Aud Vest Res (2017);26(1):1-3.

EDITORIAL

The values of effect size in statistical decision for clinical research

Mohd Normani Zakaria*

Audiology Programme, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Malaysia

Introduction

Hypothesis testing is regarded as an essential statistical analysis in clinical research. To answer the research objectives and make conclusions, the p values are commonly reported to determine whether the findings are likely due to chance. Lower p values (e.g.<0.05) indicate that the difference between the groups is real and does not occur by chance. While the hypothesis testing offers known benefits in statistical analyses, it does have some drawbacks worth to be highlighted. As suggested by many statisticians and researchers, in conjunction with hypothesis testing, authors are encouraged to report effect size, particularly in clinical research.

Advantages of effect size

Unlike p value, effect size provides information on the magnitude of effect in the given samples. Depending on the variables involved, there are several types of effect sizes. Due to simplicity, Cohen's effect size (d) is perhaps the most commonly reported effect size. For interpretation, effect sizes of 0.2, 0.5 and 0.8 are categorized as small, medium and large, respectively.

Consider auditory brainstem response (ABR) data recorded from young males and females, as shown in Tables 1 and 2. The mean and standard deviation for wave V latency for both male (n=10) and female (n=10) groups are revealed in Table 1. As shown by an independent t test, males are found to produce a significantly longer mean wave V latency than females (p=0.02). This finding is further supported by a big Cohen's effect size (d=1.18). This demonstrates that the gender difference in wave V latency is genuine and 20 subjects are sufficient to obtain the desired statistical results.

Table 2 shows the mean and standard deviation of wave V amplitude for both female (n=10) and male (n=10) groups. Even though females revealed a higher mean amplitude (0.35 µv) than males (0.26 µV), this difference is found to be insignificant (p=0.10). Nevertheless, the Cohen's effect size is large (d=0.81), indicating that the difference between the two groups is large. In this regard, if only the p value considered, a conclusion such as "no significant difference in wave V amplitude was found between females and males" would be made. This conclusion, nevertheless, can be debatable as the effect size is large (that indicates the difference between the two groups does exist).

On the other hand, as shown in Table 2, if all values are kept constant but more subjects are added (20 males and 20 females), the p value is now significant (0.01) and the large effect size remains (d=0.81). This clearly shows that the effect size is not affected much by the sample size and the true difference (if any) can be revealed even with a small sample size. The p value, on the contrary, is highly affected by the sample size and insignificant results can be obtained if the sample size is small. To achieve statistically significant results (p<0.05), the sample size must be sufficiently large.

* **Corresponding author:** Audiology Programme, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, 16150, Malaysia.
Tel: 00609-7677691, E-mail: mdnorman@usm.my

Table 1. Analysis of wave V latency of auditory brainstem response (ABR)

	Data entry				Raw difference						Standardized effect size					
	Male		Female		Pooled standard deviation	P-value for difference in SDs	Mean difference	P-value for mean diff (2-tailed T-test)	Confidence interval for difference		Effect size	Bias corrected (Hedges)	Standard error of E.S. estimate	Confidence interval for effect size		Effect size based on control group SD
	n	Mean (SD)	n	Mean (SD)					Lower	Upper				Lower	Upper	
Wave V Latency	10	6.91 (0.26)	10	6.62 (0.23)	0.25	0.36	0.29	0.02	0.06	0.52	1.18	1.13	0.48	0.19	2.08	1.26

The above examples reveal one clear advantage of effect size, i.e. good study outcomes can still be obtained even when the sample size is small. From a practical viewpoint, getting large sample sizes can be laborious, particularly in clinical research. Factors such as the nature of data collection, funding and dropout rate should be considered.

For a study that utilizes hearing screening tool (e.g. otoacoustic emission) for collecting the data, getting a large sample size is easier as the testing time is only a few minutes per subject. In contrast, it is unrealistic to have a large sample size when the subjects are tested with the diagnostic ABR (as the testing time can be at least one hour per subject). As researchers, having large research grants is the ultimate aim. However, some research grants are difficult to procure and the funding for research may not be sufficient. Inevitably, this limits the recruitment of a higher number of subjects in the research. The dropout rate is also an important factor and it is perhaps more prominent in experimental research, in which the effectiveness of new treatment methods is studied. Since frequent visits to respective research centres or clinics are typically required for completing the data collection, subjects may not be able to give full commitments leading to incomplete data and small sample size.

When conducting clinical research, getting “clinical significant” outcomes is more favorable than having only “statistical significant” results. The hypothesis testing is clearly about getting statistical significant outcomes and the data may not be beneficial clinically. Even though debatable, the effect size is a better option and the data may be used for making clinical

Table 2. Analysis of wave V amplitude of auditory brainstem response (ABR)

	Data entry				Raw difference						Standardized effect size					
	Female		Male		Pooled standard deviation	P-value for difference in SDs	Mean difference	P-value for mean diff (2-tailed T-test)	Confidence interval for difference		Effect size	Bias corrected (Hedges)	Standard error of E.S. estimate	Confidence interval for effect size		Effect size based on control group SD
	n	Mean (SD)	n	Mean (SD)					Lower	Upper				Lower	Upper	
Wave V Amplitude	10	0.35 (0.12)	10	0.26 (0.10)	0.11	0.31	0.09	0.10	-0.02	0.20	0.81	0.78	0.49	-0.18	1.73	0.90
Wave V Amplitude	20	0.35 (0.12)	20	0.26 (0.10)	0.11	0.22	0.09	0.01	0.02	0.16	0.81	0.80	0.33	0.15	1.44	0.90

judgments.

For a study to be recognized and valid, its type II error should be low, so that its statistical power would be high (>0.80). Having high statistical power implies that the difference between the groups (if any) is genuine. To calculate the statistical power of a particular study, the effect size is required. The p values, on the other hand, are only related to type I error (incorrect rejection of a true null hypothesis). For an optimum decision making, both p values and statistical power should be taken into account. Furthermore, depending on other variables, in cases where the p values are insignificant (>0.05), even medium effect sizes would produce high statistical power (>0.80). This further supports the superiority of effect size in the data analysis.

Conclusion

Based on the aforementioned points, having the effect size is clearly useful in statistical decisions. Thereupon, I urge readers, researchers and colleagues to include effect size (whenever applicable) to support decision making in research.