in the model it is the squared difference between the actual value and the fitted value we always expect that our model performance will be good. So for that, we expect our error term in the model will be as less as possible. Many times researchers add irrelevant predictors to build a predictive model, but this is not a good practice to do so [14,15] only the relevant predictors must be added to predict the outcome variable. There is numerous problem we may face while building a predictive model. Missing data is a measure problem among them [10, 16, 17]. So we must be aware of this kind of situation in case of missing data how to deal with such kind of situation.

VALIDATING A PREDICTIVE MODEL

To establish an outstanding predictive model, it must not only have the predictive capability but also it should execute well in a validation [7, 18]. Many times researchers create predictive modeling to predict their response variable but they do not focus on the validation of the model or in a simple way we can say, how the model is working with the real data. A model's presentation may vary considerably among derivation and validation cohorts for numerous reasons for example overfitting of the model, and the absence of essential independent variables [19]. When a validation study indicates unsatisfactory consequences, investigators frequently used to decline the initial model and build up a new predictive model by the validation cohort data. There are more than 60 predictive models were established in the literature to predict breast cancer. This move neglects the information captured from previous studies and predictive models. There are numerous methods to modernize previous predictive models with data from the patients of the validation cohort, but these are unluckily hardly ever utilized. Many models were created according to the study variables but they are rarely used by the researchers.

EXAMPLE OF PREDICTIVE MODELLING IN CLINICAL RESEARCH

Many mathematical and statistical models have been introduced in the literature to provide a better forecast of outcome variables [20], for example predicting a birth weight of a newborn baby, predicting whether the newborn baby will be an underweight or normal category, predicting whether the woman will face a stillbirth or not, etc. Numerous examples have been described in the literature to predict clinical outcomes. While considering a linear regression approach to build the model we need to look out for some important terms like, what % of the variation is explained by the independent variable? Whereas while building a classification model we need to verify the sensitivity, specificity, area under the curve, ROC curve, model fit, and model summary. Now let us consider one example to interpret the predictive model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \cdots \beta_n x_n + \varepsilon \quad (1)$$

Let us consider y as our outcome variable which is birth weight here. To predict the birth weight we will use the linear regression approach here. Since birth weight is a continuous variable.

$$Y = 1153.825 - 44.417x_1 - 3.060x_2 + 8.552x_3 + 74.602x_4 \\ -80.152x_5 + 24.491x_6 + 0.826x_7 - 1.096x_8 - 190.815x_9 + \varepsilon \quad (2)$$

Y= 1153.825-44.417(*bmi*)-3.060(socio economic status) + 8.552(gestational age) + 74.602(immunized) 80.152(supplements taken) 24.491 (booked) 0.826 (mode of delivery) – 1.096(heart rate – 190.815 (smoking) + ε (3)

Table 2: Classification Table for Stillbirth Prediction

Observed	Predicted		
Stillbirth Status	No(0)	Yes(1)	Percentage Correct
No (0)	194	16	92.4
Yes (1)	54	66	55.0
Overall Percentage			78.8

Now from the above regression model, we can predict the birth weight of a baby. However, the p-value plays a significant role to predict the significant independent variables for our response variables. In the above example, we have considered nine independent variables to predict the birth weight of the baby. However, the selection of variables depends upon the researchers. Similarly, now let us consider one classification problem to understand how it works and how researchers create classification modeling to predict the outcome will fall which of the category. Here we have taken the example of a binary logistic regression model to predict the outcome variables.

$$P(y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}} \quad (4)$$

$$= \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots\beta_i x_i)}} \quad (5)$$

Now from the above classification matrix, we can see that we have performed the binary logistic regression model and the result of the confusion matrix is given above. Now we can say that the sensitivity of the model is (55%) whereas the specificity of the model is (92.4%) and the overall accuracy of the model is 78.8%. The graph of sensitivity and specificity can be obtained by a ROC curve.

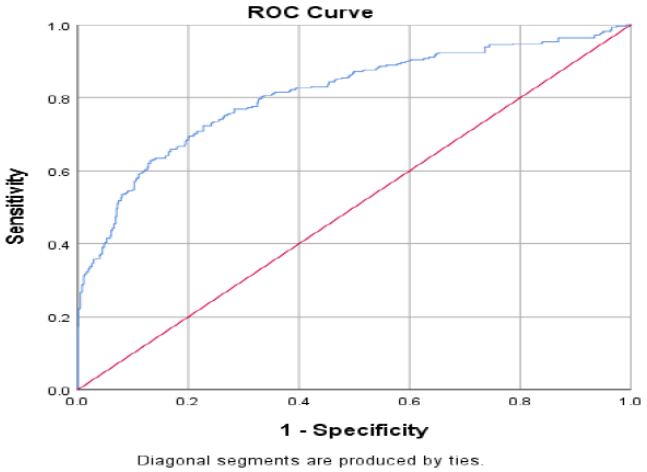

Figure 2: Roc curve for sensitivity vs. specificity (source spss).

DISCUSSION

In today's scenario, we can find numerous applications of predictive modeling in healthcare management [23, 24] here in this review we discussed some important parts of healthcare, where we use the models most frequently. By using a predictive model we can estimate the exact cost of insurance for a specific individual it is also possible to define how reasonable it would be to give a particular medical insurance plan to the applicants using various parameters such as age, sex, region, insurance history, medical history, any habit, heredity, etc. Now day's predictive models are giving a high efficiency in the field of radiology. It can be done by using machine learning and artificial intelligence techniques. For this, we considered one of the most important examples in the university of Montreal hospital center. This hospital center uses AI techniques that find anatomical changes in patients and detect disease-specific markers regarding x-ray photographs. This helps to prepare patients for surgical intervention based on prediction.

Usually taking a decision based on a predictive model is often complicated it requires a lot of consideration before reaching a course of action in patient care. In today's scenario, lots of predictive models and machine learning models are available to take decisions based on healthcare data. As we discussed in this paper predictive modeling requires many aspects for development and validation. Currently healthcare researchers are using a predictive model for some major diseases e.g., diabetes, heart disease, cancer, chronic kidney disease; lower back pain, maternal mortality, and stillbirth. Hence it is very useful to predict the outcome of any disease so that we can take necessary action to prevent the disease.

CONCLUSIONS

In our point of view, predictive models may not substitute clinical judgment; they only provide an estimate of the future path of certain diseases and provide us with important information in healthcare

management. Many models have been adopted in clinical practice to identify risk factors and admission risks associated with the disease. Moreover, more percussion should be taken, to build an accurate predictive model on which we can relay. Blindly relying on these models is not a good practice. Studying the clinical impact of predictive modeling is our future work.

REFERENCES

[1] Toll DB, Janssen KJ, Vergouwe Y, *et al*. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 2008; 61: 1085-1094. https://doi.org/10.1016/j.jclinepi.2008.04.008

[2] Waljee AK, Joyce JC, Wang SJ, *et al*. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clin Gastroenterol Hepatol 2010; 8: 143-150. https://doi.org/10.1016/j.cgh.2009.09.031

[3] Singal AG, Mukherjee A, Higgins PD, *et al*. Machine learning algorithms outperform conventional regression models in identifying risk factors for hepatocellular carcinoma in patients with cirrhosis. Am J Gastroenterol 2013; 108: 1723-1730. https://doi.org/10.1038/ajg.2013.332

[4] Singal AG, Yopp AC, Gupta S, *et al*. Failure rates in the hepatocellular carcinoma surveillance process. Cancer Prev Res (Phila) 2012; 5: 1124-1130. https://doi.org/10.1158/1940-6207.CAPR-12-0046

[5] Singal AG, Rahimi RS, Clark C, *et al*. An automated model using electronic medical record data to identify patients with cirrhosis at high risk for readmission. Clinical Gastroenterol and Hepatol 2013; 11: 1335-1341. https://doi.org/10.1016/j.cgh.2013.03.022

[6] Moons KG, Royston P, Vergouwe Y, *et al*. Prognosis and prognostic research: what, why, and how? BMJ 2009; 338: b375. https://doi.org/10.1136/bmj.b375

[7] Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ 2009; 339: b4184. https://doi.org/10.1136/bmj.b4184

[8] Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. Am J Gastroenterol 2010; 105: 1224-1226. https://doi.org/10.1038/ajg.2010.173

[9] Siegel CA, Siegel LS, Hyams JS, *et al*. Real-time tool to display the predicted disease course and treatment response for children with Crohn's disease. Inflamm Bowel Dis 2011; 17: 30-38. https://doi.org/10.1002/ibd.21386

[10] Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013, ISBN 3-900051-07-0, http://www.R-project.org/

[11] Breiman L. Random forests. Machine Learning 2001; 45: 5-32. https://doi.org/10.1023/A:1010933404324

[12] Liaw A, Wiener M. Classification and regression by random Forest. R News 2002; 2: 18-22.

[13] Royston P, Moons KG, Altman DG *et al*. Prognosis and prognostic research: developing a prognostic model. BMJ 2009; 338: b604. https://doi.org/10.1136/bmj.b604

[14] Greenland S. Modelling and variable selection in epidemiologic analysis. Am J Public Health 1989; 79: 340-349. https://doi.org/10.2105/AJPH.79.3.340

[15] Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. J Clin Oncol 2012; 30: 3297-3303. https://doi.org/10.1200/JCO.2011.38.7589

[16] Kaambwa B, Bryan S, Billingham L. Do the methods used to analyse missing data really matter? An examination of data from an observational study of Intermediate Care patients. BMC Res Notes 2012; 5: 330. https://doi.org/10.1186/1756-0500-5-330

[17] Waljee A, Mukherjee A, Singal A, *et al*. Comparison of modern imputation methods for missing laboratory data in medicine. BMJ Open 2013; 3: pii: e002847. https://doi.org/10.1136/bmjopen-2013-002847

[18] Altman DG, Vergouwe Y, Royston P, *et al*. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; 338: b605. https://doi.org/10.1136/bmj.b605

[19] 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: 774-781. https://doi.org/10.1016/S0895-4356(01)00341-9

[20] Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128-138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Received on 28-07-2022 Accepted on 10-09-2022 Published on 19-09-2022

https://doi.org/10.6000/1929-6029.2022.11.09