**A research about the ADRs on Clozapine**

abstract: Clozapine, a kind of atypical antipsychotic medication, is widely used for schizophrenia and have satisfactory effects, the drug decrease the rate of suicidal behavior efficiently, and have better result than the typical antipsychotics. However, this drug may cause several ADRs since it is related to a high risk of low white blood cells, in fact, many side effects such as Agranulocytosis, Weight Gain, Constipation, Hypersalivation, Seizures, venous thromboembolism, Mypcarditis, Nocturnal enuresis and Pneumonia may occur. Based on the data from patients who have schizophrenia but treat with different drug, we try to analysis and calculate the probability that patients use Clozapine will cause those ADRs and build a statistical model to fit it.

**1. Introduction**

**1.1 Clozapine**

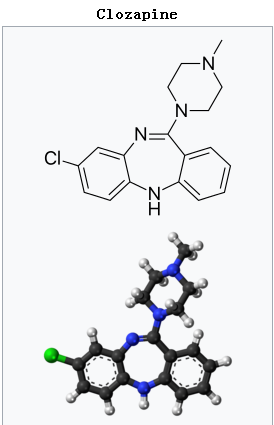
Clozapine (figure 1) was firstly made in the year 1958 by a Swiss pharmaceutical company: Wander AG, and sold in the year 1972, it is the first atypical antipsychotic which plays a significant role in the treatment of schizophrenia. [1]

Figure 1

However, it was not successful when it was made, in the year 1962, people did the first trial in human, this drug was considered as a failure, later, trials in German in 1965 and 1966 were successful, after the successful trials in 1972, Clozapine was approved to appear on the market[2]. Comparing with the other medicine in treating schizophrenia, Clozapine have better efficacy, in a 2013 research in a comparison of 15 different medicines in treating mental illness and schizophrenia symptoms, the result showed that Clozapine was ranked first and 25% better than amisulpride which was ranked second and 33% more effective than olanzapine which was ranked third[3]. Clozapine may increase the risk of catching many ADRs despite that it does well in treating schizophrenia, as a matter of fact, some side effects including constipation, bed-wetting, sedation, tremors, hypersalivation and weight gain are common but nothing serious, but some are potentially fatal, such as agranulocytosis, central nervous system depression, seizure and myocarditis et.[4][5]

**1.2 logistic regression**

Logistic regression is a regression model which have the categorical dependent variables, for the simplest case: binary logistic regression, the output just contains two values: 0 and 1 which can represent several opposites: win and lose, survival and death and so on. Logistic regression provides an excellent method to quantify the nominal variables and is widely used in economic and medical domain.[6][7]

Logistic regression model was firstly developed by David Cox in 1958, differ from the linear regression, logistic regression model get a dichotomous outcome, take the simplest binary logistic regression as an example, the logistic function of a logistic regression model can be written as , while is parameter matrix, X is factor matrix and a is intercept term. if , , we classify the point into the first group and the outcome is 0, if , the outcome is 1. The application of logistic regression in medical domain has both advantages and disadvantages, advantages includes that the logistic regression requires less formal statistical training, and is able to implicitly detect the complex nonlinear relationships between dependent or independent variables, the requirements for the possible variables are not strict but can detect all interactions among variables. The disadvantages of logistic regression are also obvious, it brings greater computation burden and faces an increasing risk of overfitting, the development of the model is limited.[8]

Comparing with the normal linear regression, logistic regression uses cumulative logistic distribution function to estimate the probability to determine which category does a new point belongs to, it can be considered as a special case of the generalized linear model, however, they are based on different assumptions, logistic regression model assume that the conditional distribution is a Bernoulli distribution rather than a normal distribution, in addition, the predicted values of a logistic regression model is probabilities which are restricted to [0,1]. The goodness of the model is usually judged by two parts: If the model is built through statistical software like SPSS, it is judged by statistics such as p-value, decisive factor et. If the model is built by machine learning methods such as gradient descent, bootstrap or split-sample validation, we usually evaluate the model by calculating F1-score, sensitivity and specificity. [9]

**1.3 previous research**

There are many researches in medical domain with logistic regression since 1990s, the components that may influence the accuracy of the model are presented constantly and researchers keep trying to overcome them. In the year 2001, a research group from California examined the usage of LR in medical domain and wrote a report. The report shows that there were 15 peer-reviewed English-language articles published between 1985 and 1999 which had used the logistic regression model in medical research, the topics are concerned about patients' interest in genetic testing for cancer susceptibility. The articles reported the relationship between the interest and 10 criteria with logistic regression model, however, based on the plenty calculation and test with real data, the result seems to be disappointed, the ratio of number of outcome events to predicted variables was so small so that the accuracy of the model tended to be suspect, also, there is no validation analysis and goodness-of-fit measure reported in those articles. This article aimed to point out the potential risk and problems of the widely use of logistic regression in medical research, it reminds us that although logistic model is a pretty excellent method to get the dichotomous outcome, we need to test the accuracy, report the goodness-of-fit and do the validation analysis after the building of a model, find out any possible reason that result in the bad result and take more factors into consideration. In the year 1996, a research group tried to perform a Monte Carlo simulation to evaluate the effect of the number of events per variable(EPV) analyzed in logistic regression model, the data and events were based on a trial with 673 patients in which 252 deaths occurred and 7 variables were considered to be related. In their simulation, they tried EPV = 2,5,10,20, and 25 and randomly selected 500 samples from 673 patients(with replacement) for each EPV. Firstly, they used a logistic regression for the full sample and calculated the probability of dying and survivors:

, P(death) = 1/[1+ exp{-()}], P(survive) = Q = 1 - P(death), then, they started their simulation, for each selected patient, they calculated , if they were selecting a death, they calculated , the algorithm for selecting death was based on cumulative death and survival selection probabilities, , also they have a random uniform number series u between 0 and 1, if , the patient i was considered as death, the selection would stop if 7\*EPV death was obtained. For example, when EPV = 2, they needed at least 2\*7 = 14 deaths. For each EPV, they simulated 500 times, after their simulation, they drew a conclusion that when EPV is larger than 10, there will be no problem occurred, when EPV is lower than 10, the bias of our logistic model will increase, and the low-EPV is the main reason that affected the validity of the logistic regression model[10]. Based on the previous research, in the year 2014, another research group examined the influence of EPV on the relative performance of three different methods which are used to access the predictive accuracy of a logistic regression model: apparent performance in the analysis sample, split-sample validation, optimism correction using bootstrap methods. The data they used was a single dataset of patients hospitalized with heart failure, they examined the data by comparing the estimates of discriminatory performance from these methods to those for a large independent validation sample arising from the same population, the results proved their assumption: the apparent performance became better as EPV increased, the difference between the result of bootstrap and validation was minimal when EPV is larger than 20. As a result, they claimed, they encouraged to use modern validation procedures such as bootstrap optimism correction in studies of model development[11]. At the same year, a Chinese medical group examined the quality of multivariable logistic regression model in 316 articles, and set a individual standard for quantify the goodness of each model. Their standard included 13 items: selection of independent variables, a description of fitting procedure, coding of variables, the check of interaction, the check of collinearity, statistical significance, goodness-of-fit, check for the outliers, whether complete identification of the statistical software that was used, whether have sufficient EPV (event per variable), participation of statisticians or epidemiologists, conformity with linear gradient for continuous or ranked variables and statistical analysis. Each standard worth 1 score, as a result of their research, the highest score was 9 and the lowest score was 1, most of the articles achieved less than 6 (85.1%).[12] These previous researches provide a guidance for us in our project, in order to avoid the large error and meaningless result, we are going to build our model in the scientific method and check our result carefully.

**2. data source and research aim**

We try to find out the potential associations between the probability of having side effects and the difference among patients with schizophrenia, the difference includes whether using Clozapine and other basic information such as age, gender, ethnicity and so on. The side effects we work on includes 8 ADRs: Agranulocytosis, Constipation, Hypersalivation, Nocturnal enuresis, Pneumonia, Seizures, Venous thromboembolism and Weight gain. Now we gather the useful data from the electronic medical records of patients with schizophrenia. Database 1 contains about 11000 patients, 6600 of them does not use Clozapine while the rest were treated with it, and we also load the basic information including age, gender, ethnicity, marriage, occupation, religion, first language, mobility, hearing impairment, visual impairment which may be a component that affect the result. We have another database that contains a three-week-time-window of over 40 thousand treating records, in which shows the number of ADRs mentioned by doctors within 3 weeks, from 3 to 6 weeks, from 6 to 9 weeks after treatments.

**3. Model building**

In the model building part, we use SPSS to help us build the logistic regression model and calculate the significant statistics which could prove the efficiency of our model. The most important variable of our model is whether using the Clozapine, figure 2 shows the comparison between patients whether using the Clozapine, the left bar in each plot is the number of patients who had side effects after treating without Clozapine in every 1000 patients, the right bar in each plot is the number of patients with Clozapine, it is obvious that Clozapine played an important role in causing the side effects, patients would face a much more risk of having each kind of ADRs, especially constipation, weight gain, seizures and hypersalivation. The value of this variable is binary, '1' means the patient used Clozapine and '0' otherwise.

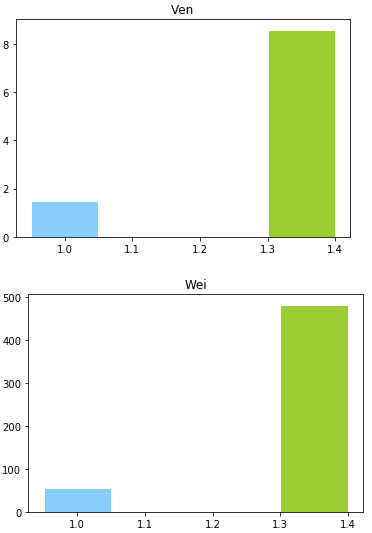
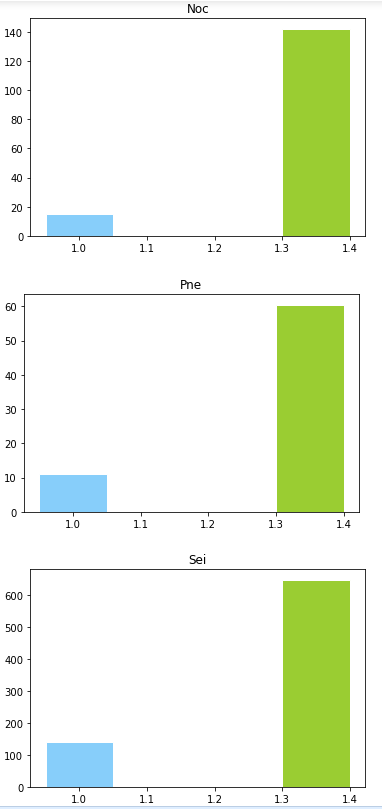
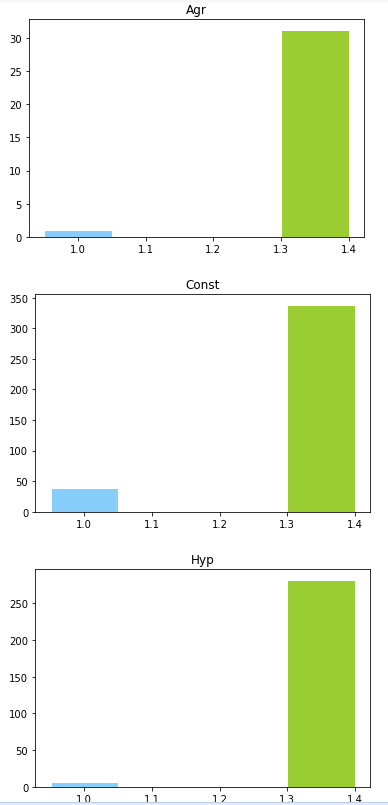


figure 2

The second important component we choose as our original variable is ethnicity, many previous researches have proved that there is a huge difference in constitution between different ethnicities, the risks of having different diseases are also different. The values of ethnicity variable are categorical and the number of different value is 10, therefore, we introduce 9 dummy variables to represent ethnicity. In addition, we also add gender and age as our original variables, since it is almost a common sense that as people getting old, the immune system of our body will become weak, the physical difference between male and female are also obvious and widely known. The method to measure gender is simple, '1' represent male and '0' otherwise, while the representation of age is more complex, despite that age is a numerical variable, we still not sure whether the association between age and risk of having side effects is linear, so we try to use two methods to measure age. The first method, directly consider the relationship is linear. The second method, according to the medical research result, our lives can be divided into three periods: young (age <=25), mid age(age in 25-50) and elder (age >50), in the each period, the function and immune ability are similar, in this method, we introduce two dummy variables, we try both methods in model building and select one with better result. Also, we take the marriage and occupation into consideration, however, we do not use them as variables since the proportions of the occurrence of each side effect under different situation are nearly the same.

When analyzing the known data, we find data in database 1 is better for training than database 2 since the distribution and number of positive and negative events are more evenly, although the number of data strips in database 2 is more than 40 thousand, the proportion of ADRs take place is less than 1%, so we use database 1 for building model and database 2 for checking our model.

As a result, after data preprocessing, we delete the data strip with missing values in variable columns and we get 8 models for each ADR:

The parameters of each variable in each model are not all satisfying, we judge whether an item is reasonable by hypothesis test and calculating the p-value, p-value gives us the probability that the parameter we get is wrong, normally, we require the p-value of each parameter should be lower than 0.05, the reasons cause the large p-value are sundry, maybe there exists collinearity, or the variable is not directly related to the dependent variable. Based on our assumption, we delete the variables which p-value does not meet the statistical requirements in each model and re-build the model, if there are more than one variables with large p-value, we firstly delete the one with larger p-value and re-build the model and check whether there are still some variables with unsatisfying p-value, repeat this step until all variables are significant. For dummy variables, we retain all of them if the overall p-value is lower than 0.05 despite that some of them do not meet the statistical requirement, since dummy variables are meaningless if they are incomplete. After checking and re-building the model, we finally get the following 8 new models:

the models above all meet the statistical requirement, and in order to check whether it is good enough for prediction, we use the data in database 2 to do an informatics test. In informatics domain, we usually calculate the F1 score to show the accuracy of a model, F1 score is a measure that can be calculated by (precision \* recall) \*2 /(precision + recall) where precision = TP/(TP+FP) and recall = TP/(TP+FN). the results is shown below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Agranulocytosis | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 40069 | 35 | 0 | 0 | 0.999 | 1 | 0 | 1 | 0.999 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Constipation | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 39518 | 397 | 1 | 188 | 0.990 | 0.995 | 0.003 | 0.995 | 0.992 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hypersalivation | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 39787 | 317 | 0 | 0 | 0.992 | 1 | 0 | 1 | 0.996 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nocturnal enuresis | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 40038 | 66 | 0 | 0 | 0.998 | 1 | 0 | 1 | 0.999 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pneumonia | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 40055 | 51 | 0 | 0 | 0.999 | 1 | 0 | 1 | 0.999 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Seizure | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 20512 | 26 | 816 | 18750 | 0.998 | 0.522 | 0.969 | 0.522 | 0.686 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Venous thromboembolism | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 40099 | 5 | 0 | 0 | 0.999 | 1 | 0 | 1 | 0.999 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weight Gain | TP | FP | TN | FN | precision | recall | specificity | sensitivity | F1-score |
|  | 29993 | 236 | 321 | 9554 | 0.992 | 0.758 | 0.576 | 0.758 | 0.859 |

It seems that the F1-scores are all in a high level and completely acceptable, but is it true? We can intuitively notice that some model can just report the positive, in fact, the logistic regression model consider each side effect would happen if the result of model is bigger than 0, some models we received from logistic regression such as 'Pne model',

assume that a patient has taken Clozapine and is an elder, the result is still lower than 0, the system will say that he will not have the Pneumonia, this method of prediction has the lowest probability of wrong prediction, however, we think a model that just can provide a single outcome is meaningless, we prefer the higher wrong prediction rate rather than a meaningless model. So, we want to adjust our model again.

The models are convincing in statistical domain, in order to make it better in application, we try to modify the threshold of each model, at present, the model judge whether one ADR would occur by the sign of logit function, as , the threshold for p is 0.5, and it can be explained that if the probability of the occurrence of one ADR is larger than 50%, we consider it will take place, in order to let each model can report two outcomes, we adjust the threshold and try to find a new measure to balance the sensitive and specificity. The bigger the sensitivity and specificity are, the better our model will be, however, from previous researches and informatics source, we could not find a suitable measure which requires true positive rate and true negative rate both stay in a good level, as a result, we have to define a new measure, similar to the calculation of F1-score, we define the new measure N as 2\* specificity \* sensitivity/(specificity + sensitivity), and we keep adjusting the threshold to find the most applicable model with highest N for each ADR.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hypersalivation | TP | FP | TN | FN | precision | recall | specificity | sensitivity | N |
|  | 21877 | 39 | 278 | 17910 | 0.998 | 0.550 | 0.877 | 0.550 | 0.676 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nocturnal enuresis | TP | FP | TN | FN | precision | recall | specificity | sensitivity | N |
|  | 27385 | 25 | 119 | 12575 | 0.998 | 0.685 | 0.826 | 0.685 | 0.749 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Seizure | TP | FP | TN | FN | precision | recall | specificity | sensitivity | N |
|  | 22569 | 132 | 710 | 16693 | 0.994 | 0.575 | 0.843 | 0.575 | 0.684 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Constipation | TP | FP | TN | FN | precision | recall | specificity | sensitivity | N |
|  | 26041 | 92 | 306 | 13665 | 0.990 | 0.656 | 0.769 | 0.656 | 0.708 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weight Gain | TP | FP | TN | FN | precision | recall | specificity | sensitivity | N |
|  | 25654 | 113 | 444 | 13893 | 0.995 | 0.649 | 0.797 | 0.649 | 0.715 |

**4. Analysis**

We fail to find better model for agranulocytosis, Pneumonia and Venous thromboembolism, the reasons are varied, firstly, the number of negative event is too little, the occurrence of Venous thromboembolism is just 5 times among more than 40000 records, keep adjusting the threshold will lead to the large amount of false predictions. Secondly, the models of these three ADRs is not good enough to describe and predict the situation of having side effect, take agranulocytosis as an example, the model says that the occurrence of agranulocytosis is just related to whether using the drug regardless that the patient is young or old, male or female, black or white, and if we adjust the threshold to some number, the result will shows that all patients take Clozapine will have agranulocytosis, it is obviously not realistic and unreasonable, similarly the model of pneumonia and Venous thromboembolism are not good as well. The probable reason lead to a not good model because of the low EPV in training set. The number of events for each ADR among the training set (about 10000 data strips) is:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| agranulocytosis | Constipation | Hypersalivation | Nocturnal enuresis | Pneumonia | Seizure | Venous thromboembolism | Weight Gain |
| 141 | 1671 | 1277 | 699 | 314 | 3512 | 44 | 2387 |

The origin model contains 13 variables, the EPV of Venous thromboembolism even less than 5, agranulocytosis, Pneumonia and Venous thromboembolism have the lowest EPV in our research, and this is the main reason leads to the unsatisfying model.

For those acceptable models, we provide the statistical models and one applicable model for each ADR, actually, the threshold for each ADR is flexible, it depends on what we need. For example, for the fatal and serious side effect, we prefer to predict all of them, even there are many false predictions, we still want to drop the risk to the lowest level, under this circumstances, we need to adjust the threshold until more than 95% of negative events could be recognized. For each ADR which are not fatal, the cost functions for each patient are different, the outcomes and conclusions for each patient may not the same as well.

**5. Discussion**

Throughout our models, there still exists problems, actually, the causes lead to one disease are diverse, such as whether the patient has another disease and takes some other drugs, however, in our research, we cannot take those factor into consideration, because the factors worth consideration are many, with the increase of variables, EPV will become smaller which may leads to the bad result, also, it is not easy to quantify those categorical variables.

Overall, our models are acceptable in the prediction of five ADRs: Constipation, Hypersalivation, Seizure, Weight Gain and Nocturnal enuresis, the more common an ADR is, the better our prediction will be, and our model system could be used in real life.

**6. acknowledge**

The study is supported by Honghan Wu and Kings College Hospital, and we would like to thank Dr Honghan Wu for providing the data.

**7. References**

[1] Clozapine from wikipedia (https://en.wikipedia.org/wiki/Clozapine)

[2]  Crilly, John (2007-03-01).[*"The history of clozapine and its emergence in the US market a review and analysis"*](http://hpy.sagepub.com/content/18/1/39).History of Psychiatry.**18**(1): 39–60.

[3] Leucht, Stefan; Cipriani, Andrea; Spineli, Loukia; Mavridis, Dimitris; Örey, Deniz; Richter, Franziska; Samara, Myrto; Barbui, Corrado; Engel, Rolf R; Geddes, John R; Kissling, Werner; Stapf, Marko Paul; Lässig, Bettina; Salanti, Georgia; Davis, John M (2013). "Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis".*The Lancet*. **382** (9896): 951–62.

[4] Baggaley M (Apr 2008). "Sexual dysfunction in schizophrenia: focus on recent evidence". *Human Psychopharmacology*. **23** (3): 201–209.

[5]  [*"Clozapine"*](https://sites.google.com/site/pharmacologymcqs/clozapine-clozaril)*.*[*Archived*](https://web.archive.org/web/20131110214051/https:/sites.google.com/site/pharmacologymcqs/clozapine-clozaril)*from the original on 2013-11-10.*

[6] Boyd, C. R.; Tolson, M. A.; Copes, W. S. (1987). "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score". *The Journal of trauma*. **27** (4): 370–378.

[7] logistic regression from wikipedia (https://en.wikipedia.org/wiki/Logistic\_regression)

[8] **Jack V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, In Journal of Clinical Epidemiology, Volume 49,**

**Issue 11, 1996, Pages 1225-1231, ISSN 0895-4356**

[9] Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. pp. Chapter 3, page 45

[10] **Steven C Bagley, Halbert White, Beatrice A Golomb, Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain, In Journal of Clinical Epidemiology, Volume 54, Issue 10, 2001, Pages 979-985, ISSN 0895-4356**

[11] **Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, Alvan R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, In Journal of Clinical Epidemiology, Volume 49, Issue 12, 1996, Pages 1373-1379, ISSN 0895-4356**

[12] **Quality of reporting of multivariable logistic regression models in Chinese clinical medical journals.** [Zhang YY](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhang%20YY%5BAuthor%5D&cauthor=true&cauthor_uid=28538397)1, [Zhou XB](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhou%20XB%5BAuthor%5D&cauthor=true&cauthor_uid=28538397), [Wang QZ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wang%20QZ%5BAuthor%5D&cauthor=true&cauthor_uid=28538397), [Zhu XY](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhu%20XY%5BAuthor%5D&cauthor=true&cauthor_uid=28538397)., [Medicine (Baltimore).](https://www.ncbi.nlm.nih.gov/pubmed/28538397) 2017 May;