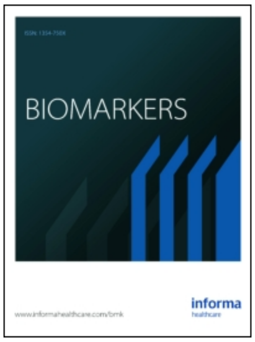


Biomarkers

ISSN: 1354-750X (Print) 1366-5804 (Online) Journal homepage: http://www.tandfonline.com/loi/ibmk20

Commentary on Modarresi, A. *et al* (2018) N-acetylcysteine decreases urinary level of neutrophil gelatinase associated lipocalin in deceased-donor renal transplant recipients: a randomized clinical trial.

David Paul Lovell

To cite this article: David Paul Lovell (2018): Commentary on Modarresi, A. *et al* (2018) N-acetylcysteine decreases urinary level of neutrophil gelatinase associated lipocalin in deceased-donor renal transplant recipients: a randomized clinical trial., Biomarkers, DOI: 10.1080/1354750X.2018.1499132

To link to this article: https://doi.org/10.1080/1354750X.2018.1499132

Accepted author version posted online: 11 Jul 2018.

Submit your article to this journal

Article views: 2

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=ibmk20

2nd July 2018

Commentary on Modarresi, A. *et al* (2018) N-acetylcysteine decreases urinary level of neutrophil gelatinase associated lipocalin in deceased-donor renal transplant recipients: a randomized clinical trial.

Accepting a paper for inclusion in a highly respected journal such as Biomarkers carries a responsibility to the scientific community. This is increasingly so given the wider societal concern about the problem of reproducibility of experimental findings in science, the perception that there is a bias against the publication of negative results and the developing unease over the use of statistical significance for the assessment of results. In this issue Biomarkers is publishing a paper by Modarresi et al (2018) reporting the results of a randomized clinical trial (RCT). This paper reports what appears to be a well-designed and well-conducted RCT. It is a simple study, clearly reported. Details of the trial design and the protocol including the statistical analysis plan are available on line at www.irct.ir (trial registration number: IRCT2014090214693N4). The paper, though, illustrates a number of issues associated with the refereeing of papers and the choice of papers to include in a journal. This commentary will use the paper to address a number of the issues that are relevant to the assessment of whether a paper is suitable for publication. These include: the statistical methods used, the availability of the 'raw' data for independent analyses, sample size and power calculations and the role of statistical significance in the assessment of results.

Acute kidney injury (AKI) disease is a common complication after kidney transplants, particularly from dead donors. Modarresi et al (2018) carried out an RCT to investigate whether, the possible renoprotective effect of N-acetylcysteine (NAC) administered to recipients could be identified by a reduction in the levels of a biomarker, urinary Neutrophil Gelatinase Associated Lipocalin (u-NGAL). u-NGAL is an early and sensitive marker of AKI after kidney transplantation. The paper reports a top line result in the abstract as "NAC significantly reduced u-NGAL levels compared to placebo (P value=0.02), while improvement in early graft function with NAC did not reach statistical significance." The main conclusion was that NAC administration can reduce AKI but that a larger sample size was needed to reach a statistical conclusion on the improvement in early graft function.

Background to the paper.

The study investigated 70 patients who received kidneys from dead donors in a double-blind, randomized, placebo-controlled trial. Patients were administered either 600 mg oral NAC or placebo twice daily from day 0 to day 5 and various biomarkers measures and early graft function assessed. The primary outcome was u-NGAL levels measured at baseline and on the first and fifth day after the kidney transplant. There were a number of secondary endpoints: including serum creatinine levels, estimated glomerular filtration rate (eGFR) and early graft function.

In the review process a referee commented that the paper relied very heavily on a single 'significant' finding (P=0.02) obtained using a sophisticated (and potentially appropriate) statistical method, the General Estimating Equations (GEE). The referee thought that the methods used were not described in sufficient detail and felt that the results supplied did not provide sufficient confidence that the small effect detected was a clinically important finding. The authors were also asked to provide the 'raw' data and more details of the statistical

Accepted Manuscript

analysis they carried out using the statistical package SPSS. The authors kindly made available the data, the SPSS instructions and the results of their statistical analyses. They also provided a comprehensive Statistical Analysis Plan (SAP). The authors were also asked to provide a fuller description of the randomization process and blinding precautions they used which they included in the next version of the paper. From the information received the study looks as though it was well conducted and there is no reason to believe there is anything wrong with the statistical analyses.

Statistical Analyses carried out

The statistical analysis carried was described in the SAP (lodged with the protocol). The authors used the Generalized Estimating Equation (GEE) approach to investigate the u-NGAL measures made at different times. The result of fitting a GEE using SPSS software showed that the 'between groups' component is identified as a significant term in the GEE model (P=0.02) as shown in their Table 2 and that the estimate of this effect was -197.8pg/ml.

Those more familiar with a ‘traditional’ approach of using t-tests for comparing two groups would have found that the ‘simple’ one-way anova/two sample t-tests on the u-NGAL1 and u-NGAL5 data show no significant difference between the two groups (diff -84.1 pg/ml, 95% confidence interval (CI) -274.4 to 106.3; P=0.38 for u-NGAL1; and diff -142.4 pg/ml, 95%CI -360.9 to 76.1; P= 0.20 for u-NGAL5.) However, the two groups differ significantly at the baseline, u-NGAL0; (diff 139.1 pg/ml, 95%CI 22.9 to 255.3; P=0.020).

An informal definition of a confidence interval is that it is a range of values, derived from sample statistics, which is likely to contain the value of an unknown population parameter. A longer more precise definition can be found in statistical texts. A negative value relates to lower values for the treated group compared with those for the control/placebo group.

The mean of the differences (or 'change') between baseline and u–NGAL1 were 223.2pg/ml with a 95% confidence interval of (50.0 to 396.4 and P=0.012) and for u-NGAL5 were 281.6pg/ml with a 95% confidence interval of (62.0 to 500.9 and P=0.013). Modarresi et al included the baseline u-NGAL values as a covariate in their GEE model where the group term had a P value of 0.017 with a mean difference of -197.9 pg/ml and a 95% confidence interval of (-360.2 to -35.5). (Note the small difference compared with Table 2 probably relates to rounding.) Not including the covariate in the GEE model would have resulted in a P-value of 0.19 (mean difference -113.3 (95%CI -281.2 to 54.7).

The Generalized Estimating Equation (GEE) approach

Many statisticians feel that modelling of data is a more suitable approach than the hypothesis testing approach. This can be difficult for the non-statistician who has grown up with an approach based upon relatively simple tests, hypothesis testing and P values. In passing, it should be noted that the t-test is special case of the analysis of variance, which in turn is a special case of the General Linear Model (GLM), which, in ‘Russian Doll’ fashion, is a special case of the Generalized Linear Model (GLZ) (Nelder & Wedderburn, 1972). Most of the ‘standard tests’, t-tests and ANOVA were developed when the calculations were done on manual calculating machines using algebraic equations (familiar to readers of old statistical text books) that made these calculations feasible. The modelling approaches use algorithms

Accepted Manuscript

based upon matrix algebra thus opening a much wider modelling approaches. A challenge for both scientist and statistician is how to integrate modern statistical methodology into the routine analysis of biological data.

The GEE is one of several modelling approaches developed to analyse repeated or longitudinal measures. Such measures tend to be correlated with each other and violate some of the assumption underlying and limiting the usefulness of ‘traditional’ anova methods. Failure to take these correlations into account can result in the estimates of standard errors of the model parameters being incorrect and consequent errors in the interpretation of the 'significance' of effects. Considerable effort has gone into developing statistical modelling approaches which can cope with more complex data and provide a generalized approach to the modelling of data. Non-statisticians tend to think of these methods as new or novel but many of them trace back to the 1970s and earlier. GEE, for instance, was first introduced in papers in the mid-1990s.

The GEE approach is an alternative method to or an extension of using GZMs analysing correlated and non-normally distributed data such as found with repeated measures or longitudinal studies. GEE is called a population averaging method (and is an alternative to random effect models) using ‘marginal models'. It estimates the 'population-average effect of covariates on the response of interest' and 'not on any random effects or previous responses'. GEE simplifies the mathematics needed for this estimation by using the group means and within group correlations in its calculations. GEE uses a 'working correlation matrix' to 'model the correlation between successive measurements'. GEE has several options which can give different results. In a paper such as this where emphasis is put on a result with a P <0.05 it is important to be able to see exactly what options have been chosen. The SPSS version of GEE allows several different assumptions to be made about the correlations. (In this set of data changing the assumptions about the correlations has no effect on the results.)

GEE as used here has similarity to an analysis of covariance (ANCOVA). Readers familiar with General Linear Modelling/Analysis of Variance (GLM/ANOVA) approaches would see that when an Analysis of Covariance (ANCOVA) is carried using the u-NGAL1 or u-NCAL5 (day 1 or day 5) data as the variate and the day 0 data (u-NGAL0) as the covariate the 'between group' component is significant at P=0.042 (u-NGAL1) and has P=0.063 (u-NGAL5) for days 1 and 5 respectively. In effect, both, the ANCOVA and the GEE analysis use baseline as the covariate to 'correct' for the baseline differences to give a test between the two groups. Therefore, the result only becomes ‘significant’ when either the GEE or ANCOVA approach is used with a covariate. As the authors state "The comparison is only significant when the values are corrected in the GEE model for the baseline values". Note that the specific use of the baseline in GEE analysis as a covariate was not explicitly stated in the SAP.

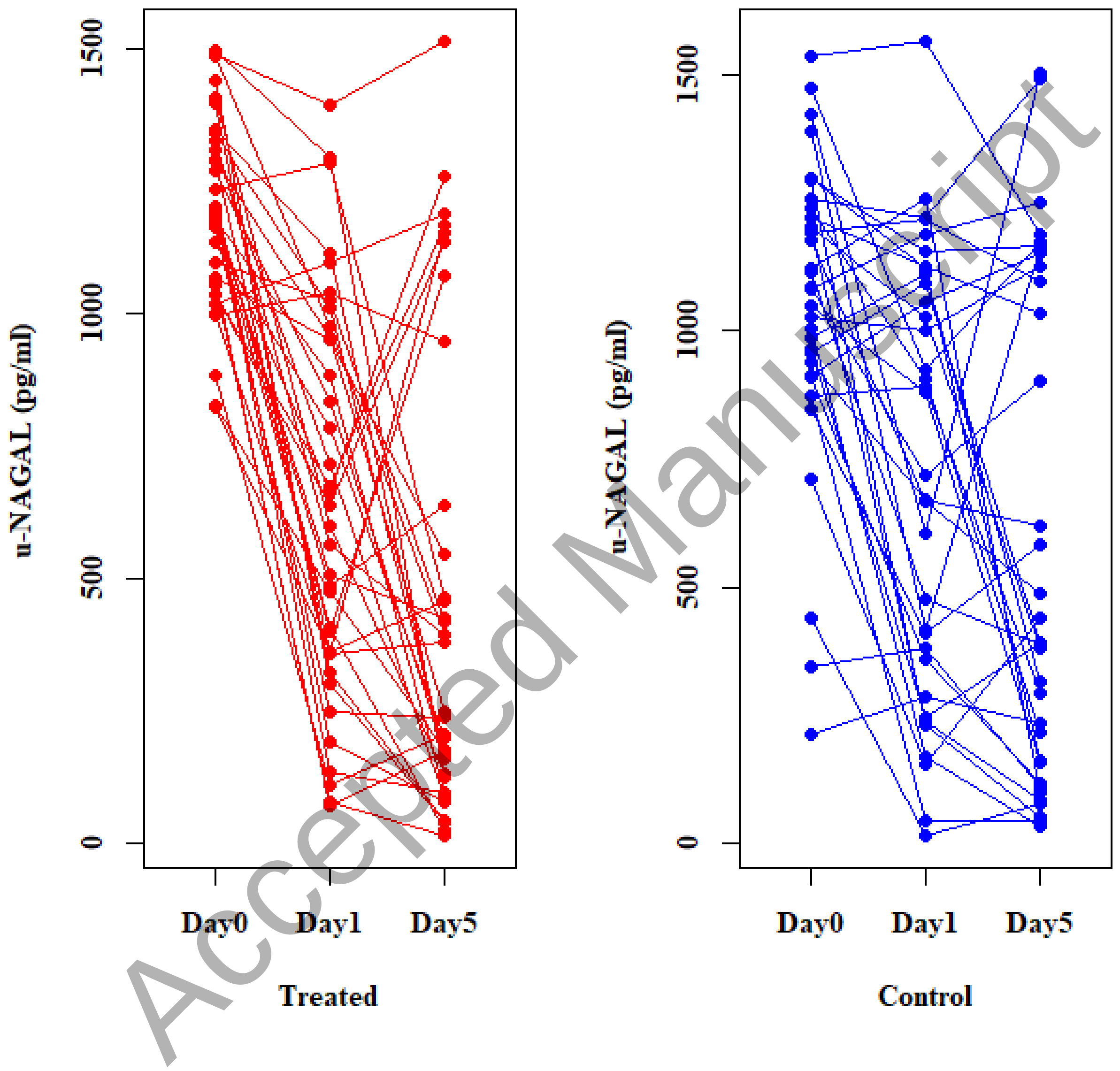
Looking at the raw data

Going back to the 'raw data'. Figure 1 is the crux of the paper. Data in red are from the treated group, blue from the control/placebo group. Using a 'standard' two sample t-test, there is a significant difference between the groups at day 0 (P=0.020) but a non-significant difference at P=0.38 at Day 1 and P=0.20 at Day 5. Looking more closely at the raw data there are 3 individuals in the control group with much lower baseline values than the others. Remember

Accepted Manuscript

that the subjects in the ‘blue data’ group did not receive the drug that was expected to reduce the level of u-NGAL Note also that there may be a 'basement effect' as the values cannot go below zero.

Figure 1



Accepted Manuscript

Figure 1: 'Spaghetti plot' of u-NGAL (pg/ml) values as 0, 1, and 5 days for treated group (red) and control/placebo group (blue)

If the individual in the control group with lowest baseline value is removed from the GEE with covariate analysis the P value becomes P=0.15; removing just the second lowest and P =0.145 and, just the third lowest, P =0.12. So, whether the P value associated with test of the group effect is less than the ‘magic’ P=0.05 threshold depends on whether the data from one or other of the 3 subjects with the lowest baseline scores in the control are included in the analysis or not. This pattern of results - no effect in a t-test and then a significant effect in the ANCOVA - may be real but it may also be an artefact related to the difference between the two groups at baseline and possibly the effects of some very low baseline measures in in the control group together with a 'basement effect' meaning that low value cannot fall as far as high values.

The bottom line result is that after treatment the value at day 5 is approximately 200 pg/ml lower in the treated group than the control group and that this is significant at P<0.05) with an exact P of approx. 0.02 after correcting for the baseline (u-NGAL0) using the GEE approach. However, when the ‘raw data’ are examined and the effect of a single individual are investigated this ‘picture’ may not appear convincing in terms of a biologically/clinically important result other than it has got over the P=0.05 hurdle.

The authors concluded that their “…study showed that NAC administration in deceased-donor KT recipients can reduce tubular kidney injury, evidenced by u-NGAL measurements. Improvement in early graft function needs a larger sample size to reach a statistical conclusion.” A closer look at the results suggest that care is need in interpreting the evidence from the u-NGAL measurements and that this needs to be taken into account in the design of further studies to investigate early graft function further.

This paper raises and illustrates a number of points that can be generalized across many investigations.

Sample size and Power Calculation

In their SAP the authors carried out a sample size calculation before conducting their study using an effect size based upon a standardized difference (difference/ standard deviation), of 0.75, a significance level of 5%, power of 80% in, presumably, a two-sided test. Note that a standardized difference of 0.80 is considered to be a large effect size by Cohen (1988), the developer of the approach. Standard sample size packages (Minitab and GPower) estimate 29 per group. The authors increased the number per group they derived of 28 by 30% to 36 to take into account attrition.

'Effect size' has been equated to a 'clinically relevant difference' (crd) or a 'minimal clinically important difference' (MCID) where it is ‘smallest change in an outcome that a patient would identify as important’ (Jaeschke et al.; 1989). Schunemann & Guyatt (2005, p.594) recommended widening the usage by dropping the ‘clinical’ so that it is the 'minimally important difference' (MID). The effect size approach was developed by Cohen (1988) but has been criticised by Lenth (2001, 2007) and others because it avoids the question of defining what is the size of the biologically important effect that the study aims to detect and falls back on ‘if it is significant I’ve got a positive effect and I can report it.’ There is a greater chance of a study being published if the results are considered positive than negative. There remains a tendency to dichotomize an effect as having occurred or not especially using significance tests. This can lead onto variability in a small effect being converted by this dichotomization into a lack of reproducibility.

Accepted Manuscript

The results of this study suggest that although a significant difference was found the size of the effect detected was below that used for the study design. Based upon estimates of the within group variability in the pre-treatment groups an estimate of the standard deviation (SD) of the u-NGAL is about 300pg/ml (actual value is 297) and the standardized difference detected is approximately 0.66SD units; below the 0.75SD units which could considered to be the MID. These SDs are based upon sample sizes of over 30 and although estimates of SD have wide variability when sample sizes are small these n's of more than 30 should provide some degree of precision in the estimate. (The SDs of the placebo group are larger on Days 1 and 5: 423 and 483 pg/ml with observed effect sizes of 0.47 and 0.41 respectively.) It is important to realise that the effect observed in a study is a single sample from a distribution of possible effects. This single sample could be at the upper end of the distribution of a small 'real' effect or the lower end of the distribution of a large 'real' effect. The one observed though is considered the 'best' estimate.

The argument that the authors need to carry out a bigger study is, therefore, less convincing. They designed a good study and showed the effect observed (if not an artefact) was lower than a 'minimally important difference'.

Statistical significance as a flawed paradigm

In the paper the authors state "P values <0.05 will be considered as the significance level." The abstract makes clear that the paper depends heavily on achieving a ‘statistical significance’ which is increasing viewed by statisticians as a flawed concept.

The criticism of the over-reliance on the use of significance testing and P-values in the interpretation of results is not new. This occurred before the recent surge in articles in both the scientific and popular press linking 'significance testing' to the ‘crisis of reproducibility’ in science. Twenty-five years ago, Cohen (1994) wrote “After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists”. Others like Fidler et al (2004) took a more measured/philosophical approach to what Cohen termed the Null Hypothesis Significance Testing (NHST) approach. For over 30 years statisticians have argued that an estimation approach based upon confidence intervals is better than the NHST/P value approach and should be reported in preference (Gardner & Altman, 1986).

Often more important, though, than whether the differences are significant is whether the assumptions behind the test have been met. In particular, whether factors such as were the samples all collected at the same time and effectively randomized to avoid batch effects and biases being introduced? These can easily arise in large observational studies introducing artefacts which can result in small differences becoming highly significant. Does, for instance, the statistical significance disappear if the outliers are excluded from the analyses? In this study considerable effort was put into the randomization and blinding process.

ANCOVA .v Change from Baseline

The finding that the two groups differ significantly at the baseline before a study starts, as here, can cause some confusion. In an RCT the assumption is that the groups were adequately randomized and this is one of the one in 20 events associated with the Type 1 error. It has been suggested that in these types of circumstances the randomization process should be repeated. Others, such as Bland (2004) argue that such events will occur as part of the true

Accepted Manuscript

randomization process and that such a study should continue without change. Of course pre-experiment differences between the groups can arise from biases in group selection. Although the introduction of bias cannot be rule out the randomization in this study seems to have been thorough and done correctly.

Senn (2006), amongst others, has pointed to the advantage of using Ancova with the baseline as the covariate as opposed to a simple analysis of the change scores (the difference between the before and after scores).This still applies even when the groups differ at baseline because any biases are also likely to affect the change data as well

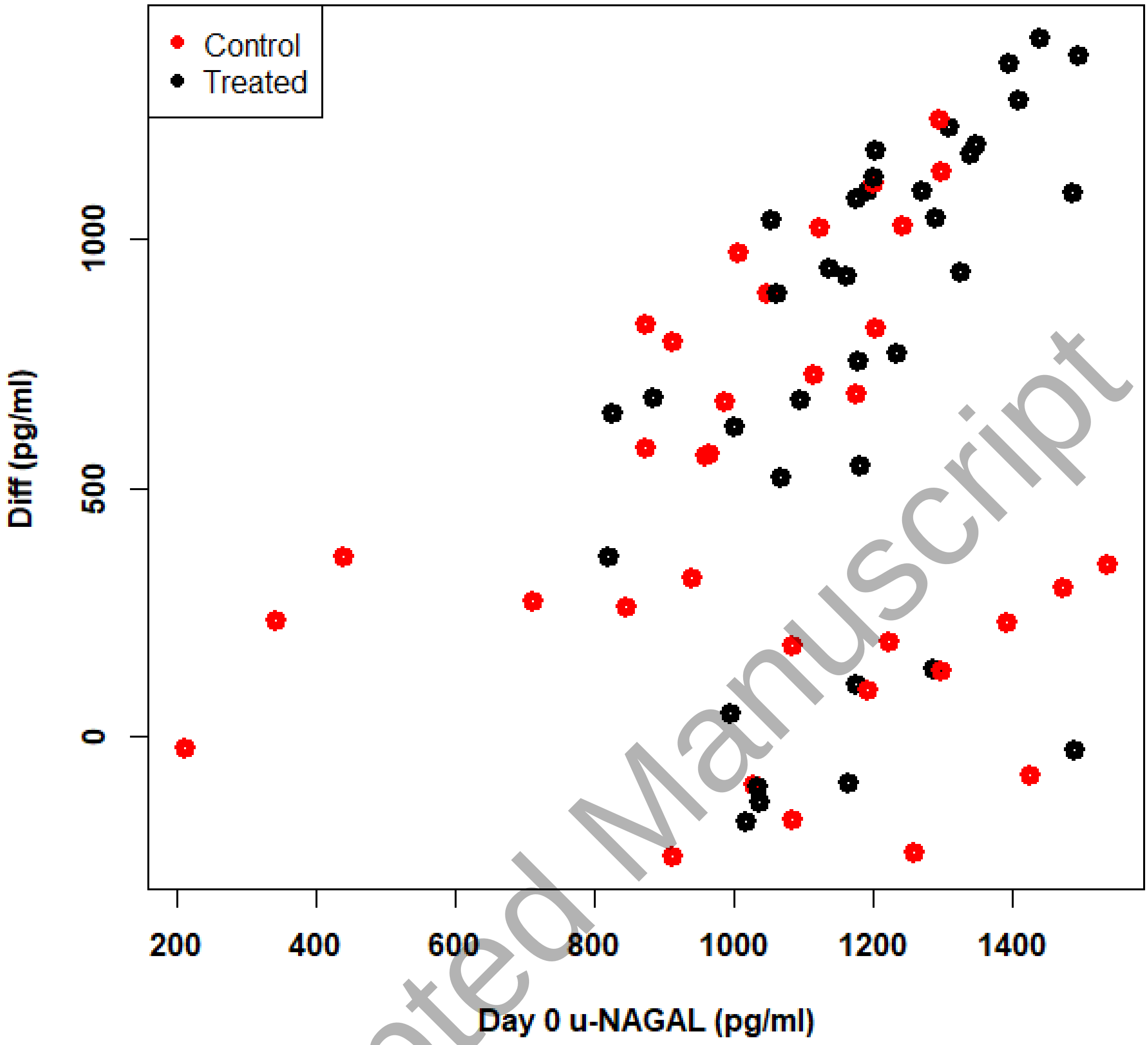
Heterogeneity of Response

Results of repeated measures studied are usually presented in tables or figures in summary form. However, presentation of individual responses in the form of 'spaghetti' or 'ladder' graphs can be informative. The figure of the raw data shows that the responses are heterogeneous. As well as showing that the four lowest baseline values are in the control group, it shows that 13 out of 34 controls have baseline levels below 1000 while only 4 out of 36 treated do. (*Post hoc* Fisher exact test P =0.012.) This imbalance needs to be at least considered in the context of a 'basement effect' where many control values have less 'room' to fall.

A plot of the change score against the baseline (Figure 2) indicates the presence of two groups (as well as the three control individuals with baseline values of <500). One group of 48 of individuals ((29 treated: 19 placebo) show a clear linear relationship while a second smaller group of 20 individuals (8 treated: 12 placebo) had very small reductions or even increases.

Figure 2

Accepted Manuscript



Accepted Manuscript

Figure 2. Plot of difference (Diff) between Baseline (u-NGAL0) and Day 1 (u-NGAL1) versus Baseline (u-NGAL0). Group 1(black), 'Treated'; Group 2 (red), 'Control/Placebo'

The presentation of repeated measures in 'spaghetti-type' graphs (Figure 1) so that heterogeneity of response can be identified is recommended. As mentioned previously, such figures can also provide a good guide to considerations of biological/clinical importance compared with statistical significance.

Conclusions

Biomarkers would normally need stronger evidence than a single P-value from a study before accepting a paper. Particularly when the effect which, even if not an artefact, was appreciably smaller than what the authors would have considered an important effect based upon their study design.

However, *Biomarkers* felt this was a well-conducted study that deserved to be published. It is an interesting paper in that it illustrates some of the points that need to be discussed when deciding whether to publish results based mainly on a single significant result (even though on the designated primary endpoint) is not an easy one. The commentary is, therefore, not a criticism of the authors' work but an attempt to educate people on some of the complexities of the process.

This paper illustrates a number of issues: the need for good study design, the finding that a result is statistically significant in some tests but not others, the dependence of the results on statistical analysis of the data taking into account that the groups differed before the study started, the use of more sophisticated modelling compared with traditional statistical tests, the identification of significant effects below those the study was designed to detect, the use of a cut-off P-value to define statistical significance and a positive result, the potential influence of single data points and the heterogeneity of responses.

It is important in the context of a scientific study/clinical trial that the data and statistical methods used are available to be able to check independently the results obtained. Many journals, such as *Biomarkers*, now require such information to be available as part of their requirements for transparency and to asses reproducibility. This is of increasing importance in view of the plethora of new journals which appear to have sprung up in the last few years, which are competing for papers and which, it is feared, may have lower standards in their acceptance of papers.

The use of a P value cut-off based upon the NHST paradigm for determining a positive result is unsatisfactory. The use of estimates of the size of an effect together with a confidence interval has much to commend it. Such an approach may not provide an easy binary 'call' but is a more informative description of the results. It allows a much more nuanced interpretation of the biological importance of the results and consequently addresses part of the 'reproducibility crisis' by removing the false dichotomy of positive and negative experiments. Reporting the size of effects with their confidence intervals derived from well-conducted studies also circumvents the potential for journals to introduce publication bias by equating apparently negative studies as 'failure' and the consequent search for $P<0.05$ somewhere in the results (P-hacking) so that the study is considered a 'success' and, therefore, worthy of publication.

Accepted Manuscript

References

Bland M. (2004) How to Upset the Statistical Referee. Available at: http://www-users.york.ac.uk/~mb55/talks/upset.htm

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences. Routledge.

Cohen, J. (1994) The Earth Is Round (p < .05) American Psychologist 49 997-1003

Fidler, F., Geoff, C., Mark, B. & Neil, T. (2004) Statistical reform in medicine, psychology and ecology. J. Socio. Econ. 33 615–630

Gardner, M.J. & Altman, D.G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br. Med. J. (Clin. Res. Ed). 292 746–750.

Jaeschke, R., Singer J., & Guyatt, G.H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 10 407–415.

Lenth, R.V. (2001). Some practical guidelines for effective sample-size determination. The American Statistician 55 187–193.

Lenth, R.V. (2007) Post Hoc Power: Tables and Commentary. https://stat.uiowa.edu/sites/stat.uiowa.edu/files/techrep/tr378.pdf

Modarresi, A. *et al* (2018) N-acetylcysteine decreases urinary level of neutrophil gelatinase associated lipocalin in deceased-donor renal transplant recipients: a randomized clinical trial. Biomarkers. (In Press)

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. Journal of the Royal Statistical Society, Series A, 135 370-384)

Senn, S. (2006) Change from baseline and analysis of covariance revisited. Stat Med. 25 4334-4344.

Accepted Manuscript