Eur Arch Psychiatry Clin Neurosci (1999) 249 : 21–27

ORIGINAL PAPER

B. Bandelow · E. Brunner · D. Beinroth · L. Pralle
A. Broocks · G. Hajak · E. Rüther

Application of a new statistical approach to evaluate a clinical trial with panic disorder patients

Received: 3 March 1998 / Accepted: 9 October 1998

Abstract In clinical trials in psychiatry, changes in severity are usually measured with ordinal level scales which are applied repeatedly during the trial, showing a constant decline in psychopathology scores as treatment leads to improvement. Previous non-parametric tests for repeated measures in factorial designs did not test the hypothesis that scale scores decrease constantly during the trial. A recently developed "rank test for ordered alternatives in a mixed model" was developed and applied to the data of a clinical trial in panic disorder. Thirty-seven outpatients with panic disorder and agoraphobia (PDA) were treated with imipramine (75–150 mg/day) in an 8-week open, prospective trial. Patients with intercurrent agoraphobia were instructed in practising self-exposure in their agoraphobic situations. The total score on the Panic and Agoraphobia Scale, the Hamilton Anxiety Scale (HAMA) and the Clinical Global Impression Scale (CGI) were used as the main efficacy measures. The new rank test showed significant treatment results in all scales applied. Treatment results were excellent, as was shown by a decrease in the average Panic and Agoraphobia Scale severity scores from 28.9 (range 14–45) to 13.3 (range 0–37; rank statistic $T_n = 6.7$; $p < 0.0001$). The largest effect size r_w of all clinician-rated scales was seen with the observer-rated version of the Panic and Agoraphobia Scale, although closely followed by the CGI and the HAMA. Among the self-rated scales, the Panic and Agoraphobia Scale also showed the largest effect size. All five subscores of the Panic and Agoraphobia Scale showed significant improvements. The highest treatment effect sizes were seen in the "panic attacks" subscore, followed by the "anticipatory anxiety" subscore. The new statistical test applied in this study, which has some advantages in comparison with previously applied tests, is suitable for psychiatric treatment evaluations since it can also be applied in the case of discrete repeated measurements.

Key words Panic disorder · Agoraphobia · Imipramine · Self-exposure · Rank test

B. Bandelow (✉) · D. Beinroth · A. Broocks · G. Hajak
E. Rüther
Department of Psychiatry, University of Göttingen,
Göttingen, Germany

E. Brunner · L. Pralle
Department of Medical Statistics, University of Göttingen,
Göttingen, Germany

Present address:
B. Bandelow, von-Siebold-Strasse 5, D-37075 Göttingen, Germany
e-mail bbandel@gwdg.de,
Tel.: +49-551-396610, Fax: +49-551-392004

Introduction

The statistical evaluation of clinical trials in psychiatry should be performed with non-parametric methods because the rating scales usually used in psychiatry produce data of ordinal (rank) level (e.g. with ranks such as 0 = not ill, 1 = mildly ill, 2 = moderately ill, 3 = severely ill, 4 = extremely severely ill). No psychiatrist would claim that the difference between "mild" and "moderate" would be exactly the same as between "moderate" and "severe". Note that the assignment of numbers to the levels of the grading scale is arbitrary. Instead of a 0–4 scale, one might as well use a 1–5 or a 1, 3, 5, 9, 15 scale. The results of parametric tests such as the *t*-test or the analysis of variance (ANOVA) are not independent of the choice of the grading scale. Clearly, these numbers cannot be regarded as observations from a normal distribution. This assumption, however, is required when parametric tests such as the *t*-test or ANOVA are applied. Non-parametric tests, particularly rank procedures, are independent of the choice of the grading scale and of the assumptions of a normal distribution. This is particularly important when sample sizes are small, as they usually are in psychiatric trials. Some statisticians believe that the central limit theorem might apply when the sample size is large enough (i.e. more than 40 subjects in every treatment group), thus allowing the use of parametric methods. Even in this case, one would be on the safe side by using non-parametric methods.

However, it is problematic to apply standard rank statistics for the analysis of data from a clinical trial in which repeated measures are usually being performed, i.e. rating scales are applied every week to measure treatment efficacy and not only at baseline and endpoint. One of the reasons for weekly assessments is that if a patient drops out of the trial, at least the previous week's scores are still available and the patient can still be evaluated (last-observation-carried-forward method). However, when repeated measures are being used, it would be an unnecessary expenditure of time and money to compare only the last observation with the baseline score and to discard all other observations obtained at weeks 2, 3, 4, etc. Also, it is not advisable to compare all weekly observations separately with baseline scores (i.e. week 1 with baseline, week 2 with baseline, etc.). By doing so, the risk of a type-I error would increase because of the multiple statistical tests applied. Moreover, a psychiatrist wants to have an answer to the question, "Which treatment was more successful over the whole 8-week study period?" instead of "Which treatment was more successful in weeks 2 and 5, but not in weeks 1, 3, 4, 6, and 7?"

Furthermore, one has to be concerned about the interdependencies of the weekly observations. Data observed repeatedly on the same patient over time (usually called "longitudinal data") are dependent on each other, whereas the structure of the dependencies with respect to time is usually unknown. Standard rank tests such as the Wilcoxon-Mann-Whitney or the Kruskal-Wallis tests require independence of the observations. This assumption is usually not fulfilled in clinical trials. Patients improving between weeks 3 and 4 have a good chance of further improving between weeks 4 and 5. In other words, weekly rating scores usually are highly intercorrelated. In other rank tests, such as in the case of the Friedman test (Friedman 1937) or the Page test (Page, 1963), the same type of dependencies between repeated observations for all time points is required (compound symmetry). This assumption usually is also not fulfilled in clinical trials, because scores obtained on time points that are closer together (e.g. weeks 2 and 3) correlate higher than those obtained at distant time points (e.g. weeks 2 and 8); thus, none of the standard rank tests is appropriate for the analysis of longitudinal data.

Based on some recent advances in the asymptotic theory of rank statistics under dependencies (Brunner and Denker 1994), rank statistics for the analysis of repeated measures or longitudinal data have been developed (Akritas and Brunner 1996, 1997; for a detailed description see Methods). Usually, the results of a psychiatric trial follow a certain pattern which consists of a more or less constant decrease in rating scale scores in the case of effective treatment (e.g. see Table 3). It would be less probable to find a pattern in psychiatric studies where severity scores decrease in the first week, then increase, to decline again at the end of the study.

A linear rank statistic T_n, which is a weighted sum of the rank means for the different time points, has been suggested (Akritas and Brunner 1996). The statistic T_n (analogous to the Page test) is especially sensitive to such "patterned" effects such as an increasing, decreasing or "umbrella-type" trend of the rank means, according to the selected weights of the statistic T_n. By selecting a special pattern of weights, a researcher might also test alternative hypotheses, e.g. that a certain treatment leads to an initial aggravation and then later to an improvement of symptoms. It is noteworthy that the well-known Page test uses the same method but under the assumption of compound symmetry which is rarely appropriate for longitudinal data. Thus, the Akritas-Brunner (1996) statistic T_n is a generalization of the Page test for repeated-measures designs.

The population studied in this investigation contained two sub-populations: patients with panic disorder with agoraphobia (PDA) and another group with panic disorder without agoraphobia. The questions of whether the treatment effects over time or whether the trends over time are the same for the two groups of patients can also be analysed by the rank test.

The analysis of factorial designs by rank methods is enabled by the method of "non-parametric hypotheses" which was introduced by Akritas and Arnold (1994) for continuous distributions. Later, Akritas and Brunner (1997) extended this idea to general distributions (including the case of ties, i.e. when two subjects have exactly the same score) and general factorial designs.

The new rank procedure might be the method of choice for most clinical trials in psychiatry because it overcomes some problems that occurred with previous methods. It is more appropriate for psychiatric treatment evaluations than previously applied tests, since it can also be applied in the case of discrete repeated measurements.

Less statistically experienced readers might ask why so many clinical trials in psychiatry that have been evaluated by means of the Wilcoxon-Mann-Whitney test, the Kruskal-Wallis test or even parametric methods, such as ANOVA, have been published in refereed journals (Munzel and Bandelow 1998). The answer might be that the conclusions drawn from these studies were not necessarily wrong, but may have been subject to some kind of inexactness. In most cases, a more appropriate statistical test might yield the same results as previously applied methods but come to different conclusions in other cases. New statistical approaches are being developed to draw conclusions more safely and to detect the real conditions in the population more precisely.

It may be a problem that in standard rating scales total scale scores are derived from the addition of ordinal level data, because it is not always known whether the different items of a scale contribute equally to the total score. However, if they do not, then adding item scores to a total score may even be more problematic when using parametric methods.

Our objective was to apply the new statistical approach in a clinical trial with panic disorder patients in which treatment methods of proven efficacy were being applied. Since 1964, the efficacy of the tricyclic antidepressant imipramine has been shown in many trials with PDA patients (e.g. Klein 1964; CNCPS 1992). Behaviour thera-

Table 1 Sample characteristics (n = 37)

Panic Disorder with Agoraphobia	24	(64.9%)
Panic Disorder without Agoraphobia	13	(35.1%)
Age (years)	32.7	(SD 6.3)
Gender: female	26	(70.3%)

pies involving exposure of the patient to agoraphobic situations have also been proven to be effective in PDA. After being instructed by the therapist, self-exposure can be performed by the patient himself without the presence of the therapist. Self-exposure has been shown to be as effective as therapist-aided exposure (Al-Kubaisy et al. 1992).

Patients and methods

Thirty-eight patients with PDA according to DSM-IV or ICD-10, all outpatients of the anxiety disorders unit at the Department of Psychiatry, University of Göttingen, consented to take part in a study to investigate the efficacy of imipramine and self-exposure. Diagnoses were made using the German version of the Structured Clinical Interview for DSM-III-R (Wittchen et al. 1990). One patient dropped out before week 2 and was not evaluated. Sample characteristics are given in Table 1.

Exclusion criteria were other psychiatric disorders, physical disorders, pregnancy and insufficient contraceptive methods. Patients who were taking medications withdrew from them at least 1 week (2 weeks in the case of benzodiazepine intake) before the beginning of the study.

Treatment

The study design was approved by the institutional ethics committee. The study was conducted as an open trial. Imipramine was given in daily doses between 75 and 125 mg daily on a t.i.d. schedule. The only medication allowed was promethazine 25 mg as needed in cases of acute anxiety. Only one patient took 25 mg promethazine once in the first treatment week. All patients were given a leaflet that briefly explained the symptoms of PDA. Patients with combined PDA were instructed to perform self-exposure by not avoiding feared situations (Al-Kubaisy et al. 1992). Patients had the rationale of self-exposure explained at the first session. Patients were asked to expose themselves intentionally to phobic stimuli (e.g. crowds, shopping malls, public transport, elevators). Strategies for coping during exposure were discussed. Patients were reassured repeatedly and given positive approval at each visit whenever they had reduced their avoidance behaviour.

Treatment lasted for 8 weeks. Throughout the study, patients were required to return to the outpatient department every 2 weeks for efficacy measurements. Patients who completed at least 2 weeks of treatment were evaluated.

Rating scales

Treatment outcome was assessed with the Panic and Agoraphobia Scale (Bandelow 1995). This scale is available as an observer-rated and a self-rated version, both of which were applied to the patients. Both scales have 13 identical items with 5 answer choices (from 0 to 4). Of the 13 items, five subscores were assessed separately: (a) panic attacks; (b) avoidance behaviour; (c) anticipatory anxiety; (d) disability; and (e) worries about health. The observer-rated version was used as the main efficacy measure of the study, together with the Hamilton Anxiety Scale (HAMA; Hamilton 1959) and the Clinical Global Impression Scale (CGI; NIMH 1976). Only the item "severity" of the 1–7 points version was used (1 = normal, not at all ill; 2 = borderline mentally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; 7 = among the most extremely ill patients). Additionally, the following self-ratings were filled out by the patients and used as secondary efficacy measures: State Trait Anxiety Inventory (STAI; Spielberger et al. 1979) the Agoraphobia Subscale of the Fear Questionnaire (FQ; Marks and Matthews 1979), the Mobility Inventory For Agoraphobia (MIFA; Chambless et al. 1985) and the Patients Clinical Impression Scale (PGI), a global severity scale corresponding to the CGI 1–7 Likert Scale.

Statistical evaluations

Rating scale data were obtained at five timepoints: at baseline and at weeks 2, 4, 6, and at endpoint (week 8). Treatment success was measured by comparing the marginal distributions of the scale. A last-observation-carried-forward analysis (LOCF) was performed, i.e. in the case of a drop-out, the last observation was used as endpoint score (but only if the patient had been in the study for at least 2 weeks). Additionally, patients with or without agoraphobia were assessed separately to see if the treatment had differential effects on these diagnostic groups. For simplicity we only present the mathematical derivation for one group.

Rating scale data where evaluated by means of the rank test described by Akritas and Brunner (1996, 1997). This test is appropriate for typical clinical studies in psychiatry, where ordinal level data, such as rating scale scores, are obtained in weekly assessments (repeated measures).

In this test, the observations X_{jk} (e.g. Panic and Agoraphobia Scale total scores) from all patients (k = 1, …n) and all j = 1, … 5 time points are placed in order of rank. Mid-ranks are used in cases of ties. Let $\bar{R}_{j\bullet}$ denote the arithmetic mean of the ranks of all patients at time point j. For comparing the results of two time points, e.g. j and j', the difference $\bar{R}_{j\bullet} - \bar{R}_{j'\bullet}$ of the ranks for these two time points is considered. For a large number of subjects, this difference, standardized by $s = \sqrt{(n-1)^{-1}\sum_{k=1}^{n}(R_{jn} - R_{j'k} - \bar{R}_{j\bullet} + \bar{R}_{j'\bullet})^2}$ has a normal distribution under the hypothesis of no time effect (see Akritas and Brunner 1997). If a certain treatment is successful, it is expected that the scale scores should decline gradually over time. The hypothesis "no time effect" is tested against the alternative "decreasing trend". In order to test this hypothesis, we multiply the rank means by weights w_j which reflect the conjectured decreasing trend. The weights $w_1 = 2$, $w_2 = 1$, $w_3 = 0$, $w_4 = -1$, and $w_5 = -2$ are used (note that $\sum_{j=1}^{5} w_j = 0$ is required). The statistic T_n for trend alternatives is derived from the rank means $\bar{R}_{j\bullet}$ by multiplying each $\bar{R}_{j\bullet}$ by a weight w_j. The vector of the weights $w = (w_1, \ldots, w_5)'$ reflects the conjectured pattern of the alternative. For a large sample size, the statistic $T_n = \sqrt{n}\sum_{j=1}^{5} w_j \bar{R}_{j\bullet}$ has a normal distribution with the mean 0 and the variance σ^2. The variance σ^2 is estimated from the ranks by $\hat{\sigma}^2 = (n-1)^{-1}\sum_{k=1}^{n}(Z_k - \bar{Z}_\bullet)^2$, where $\bar{Z}_\bullet = n^{-1}\sum_{k=1}^{n} Z_k$ and $Z_k = \sum_{j=1}^{5} w_j R_{jk}$. For small sample sizes, the distribution of T_n is approximated by a central t_v-distribution with $v = n - 1$ degrees of freedom (for details see Akritas and Brunner 1996).

Because this statistical method is new, an additional ANOVA was performed. Although the application of the ANOVA is problematic with this kind of data, as was explained previously, the majority of psychiatric studies use this method for statistical evaluations.

Effect sizes

When treatment effects are measured by two or more different rating scales (e.g. the CGI and HAMA), the effect sizes measured on these different scales cannot be compared directly, because the scales differ in the number of items that are added to a total score, the number of possible answer choices in each item and in their statistical parameters (mean and standard deviation). To make the effect sizes of different scales comparable, meta-analysis procedures have been developed. Rosenthal's *r* (Rosenthal 1984) is a measure for effect sizes that is independent of the different scales' properties. However, Rosenthal's *r* is dependent on statistical parameters, such as the standard deviation, and thus is not suitable for ordered categorical data. For the comparison of effect sizes of ordinal level data obtained with rating scales, we suggest a "rank version" of Rosenthal's *r*. Since the distribution of T_n is approximated by the t_v-distribution under the hypothesis, we suggest the following non-parametric effect size:

$r_w = \sqrt{\frac{T_n^2}{T_n^2 + n - 1}}$, where T_n is the linear rank statistic and *n* the total number of patients.

Note that r_w does not depend on the special choice of the grading scale since T_n does not depend on it. Moreover, r_w is a measure for effect sizes of several samples. In the case of two samples, it reduces to the two-sample rank version of Rosenthal's *r*. The proposed non-parametric effect size r_w ranges between 0 and 1, where 1 represents the highest effect size.

Additionally, the results were analysed by an ANOVA for repeated measures, a method most often used for clinical trials in panic disorder (Munzel and Bandelow 1998).

Statistical tests were performed with the "Statistical Analysis System for Windows" (SAS 6.11), MIXED procedure and a specially written SAS macro program for the rank test. This program LD-F1 is available via the World-Wide Web at ftp://ftp.ams.med.uni-goettingen.de/nonpar.

Results

Patients received a mean dose of 97.1 mg imipramine (SD 16.5) during the trial. One patient was not evaluated because he did not reappear at week 2. Of the 37 remaining evaluable patients, two terminated the study after week 2, two after week 4 and two after week 6 (Table 2). Reasons for early study termination were: unwanted side-effects ($n = 1$), lack of efficacy ($n = 1$), difficulty taking time off work ($n = 1$), patient did not reappear and could not be reached ($n = 1$), and patients remitted early and were therefore not willing to continue the medication ($n = 2$).

The mean scores of all patients are shown in Table 3. The raw scores can be obtained from the corresponding author. Results of the rank test for ordered alternatives in a mixed model are given in Table 4. The biweekly scale scores were tested against baseline. With the exception of the HAMA, FQ Agoraphobia, and MIFA, results of all scales showed significant improvement after the second treatment week. Moreover, overall improvement over the whole study period was tested by using an ordered-alternatives model for repeated measures. With this model, a constant decrease in scale scores could be shown with high significance. The results were confirmed by ANOVA for repeated measures.

Effect sizes

Treatment resulted in high effect sizes (0.51–0.75). The largest effect size r_w of all observer-rated scales was seen with the observer-rated version of the Panic and Agoraphobia Scale (Table 3), although closely followed by the CGI and the HAMA. Among the self-rated scales, the Panic and Agoraphobia Scale also showed the largest effect size. Because of the small sample size and the high number of possible comparisons, significance tests were not performed on the effect of size differences.

Table 2 Early study termination. *n* no. of patients still in the study

Time	*n*
Baseline	38
Week 2	37
Week 4	35
Week 6	33
Week 8	31

Table 3 Mean scale scores for 37 patients; last-observation-carried-forward analysis (LOCF). Standard deviations in parentheses. *P&A* Panic and Agoraphobia Scale; *CGI* Clinical Global Impression; *HAMA* Hamilton Anxiety Scale; *PGI* Patients' Global Impression; *STAI* State-Trait-Anxiety Inventory; *FQ* Fear Questionnaire; *MIFA* Mobility Inventory for Agoraphobia. r_w non-parametric effect size (pre- and postcomparison)

Scales	Baseline	Week 2	Week 4	Week 6	Week 8	Effect sizes
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	r_w
Clinician ratings						
P&A (observer)	28.9 (8.1)	23.1 (12.0)	17.2 (9.5)	15.2 (11.6)	13.3 (11.8)	0.74
CGI	4.9 (0.8)	4.4 (1.1)	4.0 (1.0)	3.6 (1.4)	3.5 (1.6)	0.72
HAMA	24.5 (8.8)	23.1 (11.5)	19.2 (9.4)	17.9 (9.7)	13.5 (10.4)	0.71
Self-ratings						
P&A (self-rating)	30.0 (9.2)	24.6 (11.4)	19.0 (10.6)	16.4 (12.5)	13.7 (12.8)	0.75
PGI	5.5 (1.0)	4.9 (1.1)	4.1 (1.2)	3.8 (1.3)	2.8 (1.5)	0.74
STAI	57.3 (13.4)	53.7 (12.1)	47.3 (10.4)	44.2 (14.3)	42.4 (14.3)	0.62
FQ Agoraphobia	16.9 (12.5)	15.5 (11.4)	14.6 (11.2)	11.7 (9.2)	9.7 (8.8)	0.58
MIFA	101.2 (43.3)	102.9 (42.0)	96.1 (40.0)	92.8 (38.9)	86.2 (40.4)	0.51

Table 4 Statistics and *p*-values for the LOCF analysis; rank test for ordered alternatives in a mixed model; one-tailed test ($df = 35$); repeated-measures analysis of variance (ANOVA). T_n linear rank statistic; *F* statistic of ANOVA; *n.s.* not significant (Bonferroni-Holm correction; Holm 1979)

Scales	Week 2 vs baseline	Week 4 vs baseline	Week 6 vs baseline	Week 8 vs baseline	Repeated measures; rank test for ordered alternatives	ANOVA
	T_n *p*	T_n *p*	T_n *p*	T_n *p*	T_n *p*	*F* *p*
Clinician ratings						
P&A	3.5	8.2	6.7	7.2	6.7	13.8
(observer)	0.0006	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
CGI	2.6	5.3	5.3	6.7	6.3	12.5
	0.006	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
HAMA	0.99	3.7	4.2	5.8	6.2	7.1
	0.16 (n.s.)	0.0004	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Self-ratings						
P&A (self-rating)	3.6	8.7	6.9	7.3	7.1	12.3
	0.0005	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
PGI	3.3	6.5	7.4	6.8	7.0	17.7
	0.001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
STAI	2.0	4.9	4.5	4.9	5.1	8.6
	0.03	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
FQ Agoraphobia	1.1	3.9	3.7	5.4	4.4	2.7
	0.19 (n.s.)	0.001	0.0003	0.0001	0.006	< 0.05
MIFA	0.6	2.0	2.6	3.8	3.9	2.8
	0.29 (n.s.)	0.025	0.007	0.0002	0.0002	< 0.05

Table 5 Subscores of the Panic and Agoraphobia Scale (observer rating). Statistics and *p*-values (LOCF analysis); rank test for ordered alternatives in a mixed model; one-tailed test ($df = 36$); T_n linear rank statistic; r_w effect size (Bonferroni-Holm correction; Holm 1979)

Subscore	Baseline Mean (range)	LOCF Mean (range)	T_n	*p*	Effect sizes r_w
Panic-attacks	2.7 (1.7–4.0)	1.2 (0–3.0)	8.0	7.3×10^{-10}	0.81
Avoidance behaviour	2.0 (0–3.7)	1.0 (0–3.0)	4.4	5.2×10^{-5}	0.60
Anticipatory anxiety	2.8 (0–4.0)	1.2 (0–3.0)	7.4	4.6×10^{-9}	0.75
Disability	1.7 (0–3.7)	1.2 (0–3.0)	4.8	1.5×10^{-5}	0.59
Worries about health	2.0 (0–4.0)	1.2 (0–3.0)	6.2	1.6×10^{-7}	0.68

Subscores of the Panic and Agoraphobia Scale

In Table 5 the improvements in the different subscores of the Panic and Agoraphobia Scale (observer rating) are shown. All subscores showed significant improvements. The highest treatment effect sizes were seen in the "panic attacks" subscore, followed by the "anticipatory anxiety" subscores (Table 5). "Worries about health", "avoidance behaviour" and "disability" were less markedly influenced by the treatment with imipramine and self-exposure. Similar results were found for the self-rated version of the scale but are not shown due to space limitation.

Patients with or without agoraphobia

Patients with and without agoraphobia were assessed separately (Table 6). There were no significant diagnosis/time interactions (Table 7), i.e. no differential treatment effects were detected in the sense that treatment was more effective or improvement occurred earlier in one group than in the other. As there was a significant main "diagnosis" effect (a significant difference in the scale scores between the patients with or without agoraphobia) in all scales except the CGI, and a significant main "time" effect (a significant decrease in scale score over time) was shown, the data show that there was a parallel decrease in scale scores in both diagnostic groups.

Adverse events

The patients complained of the following side effects: dry mouth, sedation, drowsiness, hypotension. Side effects diminished during treatment with imipramine: at week 2, 60.0% of the patients were free of side effects; at endpoint, 76.6%.

Table 6 Comparison of improvement in patients with and without agoraphobia (LOCF), means and ranges

Scale	*N*	Without agoraphobia				*N*	With agoraphobia			
		Baseline		LOCF			Baseline		LOCF	
		Mean	Range	Mean	Range		Mean	Range	Mean	Range
Clinician ratings										
P&A (observer)	13	22.3	14–31	9.9	0–30	24	32.4	21–45	15.1	0–37
CGI	13	4.7	4–6	2.8	1–5	24	5.0	4–7	3.4	1–7
HAMA	13	22.4	8–36	7.5	0–20	24	25.6	11–45	16.7	3–35
Self ratings										
P&A (self-rating)	13	25.1	14–41	10.2	0–30	24	32.6	18–45	15.6	0–40
PGI	13	5.4	4–7	2.9	2–6	24	5.6	4–7	3.8	1–7
STAI	12	57.3	30–72	35.2	26–43	24	57.3	29–76	46.3	24–74
FQ Agoraphobia	12	9.0	0–26	5.8	0–2	24	20.8	6–40	11.9	0–31
MIFA	12	76.7	49–153	64.8	49–119	22	114.6	49–230	97.0	49–196

Table 7 Main effects of diagnosis (patients with agoraphobia, N_0, vs patients without agoraphobia, N_1) and time. Interaction effects diagnosis by time. *n.s.* not significant

Scale	N_0	N_1	Diagnosis *p* (*F* approx).	df_1 df_2	Time *p* (χ^2/f approx.)	*df*	Diagnosis/time interaction (patterned) *p*/*t* approx.	*df*
P&A (observer)	13	24	9.29 0.0054	1 25.02	79.90 0.0000	2.45	1.26 0.11 (n.s.)	17.70
CGI	13	24	3.82 0.0618 (n.s.)	1 25.28	63.59 0.0000	2.86	0.12 0.45 (n.s.)	14.05
HAMA	13	24	12.43 0.0015	1 28.30	55.39 0.0000	2.96	–1.28 0.89 (n.s.)	13.45
P&A (self-rating)	13	24	4.93 0.0371	1 21.69	92.02 0.0000	2.27	0.28 0.39 (n.s.)	18.18
PGI	13	24	7.35 0.0119	1 25.30	95.02 0.0000	2.66	–0.87 0.80 (n.s.)	13.53
STAI	11	23	4.97 0.0352	1 24.27	39.12 0.0000	1.78	–1.90 0.95 (n.s.)	5.97
FQ Agoraphobia	11	24	8.04 0.0109	1 18.25	30.35 0.0000	1.97	0.34 0.37 (n.s.)	22.97
MIFA	11	22	9.94 0.0053	1 18.63	18.19 0.0001	2.06	0.80 0.22 (n.s.)	27.31

Discussion

The efficacy of imipramine in combination with self-exposure in the treatment of panic disorder was demonstrated by applying recently developed statistical methods, the "rank procedures for longitudinal data". The main limitation of the study was that it was conducted as an open trial. In panic disorder trials, a high placebo response is a well-known fact (Coryell and Noyes 1988). Another study evaluating the feasibility of this new statistical approach in a double-blind, placebo-controlled trial was published recently (Broocks et al. 1998).

The largest effect size of all observer-rated scales was seen with the observer-rated version of the Panic and Agoraphobia Scale, although closely followed by the CGI and the HAMA. Among the self-rated scales, the self-rated version of the Panic and Agoraphobia Scale also showed the largest effect size. The Panic and Agoraphobia Scale has the advantage that different aspects of panic disorder, e.g. agoraphobia or anticipatory anxiety, can be assessed separately by using the five subscores of the scale. All five subscores of the new Panic and Agoraphobia Scale showed significant improvements. The highest treatment effect sizes were seen in the "panic attacks" subscore, followed by the "anticipatory anxiety" subscores. "Worries about health", "avoidance behaviour" and "disability" were less markedly influenced by the treatment with imipramine and self-exposure.

When patients with or without agoraphobia were assessed separately, no differential treatment effects were detected in the sense that treatment with imipramine and self-exposure was more effective or improvement occurred earlier in one diagnostic group than in the other; however, the number of patients in the two subgroups may have been too small, implying a type-II error.

References

Akritas MG, Arnold SF (1994) Fully non-parametric hypothesis for factorial designs. I. Multivariate repeated measures design. J Am Statist Assoc 89 : 336–343

Akritas MG, Brunner E (1996) Rank tests for patterned alternatives in factorial designs with interactions. In: Brunner E, Denker M (eds.) Research developments in probability and statistics. Festschrift in honor of Madan L. Puri on the occasion of the 65th birthday. VSP International Science Publishers, Utrecht, pp 277–288

Akritas MG, Brunner E (1997) A unified approach to rank tests in mixed models. J Statist Planning Inference 61 : 249–277

Al-Kubaisy T, Marks I, Logsdail S et al. (1992) Role of exposure homework in phobia education: a controlled study. Behav Ther 23 : 599–621

Bandelow B (1995) Assessing the efficacy of treatments for panic disorder and agoraphobia. II. The Panic and Agoraphobia Scale. Int Clin Psychopharmacol 10 : 73–81

Broocks A, Bandelow B, Pekrun G, George A, Meyer T, Bartmann U, Hillmer-Vogel U, Rüther E (1998) Comparison of aerobic exercise, clomipramine, and placebo in the treatment of panic disorder. American Journal of Psychiatry 155 : 603–609

Brunner E, Denker M (1994) Rank statistics under dependent observations and applications to factorial designs. J Statist Planning Inference 42 : 353–378

Chambless DL, Caputo GC, Jasin SE, Gracely EJ, Williams C (1985) The Mobility Inventory for Agoraphobia. Behav Res Ther 23 : 35–44

CNCPS (1992) Cross-national collaborative panic study. Drug treatment of panic disorder. Comparative efficacy of alprazolam, imipramine, and placebo. Br J Psychiatry 160 : 191–202

Coryell W, Noyes R (1988) Placebo response in panic disorder. Am J Psychiatry 145 : 1138–1140

Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Statist Assoc 32 : 675–699

Hamilton M (1959) The assessment of anxiety states by rating. Br J Med Psychol 32 : 50–55

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statistics 6 : 65–70

Klein D (1964) Delineation of two drug-responsive anxiety syndromes. Psychopharmacology 5 : 397–408

Marks IM, Matthews AM (1979) Brief standard self-rating for phobic patients. Behav Res Ther 17 : 263–267

Munzel U, Bandelow B (1998) The use of parametric vs nonparametric tests in the statistical evaluation of rating scales. Pharmacopsychiatry 31 : 222–224

NIMH (1976) National Institute of Mental Health. 028 CGI. Clinical Global Impressions. In: Guy E (ed) ECDEU assessment manual for psychopharmacology, revised edn. NIMH, Rockville, Maryland, pp 217–222

Page EB (1963) Ordered hypothesis for multiple treatments: a significance test for linear ranks. J Am Statist Ass 58 : 216–230

Rosenthal R (1984) Meta-analytic procedures for social research. Applied Social Research Methods Series, vol 6. Sage, Beverly Hills

Spielberger CD, Gorsuch RL, Lushene RE (1979) State-trait anxiety inventory (“Self-evaluation Questionnaire”). Consulting Psychologists Press, Palo Alto, Calif.

Wittchen HU, Zaudig M, Schramm E et al. (1990) Strukturiertes Klinisches Interview für DSM-III-R. Beltz, Weinheim, Basel