Journal of Paediatric Respirology and Critical Care

Common pitfall in statistical analysis: KISS principle

Chung-Hong CHAN 陳仲康 and Daniel Kwok-Keung NG 吳國強
Department of Paediatrics, Kwong Wah Hospital, Hong Kong

Randomised controlled trial (RCT) is often regarded as providing better evidence than cohort study and case control study. The reason for RCT to be the gold standard is not commonly discussed, i.e. RCT allows extremely simple statistical analysis. As the distribution of potential confounding factors is evenly distributed for each treatment arm by random allocation, one can directly compare the efficacy by simple bivariate analysis such as t-test. We don't need to adjust for the confounding factors using multivariate analysis such as multiple regression that increases the complexity of statistical analysis. Complex statistical analysis requires higher level of technical knowledge of the analysts as well as more difficult for the readers to understand. The greatest discoveries in recent history of medicine were usually analysed by very simple statistical methods. "Keep It Small and Simple" (KISS) principle should be borne in mind during study design and data analysis planning. We should aim at very simple analysis that displays the relationship elegantly.

The problems of overly complex statistical methods included:

1. Not universally available in all statistical packages. For example, Bayesian analysis is not available in SPSS
2. The results of complex statistical method may not be consistent between different statistical packages because they use different algorithms, for example, linear mixed effect model in SAS, SPSS and R give different results for the same data.
3. Not easy to interpret
4. Not easy to understand

Hence, it is important to design a study well instead of using complex statistics to compensate for a flawed study design. One should always use the simplest method of analysis as illustrated by the example listed below.

We had six asthmatic children who performed the morning peak expiratory flow rate (PEFRam) and asthma control test (ACT) score daily for 8 days. We would like to demonstrate the relationship between PEFRam and ACT score. A straightforward way is to plot the 46 readings (8 days x 6 subjects) of PEFRam and ACT scores and determine the correlation coefficient. However, this method is statistically invalid due to the fact that the assumptions of independence of observations and homogeneity of variance are completely violated because the 48 points on the scatterplot were not provided by 48 subjects. In order to adjust for this, we may introduce a new dummy variable called "subject" and adjust for the effect of "subject" by multiple regression. But it is still faulty. In order to adjust for repeated measurement of PEFR and ACT within each subject, we need to use the linear mixed effect model. Linear mixed effect model is the standard answer to this problem in biostatistics textbook. But it is too complex and the algorithm is not universally accepted.

By KISS principle, a very simple way is to deduce a correlation coefficient of PEFRam and ACT for EACH subject. Then we have six correlation coefficients. We can calculate the sample mean of these six correlation coefficients and use a very simple one-sample t-test to determine the population mean correlation coefficient. If the projected population mean is more than zero. We could than conclude that PEFRam is correlated with ACT scores.

We can exploit the KISS principle further by using an even simpler method. We can calculate the mean PEFRam and mean ACT from the 8 readings for each subject. And then we correlate the 6 pairs of mean PEFRam with mean ACT.

Computer exercise:

In SPSS with our data, we can obtain the correlation coefficient for each subject by the following code:

```
SORT CASES BY subj_no .
SPLIT FILE
  SEPARATE BY subj_no .
CORRELATIONS
  /VARIABLES=pfr act
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

The variable "cor" is the correlation coefficient from previous analysis. We can test the mean coefficient correlation is significantly different from zero by the following code:

```
SPLIT FILE
  OFF.
T-TEST
  /TESTVAL=0
  /MISSING=ANALYSIS
  /VARIABLES=cor
  /CRITERIA=CIN (.95) .
```

We can obtain the mean PEFRam and mean ACT for each subject by the following code:

```
SORT CASES BY subj_no .
SPLIT FILE
  SEPARATE BY subj_no .
DESCRIPTIVES
  VARIABLES=pfr act
  /STATISTICS=MEAN STDDEV .
```

The variable "mean_pfr" and "mean_act" are mean PEFRam and mean ACT from previous analysis. We can calculate the correlation between the two by:

```
SPLIT FILE
  OFF.
CORRELATIONS
  /VARIABLES=mean_pfr mean_act
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Answers to Radiological Quiz on page 11

Answer: C

Laryngoscopy in the infancy period together with the histology confirmed the diagnosis of laryngeal papillomatosis. In view of the subsequent extensive pulmonary involvement, the child was referred to our unit for interferon treatment in addition to the laser therapy.

CT films detected numerous pulmonary nodules with cavitations, suggestive of papillomatosis involvement. Furthermore intramural papillary lesions are also visualised in the trachea.

A baby can acquire the infection during the birth process when he/she is born through vaginal route, if there is genital wart. Most infected babies are asymptomatic after birth and the papillomas will grow gradually. Thus all those babies should be followed up to monitor any hoarseness of voice or other symptoms such as stridor.

Laser therapy is indicated for the localized lesions. For those patients with widespread pulmonary lesions, the adjuvant therapy using alpha-interferon injection was considered previously. However due to questionable efficacy of the interferon treatment as well as the possible neurological complication of spastic diplegia, the interferon therapy is not treatment of choice now. The current main adjuvant therapy is the administration of anti-viral agent, the newest one being cidofovir.