Eur Arch Psychiatr Neurol Sci (1990) 239:391–394

European Archives of Psychiatry and Neurological Sciences

Letters to the Editors

Comments on the Paper by Möller et al. (1989)*: Problems in Single-case Evaluation

Joachim Krauth

Psychologisches Institut IV, Universität Düsseldorf, Universitätsstrasse 1, D-4000 Düsseldorf, Federal Republic of Germany

Received December 22, 1989

Summary. Möller et al. (1989) investigated the effects of sleep deprivation on depressive patients using a single-case analysis based on Kleiter's 1986 approach. Here it is discussed why it is so difficult to perform single-case analyses and why Kleiter's approach in particular may lead to wrong interpretations of the data. It is indicated that any kind of single-case analysis is based on assumptions which possibly do not hold and that it is therefore always precarious to base substantial conclusions on this type of analysis.

Key words: Single-case analysis – Time-series

Introduction

In the last few years, many authors in clinical research have brought forward well-founded arguments in support of single-case studies and have pleaded for the use of studies of this kind instead of the typical group designs. Likewise a number of methodological papers offering procedures for the evaluation of single-case studies have appeared recently. It is therefore not surprising that Möller et al. (1989) followed these arguments and proposed a single-case evaluation of sleep deprivation effects which is based on a time-series approach proposed by Kleiter (1986). The main object of this comment is to argue in support of my opinion that the conclusions the authors drew from their evaluation do not seem well-founded. The possible existence of better alternative approaches is discussed. Not in the least is it my intention to criticize Möller et al. (1989) for publishing their evaluation. The problem of analysing single-case studies is a very difficult issue, and it is not to be wondered at that clinical researchers make use of such an apparently sound approach, especially if this is accompanied by a computer program. It is rather my intention to put forward several reservations which might serve as a warning to other researchers who may be inclined to design and evaluate single-case studies in a similar way.

Fundamental Difficulties in Single-case Analysis

Data obtained in a single-case study must be assumed to be dependent. Even in very simple situations, this will have the effect that the more points of time considered, the larger the number of unknown parameters becomes. Subsequently, it becomes increasingly less possible to estimate these parameters. Consider, for example, a stationary time-series with multivariate normal distributed variables. In this case, the expectations μ and the variances σ^2 are the same for all points of time. If we can additionally assume that the variables are independent, we only have to estimate two parameters (μ and σ^2). With increasing sample sizes we obtain more and more information about these parameters and can either estimate them or perform significance tests with increasingly better success. However, if we allow the variables to be dependent in an arbitrary way, this results in an ever-increasing number of parameters which describe the dependence between the variables: For two points of time we have one correlation coefficient, for three points of time we have three correlation coefficients, for four points of time we have six correlation coefficients, . . . , and finally for n points of time we have $n(n-1)/2$ correlation coefficients. This means that, with an increasing number of observations, the number of unknown parameters increases even more rapidly, and the amount of necessary information about these parameters will consequently decrease very quickly. In other words, it becomes impossible to derive efficient estimates of the correlation coefficients, and it is no longer possible to perform significance tests. The situation is even worse in cases in which we cannot assume multivariate normal distributions. The dependence relations for three points of time can then no longer be described by three correla-

* Eur Arch Psychiatr Neurol Sci (1989) 239:133–139

tion coefficients alone, since we cannot exclude the possibility of a dependence between three variables which cannot be described by pairwise interdependences. For more than three points of time the situation is worse still. In the case of stationary time-series lacking the so-called property of ergodicity, it may even be impossible to derive efficient estimates of μ and σ^2 (Krauth 1980).

The only way to overcome these difficulties is to assume that the dependence structure of the data can be described by merely a few parameters. The best known way to do this is to assume so-called ARIMA (autoregressive integrated moving average) processes as done by Box and Jenkins (1976). These models assume that time-series are basically generated by linear combinations of independent random variables, so-called random shocks. In a first step, an attempt is made to identify the theoretical time-series model and to estimate the unknown parameters from a very long base-line. In a second step, these parameter estimates are used to estimate the residuals, i.e. the random shocks, of the time-series in a treatment phase. As the residuals are assumed to be independent, it seems possible to perform the classical statistical tests using the estimated residuals.

Unfortunately, we can never be sure to have identified the correct model, because any identification is based on confidence intervals and significance tests, the interpretation of which may be quite misleading. Furthermore, the estimates of the parameters will generally deviate from the true parameters. From this it follows that the estimates of the residuals can never be expected to really correspond to independent and identically distributed random variables. This is also true in those cases in which the corresponding goodness-of-fit tests do not yield significant results, because non-significant results do not allow the conclusion that the corresponding null hypothesis holds. In other words, one must always assume that a certain interdependence between the estimated residuals remains.

It might now be assumed that statistical tests are not much affected by small interdependences of the data. However, analytical results and simulation studies show that this is not the case. Even for small positive autocorrelations, parametric and non-parametric tests will yield "significant results" with a probability larger than the chosen α, though the null hypothesis is in fact valid. This can easily be seen in the usual *t*-tests for dependent or independent samples. For these tests, the estimate of the standard deviation in the denominator is generally too small for positive autocorrelations because the positive covariance terms are neglected. This leads to excessively large values of the *t*-statistics.

Only a few of the studies which investigate the influence of autocorrelation on parametric and non-parametric tests are mentioned here: Albers (1978a, b); Andersen et al. (1981); Gardner et al. (1982); Gastwirth et al. (1967); Gastwirth and Rubin (1971); Gleser and Moore (1985); Hibbs (1974); Moore (1982); Nicolich and Weinstein (1981); Tavare and Altham (1983); and Walsh (1947). Nearly all of these studies show that even very small autocorrelations can affect the validity of significance tests considerably.

Kleiter's Approach and its Criticism

Kleiter (1986) described a procedure for the evaluation of single-case studies. According to the author, this method avoids the weaknesses of the ARIMA approach. A short description of Kleiter's approach, which was used by Möller et al. (1989), is given in the following:

1. Preserving the chronological order of the measurements, a clustering based on sums of squares is performed. This yields segments of the time series which are homogeneous in a certain sense, i.e. they show a seemingly stationary behaviour.

This procedure is more of a dissection (Kendall and Stuart 1966) than a clustering because the chronological order is preserved. In any case, it must be assumed that any significance test which is used to compare different segments derived in this way will detect effects which do not really exist, with a probability which is much larger than the chosen α. The reason for this is obvious. The segments are formed in such a way that the values of the measurements within a segment are all approximately of the same magnitude and those of different segments are all of different magnitude. Since random fluctuations will always be present in real data, the proposed procedure will always yield segments of this kind, even in the presence of a stationary time series. The differences between these segments will be "detected" with a high probability by any test of significance. (This criticism is, of course, not justified in the case of experimental designs in which the segments are derived from an external criterion.)

2. For each of the segments formed in step 1, a linear deterministic trend component is estimated and separated from the rest.

Since only linear components are considered, and not parabolic, cubic, etc. trends as well, it must be assumed that these unidentified components are still present in the estimated residuals. This could result in additional misinterpretations of the data.

3. After separating the linear trend components, an attempt is made to separate those parts which cause an autocorrelation ("whitening") by fitting ARIMA models to the rest. In other words, one tries to estimate the residuals.

The identification of ARIMA models is essentially based on estimates of the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The evaluation of these estimates is done by means of asymptotic confidence intervals. It is obvious that the number of 50 points of measurement sometimes mentioned is far too small to enable reliable identification and fitting of an ARIMA model. This number might be sufficient if we could be sure that the underlying model can be described by only a small number of parameters, e.g. for an AR(1) or MA(1) process. But even in such a case – and we can never be sure that such a model holds for real data – we must assume that the estimates of the parameters will still fluctuate considerably. In other words, only very long empirical time series will enable

reliable identification and yield reliable parameter estimates. Without this assumption, no reliable estimates of the residuals can be expected.

4. The estimated residuals, which are assumed to be uncorrelated, are added to the trend components separated in step 2. The resulting values are used to compare the segments found in step 1 by means of the non-parametric Kruskal-Wallis test.

Even if it is assumed that the values calculated in place of the original measurements were indeed uncorrelated – and several reasons why one should not expect this assumption to be correct have been given above – we generally cannot conclude that these values are independent. It is a particular property of the multivariate normal distribution that independence can be concluded from uncorrelatedness. However, if we were able to assume a multivariate normal distribution for the resulting values – a rather doubtful assumption to make – there would be no need to use the non-parametric Kruskal-Wallis test.

To summarize, nobody using Kleiter's (1986) approach for the evaluation of real single-case data can be certain that any effects identified in the data are not simply artefacts. This is, of course, also true for the study conducted by Möller et al. (1989). As a dependent variable these authors used a rating scale, the Adjective Mood Scale, which obviously gives measurements on ordinal-scale level at the most. All of the four steps of Kleiter's approach described above require the assumption of at least interval-scale level. Since the values assigned to the categories of rating scales are chosen rather arbitrarily, and as they can be transformed by a monotone transformation without changing the inherent information, the results obtained are also rather arbitrary. The authors considered time series of 38 and 28 points of time, respectively. From the above, it should be clear that with time series which are so short it is neither possible to identify time series models, nor to obtain reliable parameter estimates or reliable estimates of the residuals. Thus it can neither be assumed that uncorrelated residuals have resulted nor that the results of the statistical tests are to be trusted.

Conclusions

As we have seen, Kleiter's approach cannot be recommended for single-case evaluations. Therefore, the question arises as to which methods could be used instead. As indicated in the Introduction, this question is difficult to answer, though parametric and non-parametric procedures for time-series evaluation have been described in hundreds of articles. However, the application of any of these procedures is only justified if we assume that a certain dependence model holds for the underlying process. For rather simple models, e.g. for processes with stationary independent increments (Bell et al. 1970; Krauth 1981a), it is possible to derive from the time series independent values to which one can apply the usual significance tests. If such assumptions are justified, even very short time series can be evaluated. Unfortunately, it is very difficult to find such a justification for real data. Empirical identification would require long time series and, as discussed above, we could still not be sure of achieving a correct identification.

One alternative approach which can also be used with relatively short time series was described by Edgington (Edgington 1975, 1987; Krauth 1981b). In this method, it is important that the number of points of time observed is fixed prior to the commencement of the study. Additionally, the point of intervention must be chosen in a strictly random way. It is then possible to derive small-sample and large-sample non-parametric randomization tests for some common time series designs. Though these significance tests are valid from a statistical point of view and though it is not necessary to make any statistical assumptions which are difficult to justify, some problems still remain. One problem is that in clinical research it is often not feasible to randomly select the point of intervention. Another problem lies in the interpretation of significant results. Often one cannot be sure that the found significant effect of an intervention is related to those parameters to which it is attributed. If one-sided test problems of the same direction are considered for all single-subject designs, it is possible to combine the *P*-values from the various tests in order to combine the results of the different single-case studies. One of the numerous possibilities of doing this was also described by Edgington (Edgington 1972; Krauth 1988).

All in all, the best way to establish effects of sleep deprivation on depressive patients in the study by Möller et al. (1989) might have been to plan single-subject designs as suggested by Edgington, to perform appropriate randomization tests, and then to combine the test results via the *P*-values. In this particular case, it must be assumed that the observed scores of the Adjective Mood Scale are merely on an ordinal-scale level. For this reason, a rank transform of these scores should be performed prior to conducting the randomization tests.

References

Albers W (1978a) Testing the mean of a normal population under dependence. Ann Statist 6:1337–1344

Albers W (1978b) One-sample rank tests under autoregressive dependence. Ann Statist 6:836–845

Andersen AH, Jensen EB, Schou G (1981) Two-way analysis of variance with correlated errors. Int Statist Rev 49:153–167

Bell CB, Woodroofe M, Avadhani TV (1970) Some nonparametric tests for stochastic processes. In: Puri ML (ed) Nonparametric techniques in statistical inference. University Press, Cambridge, pp 215–258

Box GEP, Jenkins GM (1976) Time series analysis, forecasting and control, revised edition. Holden-Day, San Francisco

Edgington ES (1972) An additive method for combining probability values from independent experiments. J Psychol 80:351–363

Edgington ES (1975) Randomization tests for one-subject operant experiments. J Psychol 90:57–68

Edgington ES (1987) Randomization tests, 2nd edn. Dekker, New York

Gardner W, Hartmann DP, Mitchell C (1982) The effects of serial dependence on the use of χ^2 for analysing sequential data in dyadic interactions. Behav Assess 4:75–82

Gastwirth JL, Rubin H (1971) Effect of dependence on the level of some one-sample tests. J Am Statist Assoc 66:816–820

Gastwirth JL, Rubin H, Wolff SS (1967) The effect of autoregressive dependence on a nonparametric test. IEEE Trans Inf Theory IT-13:311–313

Gleser LJ, Moore DS (1985) The effect of positive dependence on chi-squared tests for categorical data. J R Statist Soc B47:459–465

Hibbs DA (1974) Problems of statistical estimation and causal inference in time-series regression models. In: Costner HL (ed) Sociological methodology 1973–1974. Jossey-Bass, San Francisco, pp 252–308

Kendall MG, Stuart A (1966) The advanced theory of statistics, vol 3. Design and analysis, and time-series. Griffin, London

Kleiter EF (1986) HTAKA Hierarchische Trend-Abschnitt-Komponenten-Analyse. Ein Verfahren zur Analyse von Zeitreihen. Z Emp Pädagogik Pädagogische Psychol [Suppl 2]

Krauth J (1980) Possible misinterpretations when evaluating psychological time series. Arch Psychol 133:139–147

Krauth J (1981a) Nichtparametrische Ansätze bei Zeitreihenanalysen. In: Janke W (ed) Beiträge zur Methodik in der differentiellen, diagnostischen und klinischen Psychologie. Hain, Meisenheim, pp 26–45

Krauth J (1981b) Statistische Methoden der Veränderungsmessung. In: Baumann U, Berbalk H, Seidenstücker G (eds) Klinische Psychologie – Trends in Forschung und Praxis, vol 4. Huber, Bern, pp 98–131

Krauth J (1988) Distribution-free statistics: an application-oriented approach. Elsevier, Amsterdam

Möller HJ, Blank R, Steinmeyer EM (1989) Single-case evaluation of sleep-deprivation effects by means of nonparametric time-series analysis (according to the HTAKA model). Eur Arch Psychiatry Neurol Sci 239:133–139

Moore DS (1982) The effect of dependence on chi-squared tests of fit. Ann Statist 10:1163–1171

Nicolich MJ, Weinstein CS (1981) The use of time series analysis and *t* tests with serially correlated data tests. J Exp Educ 50: 25–29

Tavare S, Altham PME (1983) Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. Biometrika 70:139–144

Walsh JE (1947) Concerning the effect of intraclass correlation on certain significance tests. Ann Math Statist 18:88–96