How to analyse TEWL-measurements properly - A statistical approach

O. Kuss, T.L. Diepgen
Dermato-Epidemiology Unit, Dept. of Dermatology, University of Erlangen, Germany

Introduction
The measurement of transepidermal water loss (TEWL) is a well known and widely used bioengineering technique to assess the barrier function of the skin. In the last years considerable efforts were made to establish guidelines for the standardisation of TEWL-measurements (Tupker et al. (1997)) and it could be shown that TEWL is a stable and reliable personal characteristic by proving that inter-individual variability of TEWL is much greater than intra-individual (Smit et al. (1990)) .
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. The family of analysis of variance (ANOVA) models seem to be suited to deal with the common study designs for TEWL-measurements but it is connected with some statistical pitfalls.

Common problems in ANOVA designs for analyzing TEWL-data

1. Model assumptions violated
2. Inflation of the α-error by pairwise comparisons
3. Inadequate sample size

Serial correlation over time
Important assumptions in the ANOVA design are the normality of the data within every group, equality of variances across the groups and the independence of the sampled data within every group. Violations against the first two assumptions can be shown to be less serious because the ANOVA F-test is not greatly affected by nonnormality and unequal variances. Furthermore, the data can be transformated to achieve the desired properties.
A more serious problem arises if the data within the groups are correlated which will be naturally the case if the data consist of serial measurements over time. In that case we would expect characteristic nonrandom patterns of observations. The ANOVA F-test will be much more affected by violations against this independence assumption and, even worse, the direction of the effect depends on the nature of dependence over time. Therefore we are forced to modify the F-tests to account for this serial correlation over time.
Arnold (1981) proposed that if we can anticipate a certain form of the serial dependence in the error covariance matrices (e.g. autocorrelation or compound symmetry) that can be described with a small numbers of parameters, we can estimate this unknown parameters from the data, and calculate slightly modified F-statistics. Arnold expects this approximate procedures to work satisfactory in most cases.
Lefante (1990) showed that if we assume a linear change of TEWL over time within every treatment and estimate this change by simple linear regression, the usual F-test remains valid, at least in a balanced design.
A final possibility is to abandon the ANOVA design and move to the field of longitudinal data analysis, which would require, at least in our opinion, additional efforts in understanding the models and in carrying out the calculations.

Multiple comparison
The usual ANOVA F-test for the treatment effects tests the equality of the means in the treatment groups, i.e. if this test rejects our hypothesis we know that there is at least one treatment that significantly differs from the others but we don´t know which treatments differ from which others.
If we want to learn about this, we have to do some pairwise comparison of the means in the treatment groups. The simplest approach to this is a series of t-tests, one on every pair of means. Unfortunately, this procedure causes the serious problem of inflating the α-error, the probability of a false rejection of the null hypothesis. In our case this means to claim erroneously a difference between single treatments. It is difficult to calculate the exact probability of this inflated α-error, but it can be approximated by $1-(1-\alpha)^{k(k-1)/2}$, where k is the number of means to compare and α the significance level of the single t-test.
A variety of methods have been proposed to handle this problem. These reach from simply adjusting the significance level of the single t-test by the Bonferroni inequality over to the more powerful methods of Tukey and Scheffe and modifications of these.

Sample size choice
Sample size choice is an important, not just a statistical, but also an ethical problem in designing an experiment or a trial. On one hand, the sample size should be big enough to have a chance to detect a difference in treatments, on the other hand small enough to ensure that subjects are not unnecessarily treated with harmful irritants.
In general, sample sizes are determined via controlling the power of the utilized test at a specified alternative where the power of a test is the probability that the null hypothesis will be rejected if it is in fact false. The power of the ANOVA F-test, as it is the same for any other statistical test, depends on the distribution of the test statistic when the null hypothesis is false, that is we have different means of TEWL in the treatment groups. It turns out that this distribution is noncentral F and this distribution importantly depends on the actual means in the treatment groups, the population variance σ^2, the significance level α and the number of observations in the sample. To become concrete we have, under the assumption of a balanced design which is usually the case in the planning phase of a study, the power of the F-test as

$$\pi(\varphi; k-1, n-k) = p\,(F(k-1,n-k;\varphi) > F_{1-\alpha}(k-1,n-k)),$$

where $F(k-1,n-k;\varphi)$ is the noncentral F-distribution with noncentrality parameter φ and k-1 and n-k degrees of freedom, $F_{1-\alpha}(k-1,n-k)$ is the (1-α)-quantile from the regular F-distribution with the respective number of degrees of freedom, k is the number of treatment groups and n is the sample size to be determined. If we specify further the α-error, the β-error, the clinical significant difference in the means we want to detect and the population variance σ^2 (which we have to guess or estimate from a small pilot study) we can calculate the required sample size.
Further insight into sample size choice can be gained through the book of Desu and Raghavarao (1990).

An example

Bioengineering situation
TEWL-measurements were used to evaluate the efficacy of skin barrier creams in the prevention of irritant contact dermatitis. For that purpose a cumulative irritation model was applied. One test site received the irritant alone, three other test sites were pretreated with different barrier creams before the irritant, one further test site remained untreated and served as a control. The procedure was executed twice a day for 14 days with a break at the weekend and TEWL was measured two hours after the second irritation.

Statistical model
The suitable statistical model for the described study design where means of TEWL-measurements are compared to detect differences between the protective effect of barrier creams (if there is any) an analysis of variance (ANOVA) model. As the TEWL-measurements are classified in two ways, one for the treatment received and the other for the subjects that receive the treatments, we have to deal with two-way ANOVA. A further extension has to be made regarding the levels of treatment and the different subjects. Because the treatment conditions are fixed in advance we consider the treatment effect as a fixed effect. The subjects, however, are randomly selected from some larger population on which conclusions want to be drawn and therefore their effect on TEWL-measurements is modelled as a random effect. Models with mixtures of fixed and random effects are called ´mixed effects models´.
Under the condition of equal numbers of observations in every group we can write down the balanced two-way mixed effects ANOVA model with interaction as

$$Y_{ijk} = \mu + \alpha_i + b_j + d_{ij} + e_{ijk},$$

where Y_{ijk} stands for the k-th measurement in the j-th subject for the i-th treatment, μ is the constant unknown grand mean of TEWL in the population, the α_i represent the fixed treatment effects and b_j and d_{ij} are unobserved independent normally distributed random variables for effects of subjects and interaction between treatment and subjects. The e_{ijk} are no longer independent in terms of k and account for the serial correlation. A slightly modified F-test now can be used to test the central hypothesis of equality of treatment, that is $\alpha_i = 0$ for every i. For computational formulas and further information we refer to Arnold (1981), in practice, the calculations should be done with some statistical software package, for example SAS PROC MIXED.

Conclusions

We hope our remarks have demonstrated some of the problems that are related with a sufficient analysis of TEWL-measurements especially 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 important to attain the **right** information and not being fooled by results that are invalid due to inadequate statistical methods.

Literature

- Arnold, S.F. (1981): The theory of linear models and multivariate analysis. Wiley, New York.
- Desu, M.M.; Raghavarao, D. (1990): Sample size methodology. Academic Press, Inc., Boston.
- Lindman, H.R. (1991): Analysis of variance in experimental design. Springer, New York.
- Lefante, J.J. (1990): The power to detect differences in average rates of change in longitudinal studies. In: Stat. Med., 9, 437-446.
- Smit, H.A. et al. (1990): Variablity in transepidermal water loss of the skin: evaluation of a method to assess susceptibility to contact dermatitis in epidemiological studies. In: Int. Arch. Occup. Environ. Health, 62, 509-512.
- Tupker, R.A. et al. (1997): Guidelines on sodium lauryl sulfate (SLS) exposure tests. A report from the Standardization Group of the European Society of Contact Dermatitis. In: Contact Dermatitis, 37, 53-69.