

Proper statistical analysis of transepidermal water loss (TEWL) measurements in bioengineering studies

OLIVER KUSS AND THOMAS L. DIEPGEN

Department of Dermatology, Dermato-Epidemiology Unit, University of Erlangen/Nürnberg,
Erlangen, Germany

In irritancy studies, measurement of transepidermal water loss (TEWL) is a widely used technique to assess barrier function. Using inappropriate statistical methods, however, leads to loss of information and misinterpretation of results. In this paper, we discuss some problems and pitfalls when using a suitable statistical technique for most designs in bioengineering studies, analysis of variance (ANOVA): multiple comparisons, choice of sample size and violation of statistical assumptions. For clarification of these points, a practical example will be given. Using the proposed adequate statistical methods correctly will, although accompanied by increased complexity, increase the efficiency of bioengineering studies.

Key words: analysis of variance; bioengineering techniques; transepidermal water loss; sample size; power; multiple comparison; repeated measurements; metalworking fluids; irritant contact dermatitis. © Munksgaard, 1998.

Accepted for publication 9 April 1998

Measurement of transepidermal water loss (TEWL) is a well-known and widely used bioengineering technique to assess the barrier function of the skin. In recent years, considerable efforts have been made to establish guidelines for the standardization of TEWL measurements (1), and it has been shown that TEWL is a stable and reliable individual characteristic by proving that inter-individual variability of TEWL is much greater than intra-individual (2). In the literature on this topic, however, little effort has been made to use adequate statistical methods to attain all available information from the observed data. Analysis of variance (ANOVA) models, in their simplest forms just extensions of the *t*-test for the comparison of more than 2 groups, seem to be suited to deal with the common study designs for TEWL measurements. It may seem strange that a statistical method that compares means in several treatment groups is called analysis of variance. This is due to the fact that the variability in the individual data is partitioned into components of variation within and between treatment groups. According to the null hypothesis, that is that all population means are the same, the between and within group vari-

ances will be the same, but in the case of differing means, the between variance will be larger. The ANOVA *F*-statistic is thus defined as the ratio of the between and within variances, and so the larger the value of *F* the more the group means differ. For significance testing, the observed value of *F* is compared with values from the *F* distribution. There are, however, some statistical problems that have to be considered in the various stages of an experiment (planning, measuring, analysing). In this paper, we will discuss some of them and make proposals for their solution.

Multiple Comparison

The usual ANOVA *F*-test for treatment effects tests the equality of means in the treatment groups. When the null hypothesis of equal means is rejected, there is at least 1 treatment differing significantly from the others, although we do not know which.

If we want to determine which treatments are different, we have to do some pairwise comparisons of the means in the treatment groups. The simplest approach is to do a series of *t*-tests, 1 on

every pair of means. Unfortunately, this procedure causes serious problems with inflating the type I error, the probability of a false rejection of the null hypothesis. In our case, this means claiming erroneously a difference between single treatments. It is difficult to calculate the exact probability of this inflated type I error, but it can be approximated by $1 - (1 - \alpha)^{k/(k-1)/2}$, where k is the number of groups compared and α the significance level of the single t -test. For example, if there are 6 treatment groups we have to do 15 pairwise comparisons and would achieve (if we use an $\alpha=0.05$ as level of significance for the single t -test) a probability of 0.54 that at least 1 test erroneously claims a difference between 2 different treatments. In other words, we have a probability of 54% of achieving a statistically significant difference by pure chance.

A variety of methods have been proposed to handle this problem. The most simple, but also the least powerful, among these is the Bonferroni correction, where the experiment-wise significance level is divided by the number of tests carried out in the experiment, and pairwise t -tests are calculated with the corrected significance levels. More powerful methods exist with increased statistical complexity. There are procedures designed exclusively for the comparison of several treatments with a single control. Others no longer use an adjusted, but constant, critical value to assess statistical significance, but involve a hierarchy of testing stages where the significance criterion is adjusted throughout the stages (3).

Choice of Sample Size

Choice of sample size is an important, not just statistical, but also ethical problem in designing an experiment or trial. On the one hand, sample size should be big enough to enable detection of a difference in treatments, and on the other hand small enough to ensure that subjects are not unnecessarily treated with harmful irritants. In any case, the sample size has to be calculated before the data are collected.

In general, sample sizes are determined via controlling the power of the utilised test at a specified alternative. The power of a test is the probability that the null hypothesis will be rejected if it is in fact false or, in other words, the probability that the experiment would detect as statistically significant a real existing difference of a given magnitude. As for any statistical test, the power of the ANOVA F -test depends on the distribution of the test statistic if the null hypothesis is false, that is we have different means of TEWL in the treatment groups. Actually, to calculate the sample size required, we have to specify several quantities in ad-

vance: the type I error (which is α), the type II error (that is 1-power) and the population variance σ^2 (which can be estimated from a small pilot study or taken from the literature). The most important issue, however, is to fix the clinically relevant difference in the means we want to detect, because every difference observed between treatment groups can be made statistically significant by including more and more subjects in the study; however, not every difference thus found is of practical clinical interest.

Violated Statistical Assumptions

Important assumptions in the ANOVA design are normality of data within every group, equality of variances across groups, and independence of the sampled data within every group. Violations of the first 2 assumptions can be shown to be less serious because the ANOVA F -test is not greatly affected by non-normality and unequal variances. Furthermore, in most cases, data can be transformed to achieve the desired properties.

A more serious problem arises if more than 1 treatment is given to the subjects, as is frequently the case in bioengineering experiments. Of course, the observations are no longer independent; instead, we expect correlations in the data derived from the same subject. Note that this situation is just the generalization of the distinction between the paired and the unpaired t -test to the case of more than 2 treatment groups. By not controlling for these dependencies, variability due to the differences between subjects and experimental error are confused and, therefore, the power of the F -test to detect treatment differences decreases.

Unfortunately, the relaxation of the independence assumption induces a different assumption that should be met by the observed data, namely the homogeneity of treatment-difference variances. In case of heterogeneity, the degrees of freedom for the F -test have to be multiplied by the Huynh-Feldt correction factor ϵ , which is smaller than 1 and has to be estimated from the data. This means that this correction lowers the power of the test and can almost be regarded as a punishment for not fulfilling the independence assumption (4).

An Example

TEWL measurements were used to evaluate the irritancy of 7 different metalworking fluids (CF1-CF7). 10 healthy subjects were treated with the irritants for 24 h on 7 test sites on the forearm, and 1 additional test site was treated with water and served as a control. This in-vivo irritation test model was first described by Hüner et al. (5) and

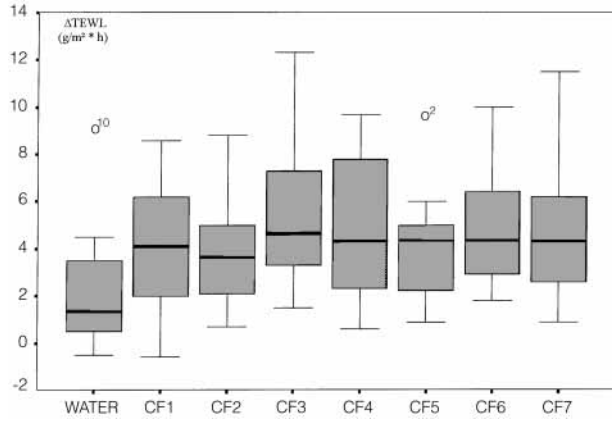


Fig. 1. Summary statistics for the Δ TEWL measurements across treatment groups.

was recently confirmed as a good in-vivo model to investigate the irritant capacity of metalworking fluids (6). TEWL was measured 3 h after the end of irritation, and the baseline TEWL before irritation was subtracted from the achieved TEWL measurements to make the actual changes in TEWL visible. Fig. 1 shows the distribution of the Δ TEWL measurements across the different treatment groups via boxplots. Boxplots show the median (the line inside the box), the 25th percentile (the lower boundary of the box), the 75th percentile (the upper boundary of the box) and outlying values that are far removed from the others. The central vertical lines extend from the box as far as the data extend to a distance of at most 1.5 box-lengths. Data points outside this range are regarded as outliers and are marked with an “o”.

The first hypothesis to be tested was the classical F -test hypothesis of equal treatment means. It was tested within 2 different models. First, the traditional 1-way fixed effects ANOVA F -test was used to assess treatment effects. This test ignores the fact that measurements come from the same subjects and yields a p -value of 0.341, that is, it detects no significant difference between the treatments.

Secondly, we carried out the test using this time the correct statistical method in this case, the repeated measures model described previously. Now that we account for the fact that every subject was treated with each of the treatments, we get a p -value of 0.080 (with the Huynh-Feldt correction: 0.099).

Suppose we regard the observed data as arising from a pilot study and want to use them to design a future study (where we consider the observed differences as clinically relevant). In the 1st, incorrect, model we would need 24 subjects in every treatment group to reach a power of 0.9, a commonly

desired value of power in trials and experiments, to detect existing treatment differences. As this test assumes that every subject is treated with only one single treatment, we would actually need 192 subjects.

In the 2nd model (the repeated measures model), we would have required 15 subjects (Huynh-Feldt: 16) to attain a power of 0.9. All in all, we achieve the same probability of detecting differences in treatments with less than 1/10 of subjects in the repeated measurements model, at least in the given situation.

To give an impression of how the required sample size changes in the repeated measurements model with the various parameters that control the power of the F -test, we refer to Table 1. To achieve the given figures, we held fixed the population variance σ^2 (8), the intra-individual coefficient of correlation r (0.4), which were estimated from our study and the level of significance ($\alpha = 0.05$). We varied the number of treatments (4, 6, 8), the desired power to detect differences in treatments (0.8, 0.9) and the clinically relevant differences ($\text{g/m}^2 \text{ h}$) of Δ TEWL between treatments that should be detected by the experiment (1–5). For example, if we plan a study that compares 6 treatments, consider a treatment difference of 4 $\text{g/m}^2 \text{ h}$ Δ TEWL clinically relevant and would like to have a 90% probability of detecting this clinically relevant difference in treatments, we have to recruit 12 probands (if we are willing to accept the other assumptions regarding population variance and intra-individual correlation). An increase in the intra-individual correlation would lead to an increase in power and, therefore, to a decrease in the required sample size. The opposite is true for the population variance: an increase leads to a loss of power and to an enhanced sample size.

Another research question of medical interest is how the different metalworking fluids compare with each other and with the control. If we do pairwise t -tests for every combination of the treatments without controlling for the overall type I error, we find CF3, CF4, CF6 and CF7 signifi-

Table 1. Required sample size for the F -test in a repeated measurement model ($\alpha = 0.05$, $\sigma^2 = 8$, $r = 0.4$)

Clinically significant differences in means of treatment groups ($\text{g/m}^2 \text{ h}$) Δ TEWL	Number of treatments					
	$k=4$		$k=6$		$k=8$	
	power		power		power	
	80%	90%	80%	90%	80%	90%
1	106	138	125	160	139	177
2	28	36	32	41	36	45
3	13	17	15	19	17	21
4	8	10	9	12	10	13
5	6	7	7	8	7	9

cantly different from the control group (at the 5%-level). This leads, however, to a rise in the overall type I error of $1-0.95^{28}=0.76$, that is, we have a probability of roughly 3/4 that at least 1 of the pairwise *t*-tests rejected the null hypothesis erroneously and are not willing to rely on these findings. Actually, the application of Tukey's method, a multiple comparison procedure which controls the overall type I error at the specified 5%-level (by enlarging the critical value to compare the test statistic to and therefore rendering a significant finding more difficult), obtains different results: it declares only CF3 to be different from the control group.

Conclusions

We hope that our remarks have demonstrated some of the problems that are related to an adequate analysis of TEWL measurements in the ANOVA design, although some of the problems carry over to other statistical methods as well. A lot of work still remains to be done to (1) attain *all* available information from the observed data and (2) even more importantly, to attain the *right* information and not be fooled by results that are invalid due to inadequate statistical methods.

It has to be noted that the given overview of methods and problems in ANOVA is far from being complete, but a satisfactory treatment of them is well beyond the scope of this paper and one must refer to the statistical textbooks. There are a huge number of experimental designs that can also be modelled within this statistical framework. To name just a few, there can also be more complicated higher-order designs if one wants to account for more factors than just treatments and subjects, there can be continuous factors which require analysis of covariance (ANCOVA), or we have several dependent variables, for example re-

peated measurements over time, which can be analysed within the context of multivariate analysis of variance (MANOVA). There are also a variety of distribution-free methods within the ANOVA context when the usual assumptions are not fulfilled. Last, but not least, a number of graphical methods to visualize results from the analysis have been proposed.

In practice, we further strongly recommend consultation with an experienced biometrician as early as the planning stages of the experiment.

References

1. Tupker R A, Willis C, Berardesca E, Lee C H, Fartasch M, Agner T, Serup J. Guidelines on sodium lauryl sulfate (SLS) exposure tests. A report from the Standardization Group of the European Society of Contact Dermatitis. *Contact Dermatitis* 1997; 37: 53-69.
2. Smit H A, Pinnagoda J, Tupker R A, Burema J, Coenraads P J, Nater J P. Variability in transepidermal water loss of the skin: evaluation of a method to assess susceptibility to contact dermatitis in epidemiological studies. *Int Arch Occup Environ Health* 1994; 62: 509-512.
3. Lindman H R. *Analysis of variance in experimental design*. New York: Springer, 1991.
4. Keselman H J, Keselman J C. Analysis of repeated measurements. In: Edwards L K (ed): *Applied analysis of variance in behavioral science*. New York: Marcel Dekker, Inc., 1993: 105-146.
5. Hüner A, Fartasch M, Hornstein O P, Diepgen T L. The irritant effect of different metalworking fluids. *Contact Dermatitis* 1994; 4: 220-225.
6. Wigger-Alberti W, Hinnen U, Elsner P. Predictive testing of metalworking fluids: a comparison of 2 cumulative human irritation models and correlation with epidemiologic data. *Contact Dermatitis* 1997; 36: 14-20.

Address:

Thomas L. Diepgen
Department of Dermatology
University of Erlangen
Hartmannstr. 14
91052 Erlangen
Germany