Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 2 | Article 7

HOAJ
Herbert Open Access Journals

Methodology Open Access

Risk of performing multiple logistic regression analysis without considering multiplicity: an overview for clinicians and practitioners

Tomoyoshi Tsuchiya

Correspondence: ttom@shimizuhospital.com

Department of Respiratory Medicine, Shizuoka city Shimizu Hospital, 1231 Miyakami, Shimizu-ku, Shizuoka, Japan.

Abstract

Background: In many of clinical studies, a lot of explanatory variables are analyzed and it is concluded that all statistically significant variables are important. However, if the multiplicity of statistical tests is not considered, significant variables are determined only by chance. To demonstrate the risk of multiple hypothesis tests, multiple logistic regression models created by random numbers are simulated.

Methods: Variables y and x1~x30, which have 600 elements per variable, are created by numbers selected randomly from (0,1) in the re-sampling method. Variable y is defined as the objective variable and variables x1~x30 are defined as explanatory variables. Multiple logistic regression analysis is performed using those objective and explanatory variables. Wald tests are performed in the statistical model, and the number of statistically significant explanatory variables (p value<0.05) is counted. The series of analysis is repeated 1000 times, and the numbers of significant variables are summated.

Results: In the 1000 simulations, the number of significant explanatory variables is 0~8 per one analysis. The average number is 1.69, and the median number is 2. In 80.1 percent of all of the simulations, at least one or more explanatory variables become statistically significant. Fifty percent or more simulations in all, the explanatory variables of two or more are statistically significant.

Conclusions: When performing exploratory research using multivariable analysis, we must be fully aware that there is a risk of false significance by multiplicity.

Keywords: Linear models, logistic regression, multivariate analysis, research methodology

Introduction

In a great number of clinical studies, the multiple logistic regression model is used to investigate the prognostic factors by an exploratory method. In many of these studies, a lot of explanatory variables are analyzed and it is concluded that all statistically significant variables (p value<0.05) are important. However, if the multiplicity of statistical tests is not considered, significant variables are determined only by chance and a risk arises of incorrect conclusions.

When mean values of more than three groups are compared, adjusting of multiple comparisons is commonly taken into consideration (i.e., Bonferroni procedure, Holm procedure). However, multiple hypothesis testing in the multiple logistic regression model is not discussed at all. It is possible that explanatory variables, which are described as important factors in published papers, are in fact meaningless. It is common in clinical research for a lot of factors associated with the development of certain diseases, for example, gender, age, smoking history, hypertension and so on, to be examined by performing multiple hypothesis tests. In such instances, by multiplicity, the factors that have p values actually much larger than 0.05 may be judged as significant.

If we test only one null hypothesis using 0.05 as cut off point of significance, it is correct to regard a p value less than 0.05 as statistically significant. However, if we concurrently test two independent null hypotheses, the probability that at least one will be significant is 1-(1-0.05)x(1-0.05)=0.098, not 0.05. If we test 10 such hypotheses, the probability that at least one of those will be significant is $1-(1-0.05)^{10}=0.40$, which is much larger than 0.05. Generally, if we perform k independent significant

doi: 10.7243/2053-7662-2-7

tests with the cut-off point 0.05, the probability that at least one of k hypotheses will be significant is $1-(1-0.05)^k$.

In this study, for the purpose of demonstrating the risk of multiple hypothesis tests, I simulated multiple logistic regression models created by random numbers.

Methods

Variables, y and x1~x30 which have 600 elements per variable, are created by numbers selected randomly from (0,1) in the re-sampling method (**Table 1**). Variable y is defined as the objective variable and variables x1~x30 are defined as explanatory variables, and then multiple logistic regression analysis (a generalized linear model with logit link function and binomial error structure) is performed using the objective and explanatory variables. If the probability of an interested event is p, the odds is defined as p/(1-p). The multiple logistic regression model is given by:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where β_i indicate the partial regression coefficients associated with the reference (β_0 is the intercept) and x_i indicate explanatory variables.

Table 1. Data set image created by numbers selected randomly from (0,1) in the re-sampling method. Variables, y and x1~x30 have 600 elements per variable.

Data	y	x1	x2	x3	...	x29	x30
No 1	0	1	0	1	...	1	1
No 2	1	0	0	0	...	1	1
No 3	0	0	0	1	...	0	1
...	...	...	...	...	...	...	...
No 599	1	0	1	0	...	0	1
No 600	1	1	0	1	...	0	0

This is the same procedure commonly described in medical research papers as 'We investigated using a multiple logistic regression analysis of 30 factors within 30 days of death in 600 cases of a certain syndrome'.

Conducting a Wald test on the partial regression coefficient of multivariable analysis, the number of statistically significant explanatory variables (p value<0.05) is counted. The series of analysis is repeated 1000 times, and the numbers of significant variables are summed.

All of the analyses are conducted using R version 3.1.0 (R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/). The R script used in this simulation is shown in the **Supplement Data** script list.

Results

One output example of the multiple logistic regression analysis of 1000 simulations is shown in **Table 2**. In this example, four of the 30 explanatory variables are determined to be significant (p value<0.05). The analysis processes are repeated 1000 times in the same way, and 1000 analysis results are outputted then summated.

In those 1000 simulations, the number of significant explanatory variables (p value<0.05) is 0~8 per one simulation (**Figure 1**). The average number is 1.69, and the median number is 2. In 80.1 percent of all of the simulations, at least one or more explanatory variables become statistically significant. In 50% or more simulations, two or more of the explanatory variables are statistically significant.

In order to confirm that the simulation models are created from random numbers, histograms of the adjusted odds ratio

Table 2. One-sample outputs of 1000 simulations. There are four significant variables in this analysis.

Variables	Adjusted Odds ratio	Lower limit (95%CI)	Upper limit (95%CI)	p value
X1	1.0361771	0.7385793	1.4536869	0.837000859
X2	0.9255818	0.6598432	1.2983414	0.654244721
X3	0.9986820	0.7119090	1.4009736	0.993906869
X4	1.1451813	0.8181678	1.6028988	0.429422806
X5	0.8550973	0.6075526	1.2035030	0.369345435
X6	0.7161831	0.5101159	1.0054936	0.053815719
X7	1.2484785	0.8921901	1.7470476	0.195480444
X8	0.7462226	0.5301450	1.0503695	0.093301760
X9	1.0210596	0.7291180	1.4298958	0.903457049
X10	0.8492224	0.6037984	1.1944031	0.347655106
X11	1.4939865	1.0634999	2.0987269	0.020613735*
X12	0.8907518	0.6324919	1.2544649	0.507822772
X13	1.0607610	0.7543560	1.4916220	0.734490154
X14	0.9892709	0.7059888	1.3862216	0.950030625
X15	0.8890902	0.6343551	1.2461181	0.494919832
X16	0.9112185	0.6501722	1.2770759	0.589302868
X17	0.7000925	0.5004476	0.9793824	0.037379660*
X18	1.1750798	0.8390807	1.6456254	0.347772921
X19	1.0859406	0.7715992	1.5283414	0.636315816
X20	1.0439284	0.7451216	1.4625618	0.802676958
X21	0.5666774	0.4033391	0.7961620	0.001060473*
X22	0.8346742	0.5965953	1.1677614	0.291532845
X23	1.1365872	0.8123382	1.5902619	0.454990820
X24	0.9826672	0.6993162	1.3808272	0.919754581
X25	0.9684741	0.6878587	1.3635678	0.854400542
X26	0.7097133	0.5055499	0.9963270	0.047566024*
X27	0.8933293	0.6380390	1.2507657	0.511244623
X28	1.1624144	0.8290538	1.6298184	0.382782848
X29	0.9268180	0.6621762	1.2972250	0.657753520
X30	1.1639330	0.8247425	1.6426219	0.387757558

The * indicates statistically significance (p<0.05);
CI: confidence interval

doi: 10.7243/2053-7662-2-7

for all explanatory variables (30x1000=30000) and p values for all explanatory variables (30x1000=30000) are developed (**Figures 2** and **3**). With the histogram of the adjusted odds ratio, about 1.0 is the highest, and there are few values by chance alone that will be 2.0 or more and 0.5 or less. The histogram of the p value shows a uniform distribution.

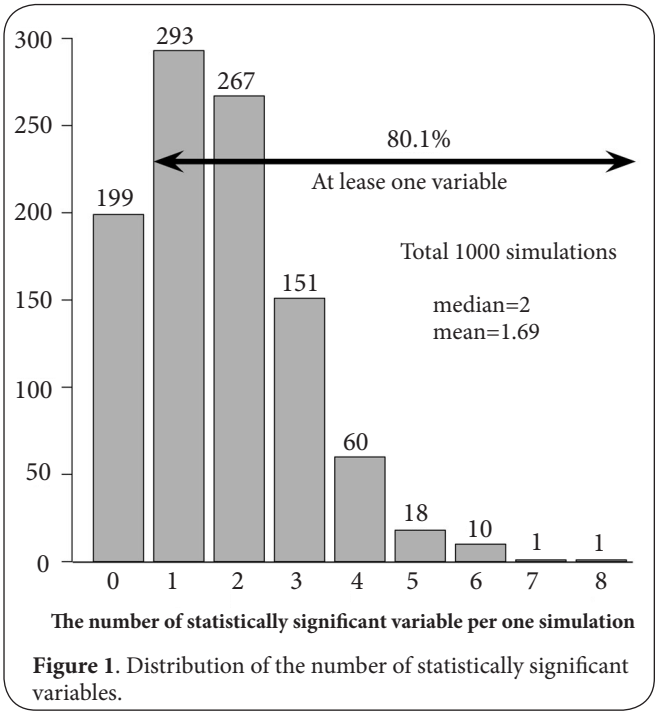

Figure 1. Distribution of the number of statistically significant variables.

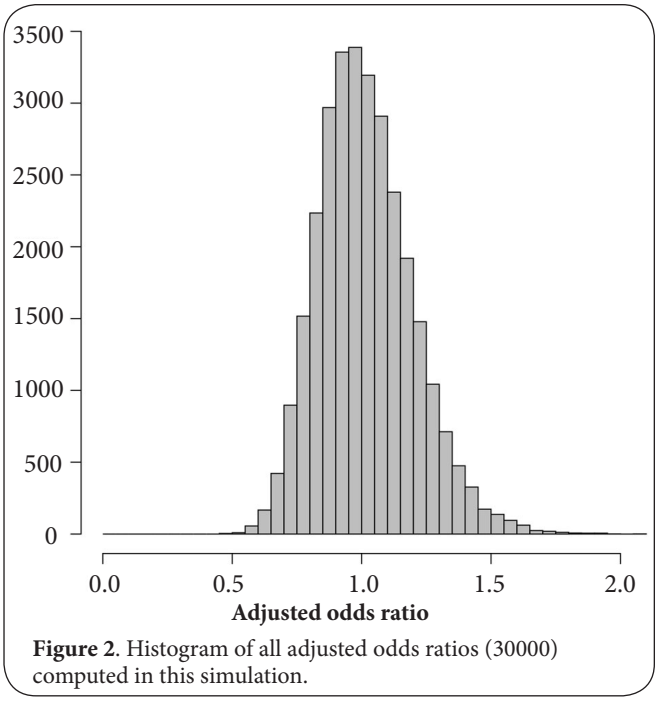

Figure 2. Histogram of all adjusted odds ratios (30000) computed in this simulation.

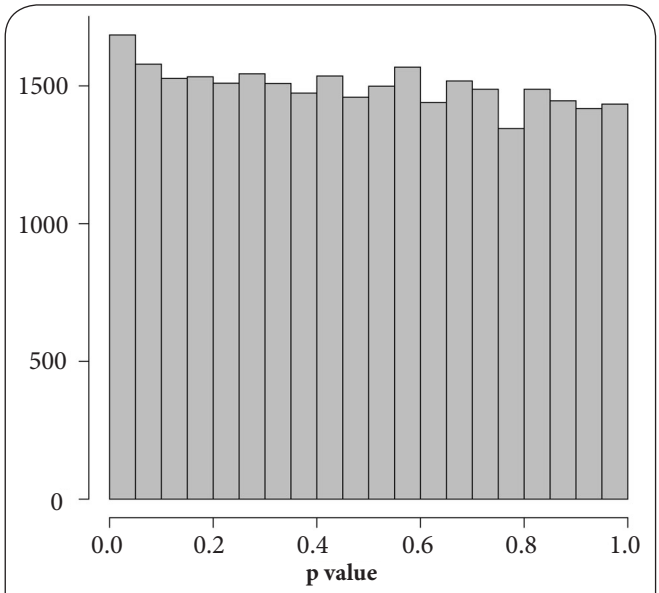

Figure 3. Histogram of all p values (30000) computed in this simulation.

Discussion

The current study demonstrates the risk of multiple hypothesis tests in exploratory clinical research. There are a few medical papers about this point [**1**-**3**], but these papers are not necessarily easy for clinicians to understand. Freedman showed that in multiple linear regression analysis using data created from random numbers, significant variables emerge from pure noise [**4**]. To develop upon that concept I present for clinicians the risk of multiple hypothesis tests in a visible manner with multiple logistic regression analysis, which is commonly used in medical research. Researchers are preoccupied with demonstrating statistical significance for publication, and may often lose the essence of their research.

As indicated in this study, statistically significant variables can be calculated using only noise, i.e., completely random data. For 30 variables, one or more variables are significant with a probability of about 80% from chance alone. Incidentally, a maximum of eight variables may be statistically significant with completely random data. If dummy-coded multiple categorical data are used, 30 explanatory variables is not such a large number. When evaluating only one null hypothesis in confirmatory study, there is no problem. However, when by evaluating a lot of null hypotheses in exploratory research, one must question whether statistically significant factors are really meaningful or not. A chance of 1 in 20 times happens quite often. Researchers are misunderstood in many cases. Confounding can be corrected by the multivariable analysis, but the chance of random is not corrected for. Important results and falsely important results by random chance are mixed in the same analysis and it is not always easy to distinguish between them.

There are several ways to avoid the risk of multiple hypo-

thesis tests. These are described in "Evaluating Clinical and Public Health Interventions" [5] as follows: (1) In accordance with the number of null hypotheses, we will adjust the significance level by the method of Bonferroni and Holm. (2) We will determine how many factors to analyze before starting the study, and will determine the main outcome. There is no need to adjust testing the primary outcome; but it is necessary to adjust secondary outcome (s). (3) We will describe clearly all the results of tests in the paper. By doing so, the readers can interpret the results properly, without adjustment. (4) We should not be swayed by the p value itself. It is not important whether p value of the test is a little under or a little over the significance level. We must sufficiently investigate the biological plausibility, consistency with findings of other studies, effect size, and so on, without being too interested in p value itself. In addition to the above, the simulation in this study indicated that an adjusted odds ratio is almost never greater than 2.0 or less than 0.5 by chance alone. Therefore, if the adjusted odds ratio exceeds this range, the results of the test are highly likely to be important.

Conclusions

Performing multiple logistic regression analysis with 30 explanatory variables, one or more variables are significant with a probability of about 80% from chance alone. A maximum of eight of 30 variables may be significant from chance alone. When performing exploratory research using multivariable analysis, it is necessary for us to be fully aware of the risk shown in this study. In interpreting the results, we should try to reduce the risk of multiple hypothesis tests.

Additional files

Supplement Data

Competing interests

The author declares that he has no competing interests.

Acknowledgement and funding

I thank the editors and reviewers for their helpful comments. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Publication history

Editor: Qiang Shawn Cheng, Southern Illinois University, USA.
EIC: Jimmy Efird, East Carolina University, USA.
Received: 10-Jul-2014 Final Revised: 31-Aug-2014
Accepted: 08-Sep-2014 Published: 17-Sep-2014

References

1. Smith DG, Clemens J, Crede W, Harvey M and Gracely EJ. **Impact of multiple comparisons in randomized clinical trials**. *Am J Med.* 1987; **83**:545-50. | Article | PubMed
2. Mills JL. **Data torturing**. *N Engl J Med.* 1993; **329**:1196-9. | Article | PubMed
3. Berry D. **Multiplicities in cancer research: ubiquitous and necessary evils**. *J Natl Cancer Inst.* 2012; **104**:1124-32. | Article | PubMed
4. David A Freedman. **A note on screening regression equations**. *The American Statistician*. 1983; **37**:152-5. | Article
5. Mitchell H. Katz. **How do I adjust for multiple comparisons? Evaluating Clinical and Public Health Interventions: A Practical Guide to Study Design and Statistics**. 2010; 140-2.

Citation:
Tsuchiya T. **Risk of performing multiple logistic regression analysis without considering multiplicity: an overview for clinicians and practitioners**. *J Med Stat Inform*. 2014; **2**:7.
http://dx.doi.org/10.7243/2053-7662-2-7