Letter to the Editor

Too much covariates in a multivariable model may cause the problem of overfitting

Zhongheng Zhang

Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China
Correspondence to: Zhongheng Zhang. Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China. Email: zh_zhang1984@hotmail.com.

Submitted Aug 09, 2014. Accepted for publication Aug 11, 2014.
doi: 10.3978/j.issn.2072-1439.2014.08.33
View this article at: http://dx.doi.org/10.3978/j.issn.2072-1439.2014.08.33

To the editor,

Many thanks for the thoughtful insights into our work by Prof. Zhang and coworkers. We strongly agree with the reader that red blood cell distribution width (RDW) can be influenced by varieties of medical conditions including but not limited to anemia, renal dysfunction, hepatic dysfunction, thyroid disease, transfusion, acute or chronic inflammation, neurohumoral activation, malnutrition (i.e., iron, vitamin B12 and folic acid), ethnicity, bone marrow depression, and use of some medications (1-4). From the perspective of controlling for confounders, incorporation of a large number of covariates into a regression model will make the independent association more reliable. As a result, some investigators suggest incorporate as much covariate as possible when the study is aiming to explore the association between a variable of interest and clinical outcome. However, the benefit of including too many covariates should be balanced with the problem of overfitting (5,6). Overfitting occurs when too many variables are included in the model and the model appears to fit well to the current data. Because some of variables retained in the model are actually noise variables, the model cannot be validated in future dataset. In essence, overfitting is caused by multiple testing in which some noise variables are entered into the model simply by chance.

Another reason that we did not incorporate so many confounding factors had something to do with technical issues. The study was a retrospective study and involved strenuous work on data extraction from electronic medical record (EMR) system. The EMR was not designed for research purpose but instead it was used for clinical practice. Some information may not be very reliable in such circumstance. For instance, the use of medications in past history may not be complete that some drugs may be omitted because it was thought to be unrelated to current disease. Furthermore, the information related to previous drug use was recorded as text, which imposed great challenge on data mining.

Finally, we acknowledge that confounding factors have not been thoroughly explored in our study and it is one of the limitations (7). As I have pointed out previously, confounding is the Achilles' heel in observational studies (8). The ultimate solution to the problem may be the randomization which, when performed in an infinitely large sample size, can balance both known and unknown confounders and make the association between the variable of interest and outcome reliable. With respect to the time interval between blood sampling and laboratory analysis, I feel sorry I cannot provide enough information for analysis (the time was not recorded in EMR). The time was actually determined by the availability of transport workers. When they are busy, the blood sample may be delayed for half an hour. However, the blood sample can be delivered to the department of laboratory within 10 minutes in most circumstances. Prof. Zhang has mentioned the impact of admission source (emergency department *vs.* floor ward) on the level of RDW. However, there is no empirical evidence to support this notion and I feel that it is trival. Additionally, many patients from emergency department are transferred from other hospitals, making them equivalent to those transferred from floor wards of our hospital.

Acknowledgements

Disclosure: The author declares no conflict of interest.

References

1. Montagnana M, Lippi G, Targher G, et al. The red blood cell distribution width is associated with serum levels of thyroid stimulating hormone in the general population. Int J Lab Hematol 2009;31:581-2.
2. Núñez J, Núñez E, Rizopoulos D, et al. Red blood cell distribution width is longitudinally associated with mortality and anemia in heart failure patients. Circ J 2014;78:410-8.
3. Ujszaszi A, Molnar MZ, Czira ME, et al. Renal function is independently associated with red cell distribution width in kidney transplant recipients: a potential new auxiliary parameter for the clinical evaluation of patients with chronic kidney disease. Br J Haematol 2013;161:715-25.
4. Yang W, Huang H, Wang Y, et al. High red blood cell distribution width is closely associated with nonalcoholic fatty liver disease. Eur J Gastroenterol Hepatol 2014;26:174-8.
5. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004;66:411-21.
6. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004;44:1-12.
7. Zhang Z, Xu X, Ni H, et al. Red cell distribution width is associated with hospital mortality in unselected critically ill patients. J Thorac Dis 2013;5:730-6.
8. Zhang Z. Confounding factors in observational study: The Achilles heel. J Crit Care 2014;29:865.

Cite this article as: Zhang Z. Too much covariates in a multivariable model may cause the problem of overfitting. J Thorac Dis 2014;6(9):E196-E197. doi: 10.3978/j.issn.2072-1439.2014.08.33