Letter to the Editor

Chance, P value and sample size

Madam, We read with interest the article by Zaman et al. in which they compared the efficacy of low (50 mCi) and high dose (100 mCi) idodine-131 for ablation of differentiated thyroid cancer remnants.[1] We feel fundamental information related to the trial is missing and the conclusions may be misleading. Correct interpretations of trial results are dependant on appropriate use of statistical methods to analyse the data. These issues should be carefully considered before commencing a trial and clearly outlined in the protocol. There are no reports of statistical calculations in the study and the conclusions are clearly facile. A clinically worthwhile difference between the two groups was not pre-specified and no evidence against the findings being merely a chance is given. We calculated the p value for this trial to be 0.343, way above the conventionally accepted alpha threshold of 0.05.

Null hypothesis for this study is: there is no difference between group A (high dose) and group B (low dose) in proportion of patients who achieve incomplete ablation. A control trial sets out to prove this. Any differences between the two groups may be due to chance (null hypothesis) and this probability is denoted by the p value. In controlled trials we conventionally accept a 5% risk of finding a difference when none exist (alpha or Type I error). When the p value of a study lies above the alpha threshold ($p>0.05$), the results are statistically not significant; as is the case in this trial. However, it is possible that genuine differences do exist but the study was not powerful enough to detect them. For example, out of initial 22 small trials of anti-thrombotic therapy compared to placebo, 17 showed non-significant statistical results, even though in a subsequent metanalysis (in which results were pooled from these trials) about 20% reduction in early mortality ($p<0.01$) was reported.[2] Type II error (beta statistics) is the probability of making this mistake. This is usually set at 20%. The power of a study is 1-- beta and is the probability of avoiding type II error. This error can be minimised by appropriate a priori sample size computations.

To calculate a sample size for this study, in addition to alpha and beta statistics, additional parameter i.e. smallest clinically worthwhile difference between the comparison groups that the authors would like to detect, is also required. This minimum useful difference is based on expert opinion and clinical judgement. There is no statistical test to compute this difference. A large sample size would be required to detect a small minimum useful difference. Similarly, as the value of alpha or beta decreases, the sample size increases. Furthermore, for sub-group analysis within a trial, as in this study, a larger sample size would be required to avoid type II error. CONSORT guidelines on reporting randomised controlled trials clearly specify the reporting of sample size calculations.[3]

In conclusion, the authors should have reported appropriate statistics to support their conclusions. Various web-based statistical calculators are available to help researchers.[4] They should endeavour to seek advice from a medical statistician before designing a study.

M. A. Rehman Siddiqui
NHS Grampian, University of Aberdeen, and Health Services Research Unit, University of Aberdeen, AB25 2ZN, United Kingdom.

References

1. Zaman M, Toor R, Kamal S, Maqbool M, Habib S, Niaz K. A randomized clinical trial comparing 50mCi and 100mCi of iodine-131 for ablation of differentiated thyroid cancers. J Pak Med Assoc 2006;56:353-6.
2. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45-57.
3. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001;357:1191-4.
4. http://www.cebm.utoronto.ca/practise/ca/statscal/ (Accessed 20th December 2006)
5. John E. Sample size estimation. [online] 2003 [cited 2006 December 20]. Available from: URL: http://www.rad.jhmi.edu/jeng/javarad/samplesize/ (Accessed 20th December 2006).

Reply by Author

Madam, Below is the reply to the comments raised by Dr. Rehman Siddiqui on our article:

A randomized clinical trial comparing 50mCi and 100mCi of iodine-131 for ablation of differentiated thyroid cancers. J Pak Med Assoc.

Statistical Analysis

Statistical analysis was done by analyzing the data on SPSS version 10.0. Chi square test was used to calculate the level of significance (p-value) for all the groups except Group A2 (100 mCi Follicular Carcinoma group), for which Fischer's exact test was used.

GROUP A (HIGH DOSE)

In Group A, out of 20, 14 patients had follow-up whole body 1-131 scan negative and 6 patients had positive whole body 1-131 scan on follow up. Hence the percentage yield is 70/30% with a p-value of 0.074. 13 patients had papillary Ca. on histopathology out of these, 9 patients had a follow-up whole body 1-131 scan negative and remaining 4 patients had a positive WBIS. Hence the percentage yield is 70/30% wit a p-value of 0.166. Remaining 7 patients had follicular Ca. on histopathology and out of them 5 had a negative and 2 had positive follow-up whole body. Hence the percentage yield is 70/30% with a p-value of 0.664.

Six month follow-up of serum Thyroglobulin level in Group A shows undetectable Tg in 14/20 (70%) patients while remaining 6/20 (30%) patients had serum T g in detectable range. Hence the percentage yield is 70/30% wit a p-value of 0.782. 7/13 (54%) papillary Ca. cases had undetectable serum and 6/13 patients had detectable serum Thyroglobulin with a p-value of 0.782. While 7/7 (100%) of follicular Ca. had undetectable serum Thyroglobulin and none of follicular Ca. patients had detectable serum thyroglobulin with a significant p-value of 0.044.

GROUP B (LOW DOSE)

In Group B, out of 20, 10 patients had follow-up whole body 1-131 scan negative and 10 patients had positive whole body 1-131 scan on follow up. Hence the percentage yield is 50/50% with a p-value of 1.0. Ten patients had papillary Ca. on histopathology and out of these, 7 had a negative follow-up WBIS and 3 had positive WBIS. Hence the percentage yield is 70/30% with a p-value of 0.206. 10 patients had follicular Ca on histopathology and out of these 3 had follow-up whole body 1-131 scan negative and 7 had positive follow up WBIS. Hence the percentage yield is 30/70% with a p-value of 0.207.

Six month follow-up of serum Thyroglobulin level in Group B shows undetectable Tg inll/20 (70%) patients while remaining 9/20 (30%) patients had __rum Tg in detectable range. Hence the percentage yield is 55/45% wit a p-value of 0.655.

5/10 (50%) papillary Ca. cases had undetectable serum and 5/10 (50%) patients had detectable serum Thyroglobulin with a p-value of 1.00. While 6/10 (60%) of follicular Ca. had undetectable serum Thyroglobulin and 4/10 (40%) of follicular Ca. patients had detectable serum Thyroglobulin with a p-value of 0.527.

P values are not statistically significant due to small sample size. However, this is an on going study and with larger sample size, a statistically better correlation in expected.

M. Zaman, R. Toor, S. Kamal, M. Maqbool,

S. Habib, K. Niaz.

NHS Grampian, University of Aberdeen, and Health Services Research Unit, University of Aberdeen, AB25 2ZN, United Kingdom.