Regular Article

Psychotherapy and Psychosomatics

Psychother Psychosom 2019;88:225–235
DOI: 10.1159/000500188

Received: November 12, 2018
Accepted after revision: April 8, 2019
Published online: May 23, 2019

Long-Term Effectiveness of Cognitive Behavioral Therapy in Routine Outpatient Care: A 5- to 20-Year Follow-Up Study

Ruth von Brachel[a] Gerrit Hirschfeld[b] Arleta Berner[c] Ulrike Willutzki[c] Tobias Teismann[a] Jan Christopher Cwik[a] Julia Velten[a] Dietmar Schulte[a] Jürgen Margraf[a]

[a]Mental Health Research and Treatment Center, Ruhr University Bochum, Bochum, Germany; [b]Faculty of Business and Health, University of Applied Sciences Bielefeld, Bielefeld, Germany; [c]Faculty of Psychology and Psychotherapy, University Witten-Herdecke, Witten, Germany

Keywords
Cognitive behavioral therapy · Long-term · Follow-up · Routine care · Clinically significant change

Abstract
Objective: Long-term follow-ups several years after receiving cognitive behavioral therapy (CBT) are scarce and most of the existing literature describes follow-up data of randomized-controlled trials. Thus, very little is known about the long-term effects of CBT in routine care. ***Methods:*** We investigated psychological functioning in a sample of 263 former outpatients who had received CBT for a variety of mental disorders such as depression, anxiety-, eating- or somatoform disorders 8.06 (SD 5.08) years after treatment termination. All participants completed a diagnostic interview as well as the Brief-Symptom Inventory (BSI) and the Beck Depression Inventory (BDI). Effect sizes and response rates according to Jacobson and Truax [J Consult Clin Psychol 1991;59:12–9] were calculated from pre- to posttreatment and from pretreatment to follow-up assessment. ***Results:*** Pre- to posttreatment effect sizes ranged between 0.75 (BDI) and 0.63 (BSI) and pretreatment to follow-up effect sizes were 0.92 (BDI) and 0.75 (BSI). Of all patients, 29% (BDI) and 17% (BSI) experienced clinically significant change at posttreatment and 42% (BDI) and 24% (BSI) at follow-up. ***Conclusion:*** The results point to the long-term effectiveness of CBT under routine conditions for a wide array of problems, especially when compared to the long-term effects of medical treatment. It is noteworthy that the results at follow-up were even better than at posttreatment, indicating further improvement. However, about a quarter of the patients did not respond sufficiently to therapy, neither concerning short-term nor long-term effects.

Introduction

Cognitive behavioral therapy (CBT) displays the strongest evidence among different treatments across a wide array of psychological problems [1–3]. Its short-time efficacy has widely been proven, and studies have also shown CBTs effectiveness under routine care conditions. However, this evidence is mostly restricted to studies with comparatively short (<2 years) follow-up periods [1].

E-Mail karger@karger.com
www.karger.com/pps

Dr. Ruth von Brachel
Mental Health Research and Treatment Center, Ruhr University Bochum
Massenbergstrasse 9-13
DE–44787 Bochum (Germany)
E-Mail ruth.vonbrachel@rub.de

There are only a few studies with longer follow-up durations. For example, in their meta-analysis of CBT for depression, Cuijpers et al. [4] included 117 randomized controlled trials (RCTs) that reported pre-post comparisons, while Steinert et al. [5] only found 11 studies on depression with a follow-up period of at least 2 years. Most long-term studies are follow-up investigations after RCTs. These studies mainly find outcomes (e.g., effect sizes) comparable to pre-post comparisons and superior to medical treatments [6]. These results, however, are extremely heterogeneous, and even though effect sizes were mostly large, about a third of the patients still classified as nonresponders at follow-up [5, 7–9]. When looking at different diagnoses, patients with panic disorder or generalized anxiety disorder often display the greatest amelioration of symptoms with improvements after termination of therapy [8], while patients with other disorders such as depression, social anxiety, or anorexia nervosa still benefit compared to medical or no treatment, but more often experience relapses, residual, or full symptoms at follow-up [7–9].

Since nearly all of these studies include patients from efficacy studies, that is, highly selective RCTs, they allow only limited transfer to patients in routine care who have more comorbidities and receive more heterogeneous treatments [1, 10–12]. Only a handful of effectiveness studies have tried to establish the benchmarks of the effectiveness of CBT under routine conditions. A meta-analysis of effectiveness trials for anxiety disorders [13] found comparable effect sizes in routine care. Effectiveness studies on depression have yielded inconclusive results, that is, some studies report similar outcomes to RCTs [14, 15], while others found smaller effects in routine care [15, 16]. One meta-analysis detected smaller effects, that is, effect sizes around 1.0 in effectiveness studies for depression and anxiety [17]. The makeup of the sample may be an explanation for these disparate results. If propensity score matching is used to make patients in effectiveness studies more similar to patients in RCTs, similar effect size estimates emerge [18]. This is in line with the meta-analysis of CBT effectiveness under routine conditions by Shadish et al. [19]. The authors found that RCTs that are more similar to routine care show most comparable results. Concerning levels of significant change, most studies on routine care found levels of clinically significant change around 50% [15, 20, 21].

Because there are fewer studies on routine care than on RCTs, only very few studies have investigated the long-term effectiveness of routine care [17]. These studies indicate stable effects – equal to follow-up studies of RCTs – with patients maintaining their treatment gains after 6 and after 12 months [22, 23]. To our knowledge, there are no studies to date investigating treatment outcome after several years in routine care.

Since long-term studies after treatment termination are scarce and long-term studies after routine care treatment even more so, little is known about the predictors of long-term treatment success. One RCT-follow-up study showed, for instance, that a quick symptom reduction during the first sessions leads to more favorable treatment outcomes in individuals with binge eating after 6 years [24]. In routine care, long-term treatment success seems to be predicted by fewer symptoms at posttreatment [25], fewer comorbidities, and less pretreatment pathology [26]. There are also some hints that a positive family environment and active coping skills promote long-term treatment gains [27]. These underlying mechanisms are, however, widely understudied.

The aim of the present study was to analyze long-term effectiveness of outpatient CBT under routine care conditions. In order to do so, we contacted former patients who had received CBT at a large university outpatient clinic in Germany. We describe effectiveness in terms of effect sizes and clinically significant changes. A secondary aim was to investigate pre- and post-therapy predictors of patients' status at long-term follow-up.

Methods

Treatment

Treatment was provided for by therapists in postgraduate training at the university outpatient clinic at the Department of Clinical psychology at Ruhr-University Bochum, Germany. To become a licensed therapist in Germany, one has to take part in a 3-year postgraduate training entailing practical CBT courses while working in an inpatient (1 year) and an outpatient clinic (2 years). During this time, prospective therapists have to regularly discuss their patients and their according treatment with a licensed CBT supervisor. Therapists in our facility had at least a master's degree in Psychology and a minimum of 1-year full-time postgraduate training in CBT. Treatments follow published CBT manuals for the respective disorders, but treatment is much less standardized than in RCTs. Treatments are paid for by the German insurance system, which routinely covers 25–60 sessions.

Recruitment

Participants were recruited among former patients. Between the beginning of data collection in 1990 and 2010, 3,285 patients initiated a treatment at the outpatient clinic. Patients were selected for the study in a stratified manner, that is, we contacted people from every year and month to reduce the risk of selection bias. We included all patients who had completed the first 5 diagnostic sessions. We sent a letter with information about the study to the pa-

tients' last known address and notified them that we would call them. During the phone call, they received more study information and were asked for their consent to participate. If a person could not be reached at the phone number or the address they had before therapy, we searched to find current contact information in official lists and online sources. If the search was not successful, we used the official registration office in Germany. After giving oral consent on the phone, participants were asked to give written consent and to fill out questionnaires. They could choose between Internet-based assessment (using a personal login via email) or printed versions of the questionnaires (and prepaid return postage). Furthermore, they took part in a diagnostic interview either by phone ($n = 172$; 59%), webcam ($n = 1$; <1%), or face-to-face ($n = 118$; 41%) at our treatment center.

Diagnostic Interview at Follow-Up

A semi-structured interview (MINI-DIPS) was used to assess current and all life-time DSM-IV diagnoses either by phone or in a face-to-face interview [28, 29]. We based our diagnostic on DSM-IV rather than DSM-5 because diagnoses at the beginning of therapy were also given according to DSM-IV using the SCID-IV Axis I Disorders [30]. Since the MINI-DIPS does not include detailed information on psychosis, psychotic disorders were diagnosed with the psychotic section of the SCID [30]. Both interviews display moderate to high retest reliabilities (MINI-DIPS $r_{tt} = 0.84$–1.00; SCID: $r_{tt} = 0.57$–1.00); the discriminant validity of the DIPS has also been shown for all diagnostic categories apart from sleeping disorders [31]. The complete interview took about 90 min. Patients' answers were coded regarding current and lifetime diagnoses.

Interviewers were graduate students in clinical psychology who had received about 30 h of training in conducting the interview. One of 2 licensed psychotherapists (RvB and AB) supervised all cases. Interrater reliability in the current study (before supervision), assessed by rating videos for a subset of 8 participants with 22 different diagnoses, was substantial with a kappa of 0.70 (95% CI 0.50–0.90). All cases were discussed, and disagreement was resolved in discussion with the supervisor.

Questionnaires

Brief Symptom Inventory

The Brief Symptom Inventory (BSI) [32] is a 53-item scale capturing symptom severity on nine different domains (somatization, obsession-compulsion, interpersonal sensitivity, depression, anxiety, hostility, paranoid ideation, phobic anxiety, and psychoticism) as well as general burden (global severity index) assessed by all items. Subjects indicate their agreement on a 5-point Likert scale ranging from 0 = *not at all* to 4 = *extremely*. The psychometric properties of the German version of the BSI are acceptable to good [33]. Internal consistencies vary vastly with α mostly between 0.60 and 0.80, but with α around 0.40 for the subscales phobic anxiety and psychoticism. Retest reliabilities are better with $r_{tt} = 0.73$–0.92.

Beck Depression Inventory

The Beck Depression Inventory (BDI)-I [34] is the most widely used instrument to assess depression. This instrument consists of 21 items with a forced choice scale (*0–3*) consisting of 4 different statements about none to severely depressed mood each. Patients indicate which statement describes best how they felt during the past 2 weeks. Richter et al. [35] reported internal consistencies of α = 0.88 for a psychiatric sample. Byerly and Carlson [36] reported retest reliabilities of $r_{tt} = 0.81$ in a sample of mixed outpatients. Its convergent validity with other self-report measures of depression and observer ratings is high. A minority of patients was administered the BDI-II during their therapy and therefore they received the BDI-II at follow-up. This questionnaire differs only in the phrasing of some of the items and shows the same cutoff points as the first version. Internal consistencies of the English version are between $\alpha = 0.89$ and 0.91 [37–39], test—retest reliability over 1–12 days in college students was $r_{tt} = 0.96$ [40]. Correlations with other self-report instruments measuring depression are moderate to high indicating good external validity [39].

Table 1. Demographic and clinical information on participants ($n = 256$)

Variable	Descriptive Statistic
Age, years, mean ± SD	52±12
Family status, *n* (%)	
Married (incl. long-term partners)	164 (65)
Divorced	28 (11)
Widowed	3 (1)
Single	57 (23)
Current diagnoses (most frequent diagnoses at follow-up), *n* (%)	
Social phobia	38 (14)
Specific phobia	36 (14)
Depression	29 (11)
Panic disorder	11 (4)
PTSD	9 (3)

PTSD, posttraumatic stress disorder.

Satisfaction with Therapy Scale

Patients' evaluation of their global therapy success was assessed with 3 items (slightly modified [41], asking "How much did the therapy help you?"; "Did the problems that led to your admission to the treatment center reoccur?; "Compared to right after your therapy it…" (answer categories are displayed in Table 1.

Data Analysis

Data were analyzed in 5 steps. First, we described the status of the participants at follow-up using descriptive statistics (M, SD). Differences between patients who agreed to participate in the study and those who did not were tested using *t* tests. Second, patients' ratings of their past therapy, recurrence of problems, and post-therapy improvements were tabulated. Third, effectiveness of the intervention at posttreatment and follow-up was investigated. We used Cohens' D ($ES = [\text{mean}_{t1} - \text{mean}_{t2}]/sd_{\text{pooled}}$) to quantify the magnitude of change from pre to post and pre to follow-up. Fourth, we scrutinized the clinical significance of these changes at an individual level using the methodology developed by Jacobson and Truax [42]. For this purpose, we first calculated the reliable change index to determine which patients showed change that was larger than chance fluctuations. This resulted in 3 groups, "reliable improvement," "no reliable change," and "reliable deterioration." We then used established cutoff points to determine whether change was clinically relevant. Specifically, we used a BSI score of

0.56 and a BDI score of 14.29 [43] as cut points for clinically levels of general symptoms and depression, respectively. The authors established these cutoffs for the BSI based on their data from 2 university outpatient clinics in Germany and for the BDI based on a meta-analysis with over 30,000 patients by Seggar et al. [44]. The 2 criteria were combined to define the following 5 outcome categories:

- "Clinical significant change" = Patients who reliably improved and had post scores below the cut point.
- "Reliable change not clinically significant" = Patients who reliably improved and had post scores higher than the cut point.
- "No change unproblematic" = Patients who showed no reliable change and had prescores below the cut point.
- "No-change problematic" = Patients who showed no reliable change and had prescores above the cut point.
- "Reliable deterioration" = Patients who showed reliable change in the opposite direction.

These categories were determined for BSI and BDI for both time points. Differences between post and follow-up with regard to the distribution across these outcome categories were calculated using Wilcoxon-signed rank test for paired data.

Fifth, we tested which variables were related to the status at follow-up as measured by BDI or BSI. For this, we used hierarchical linear regression. In the initial step, we only used the prescore as predictor for the follow-up score. In the second step, we added sociodemographic variables (age, family status, number of children). In the third step, we added information on the therapy (duration, additional therapies, number of medications, and the post score). The open-source programming language R was used for all analyses.

Results

Patients

We selected a random sample of the 648 former patients (Fig. 1). Of these, 233 (36%) could not be contacted because they had either died (n = 19; 8%) or the address was not obtainable based on our record or the registration office (n = 214; 92%). Of the 412 patients who were reached, 263 (64%) agreed to participate in the study and 149 (36%) did not agree to participate.

There were no systematic differences in severity of distress (assessed with the BSI) at the beginning of treatment, or improvement during therapy between those who participated and those who did not (all $p > 0.05$). Compared to those who did not agree to participate patients who agreed to participate had a longer therapy (1.13 vs. 1.33 years; $t[340] = -2.198$; $p = 0.03$; $d = 0.24$) and terminated therapy more recently (10.40 vs. 9.19 years; $t[325] = -2.248$; $p = 0.03$; $d = 0.26$).

At the time of the follow-up study, patients were between 28 and 82 years old (mean 52; SD 12). On average, 8.06 (Min = 5; Max = 20; SD 5.08) years had passed since the termination of initial therapy. Between the initial therapy and follow-up 101 of the patients (42%) started another psychotherapy, and 78 patients (32%) received psychopharmacological treatment, and 72 patients (28%) were receiving medication at the time of the follow-up. They had between none and 5 children (mean 1; SD 1). Most were married or lived in a long-term partnership (Table 2).

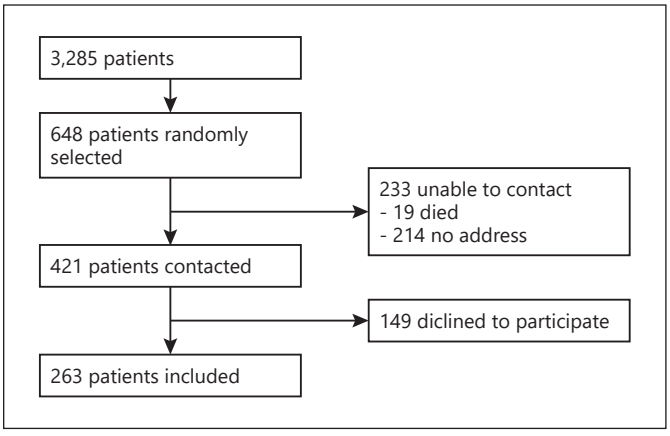
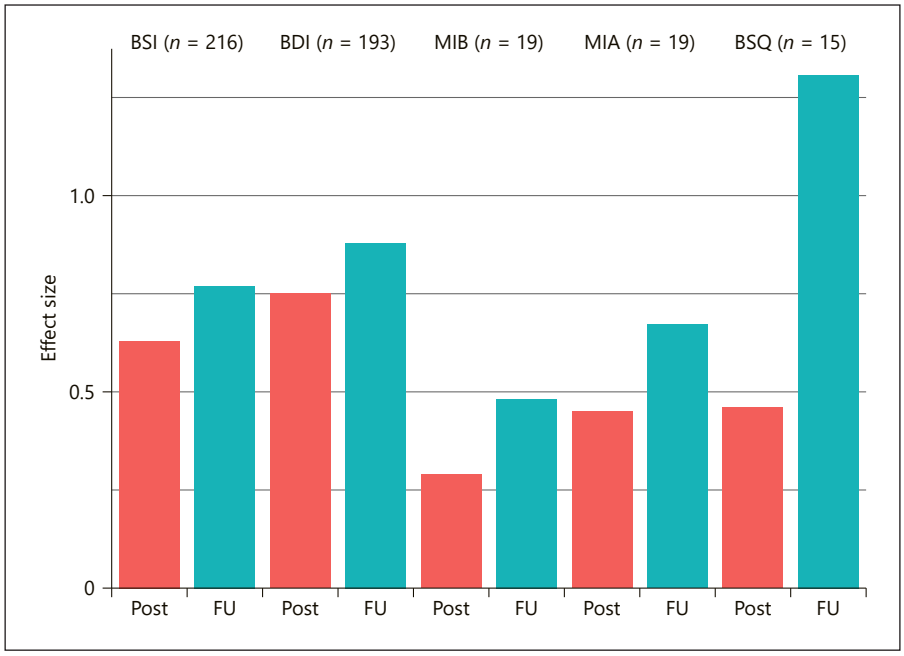


Fig. 1. Flow diagram of recruitment process.

At the beginning of therapy, data on diagnoses were available for 240 patients. The 5 most frequent diagnoses were major depression (n = 71; 27%), panic disorder (n = 44; 17%), social phobia (n = 41; 16%), specific phobia (n = 23; 9%), and obsessive–compulsive disorder (n = 13, 5%). At follow-up, 131 patients (56%) met the criteria for at least 1 current diagnosis according to the DIPS, and 51 patients (21%) met the criteria for >1 diagnosis. The most frequent current diagnoses at follow-up were social phobia (n = 38; 14%), depression (n = 29; 11%), specific phobia (n = 36; 14%), panic disorder (n = 11; 4%), and PTSD (n = 9; 3%). Comparing the frequencies of the 4 most frequent diagnoses, we found that the number of patients with depression decreased significantly from pre- to post-treatment when using retrospective information assessed with the MINI-DIPS ($\chi^2[1] = 23.35$; $p < 0.001$), as did the number of patients with panic disorder ($\chi^2[1] = 23.81$; $p < 0.001$). However, no significant decrease was observed for the number of patients with social phobia ($\chi^2[1] = 0.09$; $p = 0.77$), this was due to a large number of patients who were newly diagnosed with social phobia (n = 22) rather than a low response rate, as 25 patients initially diagnosed with social phobia were in remission. Similarly, no significant difference was found for specific phobia ($\chi^2[1] = 3.06$; $p = 0.08$), which was newly diagnosed in 30 patients and remitted only in 17.



Color version available online

Fig. 2. Effect sizes from pre- to posttherapy and posttherapy to follow-up. BSI, Brief Symptom Inventory; BDI, Beck Depression Inventory; MIA-AAC, Mobility Inventory Avoidance Accompanied; MIA-AAC, Mobility Inventory Avoidance Alone; BSQ, Body Sensations Questionnaire.

Table 2. Patients' retrospective ratings of their therapy

Helpful therapy?		Reoccurring problems?		Further improvement?	
response category	frequency, *n* (%)	response category	frequency, *n* (%)	response category	frequency, *n* (%)
More harm	4 (2)	Never again	31 (13)	Became much better	54 (23)
Not at all	17 (7)	Sometimes	54 (22)	Became markedly better	67 (28)
A little bit	27 (11)	Occasionally	65 (27)	Became a bit better	47 (20)
Some	34 (14)	Often	35 (15)	Stayed the same	25 (11)
Strong	69 (29)	Very often	34 (14)	Got worse	25 (11)
Very much	91 (38)	All the time	22 (9)	–	–

The full questions were Helpful therapy "How much did the therapy help you?"; Reocurring problems: "Did the problems that led to your admission to the treatment center reoccur?"; Further improvement: "Compared to right after your therapy it…."

The level of overall symptomatology as measured by BSI-score was 0.6 on average (SD 0.59; Min = 0; Max = 4), which is above the proposed cut point for a "dysfunctional" population [8].

Patients' Evaluation

At follow-up, patients were asked to evaluate their initial therapy, and the reocurrence of symptoms changes after therapy. About two-thirds of patients indicated that the therapy was markedly or very helpful to them, whereas 10% indicated no help or even negative effects. Judging the reoccurrence of the problems, only 14% said that the problems "never" reoccurred, while 23% indicated that the problems reoccurred "often" or "very often."

Effectiveness

Effect sizes ranged from 0.31 to 1.35 across measures and measurement occasions (Fig. 2; Table 3; online suppl. Appendix,seewww.karger.com/doi/10.1159/000500188). Notably, the effect sizes were larger at FU compared to posttherapy. Numerically larger effect sizes were found

Table 3. Descriptive statistics and effect sizes

Scale	Means (SD)			Effect sizes	
	pre	post	follow-up	pre-post	pre-follow-up
BSI (216)	1.10 (0.65)	0.69 (0.64)	0.60 (0.59)	0.63	0.80
BDI (193)	18.06 (9.72)	9.99 (9.61)	9.07 (8.66)	0.75	0.92
MIA-AAC (19)	1.78 (0.79)	1.54 (0.65)	1.42 (0.76)	0.31	0.48
MIA-AAL (19)	2.30 (0.89)	1.92 (0.89)	1.75 (0.94)	0.45	0.65
BSQ (15)	2.52 (0.86)	2.04 (0.85)	1.41 (0.73)	0.47	1.35

BSI, Brief Symptom Inventory; BDI, Beck Depression Inventory; MIA-AAC, Mobility Inventory Avoidance Accompanied; MIA-AAC, Mobility Inventory Avoidance Alone; BSQ, Body Sensations Questionnaire.

for the BDI (*ES* = 0.75) than for the BSI (*ES* = 0.63) directly after therapy.

Clinical Significance

Analysis of clinical significance showed that not all patients improved to the same extent during and after treatment (Fig. 3). Directly after treatment, for the BSI about half of the patients fell into the "no-change unproblematic" category, a quarter fell into the "no-change problematic" and "reliably deteriorated" categories, and the remaining quarter fell into the "clinically significant change" and "reliable change not clinically significant" categories. At follow-up, slightly more patients were categorized into the more positive categories ($V = 1466$; $p < 0.001$).

For the BDI, fewer patients were classified into the "no-change unproblematic" than for the BSI, and both the "clinical significant change" as well as the "no-change problematic" groups were larger for the BDI. Again, more patients were classified into the more positive categories at follow-up ($V = 1616.5$; $p < 0.02$).

Predictors for Follow-Up Status

Lastly, we used hierarchical linear regression analysis to predict patients' BDI and BSI scores at follow-up (Table 4). In the first step, only BDI-pre scores as predictors explained about 22% of the variance. Demographic variables entered in Step 2 (number of children, family status) only modestly improved fit ($\Delta R^2 = 0.05$; $F[5, 127] = 2.06$; $p = 0.07$). Variables related to therapy (additional psychotherapy after first therapy, additional pharmacotherapy, duration of initial therapy, post score) showed a marked improvement in model fit ($\Delta R^2 = 0.16$; $F[4, 123] = 8.48$; $p < 0.001$). Here patients who had higher BDI-post scores ($b = 0.39$; $p < 0.001$) had a worse outcome, but the other predictors were not significantly related to the outcome. A similar pattern of results emerged, when the analysis was performed on BSI instead of BDI scores (Table 4).

Discussion

The aim of the present study was to investigate long-term outcomes after CBT in routine care. We inspected diagnoses according to semi-structured clinical interviews as well as self-rated levels of symptomatology and treatment evaluations. Our study yielded 5 main findings. First, we found large effect sizes when comparing pre- to both post- and follow-up measurements. Second, our sample included a remarkable number of patients not attaining clinically significant levels of improvement and a very small group of patients even deteriorating. Third, slightly more patients experienced clinically significant levels of change at follow-up compared to posttreatment. Fourth, about half of the former patients, however, still fulfilled the diagnostic criteria for at least 1 psychological disorder at follow-up assessment. Fifth, regression analysis revealed that variables related to therapy explained variance above and beyond clinical and demographic data. We will discuss each of these findings in turn before discussing the limitations of the present study.

Short-Term Effectiveness

Pre-post effect sizes were comparable to other effectiveness studies in routine care. In a sample of Swedish outpatients in routine care, who showed comorbidity rates similar to those found in our study, Werbart et al. [45] report an effect size of. 0.89 on the symptom check list, which is comparable to the effect sizes we found on the BSI. The effect sizes on the BDI were 1.06 (intention-to-treat) in the meta-analysis by Hans and Hiller [17],

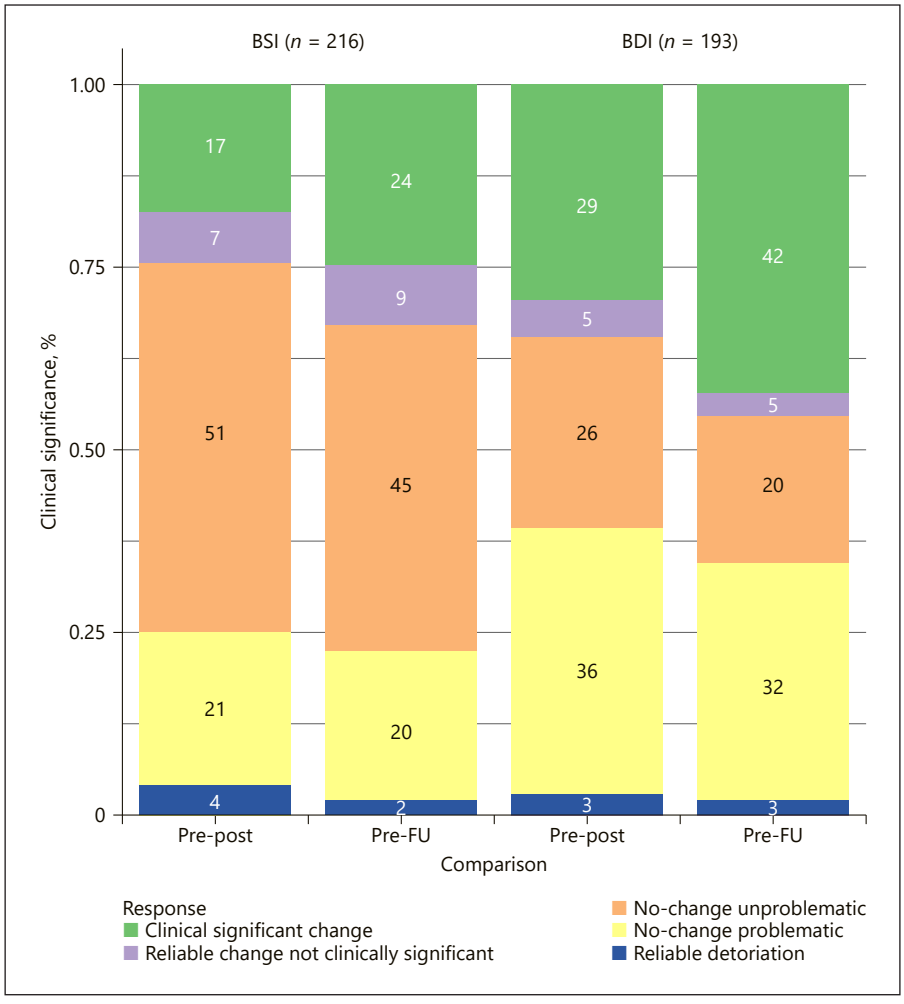


Fig. 3. Clinical Significance for BDI and BSI. BSI, Brief Symptom Inventory; BDI, Beck Depression Inventory.

which is comparable to our results. This may be due to the fact that both patients in our study and in some of the studies considered by Hans and Hiller [17] completed the BDI even when depression was not their primary diagnosis. Furthermore, we found that the overall level of distress assessed with the BSI at beginning of therapy was rather low for a clinical sample, which may have limited effect sizes.

Concerning the levels of clinically significant change, our results are lower than those in other studies from routine care that found that rates of recovered patients hover around 50% in a mix of common mental health problems [46]. However, the measures and choice of cut-off seem highly important for these results. Closer inspection of the results [46] of this review shows that the 2 studies that used the BDI and BSI as outcome measures achieved rates of clinical improvement of only 32 and 7%, respectively [46]. Also more specific studies into levels of clinically significant change for depression in routine care show rates of clinical improvement of 36 [15] and 39% [21].

Long-Term Effectiveness

Since – to our knowledge – this is the first long-term follow-up study after several years after routine care, comparisons to other studies are difficult. Our follow-up effect sizes were slightly smaller than those found 1 year after routine care. Wittmann et al. [23] reported $d = 0.97$ for the BSI and $d = 1.36$ for the BDI from pretreatment to 1-year follow-up. These effect sizes are based on a selective sample and comparable to those of our study, when corrected [23].

The follow-up effect sizes we found were smaller when compared to those from previous long-term follow-up studies based on RCTs or based on specific samples such as patients with anxiety disorders [22], which typically

Table 4. Results of the hierarchical regression analysis for BDI and BSI

Variables	BDI			BSI		
	model 1	model 2	model 3	model 1	model 2	model 3
Pretreatment score	0.44*** (0.30 to 0.58)	0.45*** (0.31–0.59)	0.21** (0.06 to 0.37)	0.38*** (0.25 to 0.51)	0.38*** (0.25 to 0.52)	0.22** (0.09 to 0.35)
Family status (reference "single")						
"Divorced"		3.85 (–1.11 to 8.81)	4.93* (0.37 to 9.49)		0.18 (–0.18 to 0.53)	0.13 (–0.19 to 0.45)
Widowed		–9.97 (–21.64 to 1.70)	–11.11* (–21.90 to –0.32)		–0.38 (–1.21 to 0.45)	–0.34 (–1.09 to 0.40)
Long-term relation		0.98 (–3.66 to 5.61)	1.31 (–2.89 to 5.51)		0.16 (–0.18 to 0.49)	0.10 (–0.20 to 0.39)
Married		–0.32 (–3.88 to 3.25)	0.29 (–3.01 to 3.60)		0.05 (–0.21 to 0.31)	0.00 (–0.23 to 0.24)
Children		0.48 (–0.93 to 1.89)	0.29 (–0.98 to 1.57)			0.03 (–0.06 to 0.12)
Therapy duration			0.46 (–0.49 to 1.42)			–0.01 (–0.07 to 0.06)
Postscore			0.39*** (0.23 to 0.55)			0.42 *** (0.28 to 0.56)
Additional psychotherapy			1.78 (–0.81 to 4.38)			0.04 (–0.14 to 0.22)
Additional pharmacotherapy			0.65 (–0.63 to 1.94)			–0.01 (–0.11 to 0.08)
Observations	134	134	134	134	134	134
R^2	0.22	0.27	0.43	0.20	0.22	0.42
F	37.34	7.77	9.15	32.53	5.98	8.75
ΔR^2		0.05	0.16		0.02	0.2
ΔF		2.07	8.48***		0.95	3.23***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; 95% CIs are given in brackets.
BDI, Beck Depression Inventory; BSI, Brief Symptom Inventory.

yield effect sizes that are higher than studies on routine care or patients with different diagnoses [1, 17]. Equal to previous long-term follow up studies, we found that effect sizes were at least equally high from pretreatment to follow-up than from pre- to posttreatment.

Improvement after Follow-Up

Our results in terms of effect size and clinical significance show that many patients report further improvements after therapy. Several studies report similar positive effects [47], also after routine care treatment [23]. As of today, it is unclear what drives these effects. It is plausible that patients might continue to apply CBT strategies such as exposure, behavioral activation, reducing safety, or avoidance behavior. Especially exposure therapy and the reduction of avoidance behavior seem to be a very promising strategy for promoting long-term success. CBT entails numerous direct instructions for the patient on how to overcome his or her problems, which patients can apply on their own after termination of treatment. Accordingly, self-reliant strategies such as homework are related to treatment success as is self-efficacy [48, 49].

At the same time, a large number of patients in our sample felt that the problem reoccurred very often. When trying to predict the status at follow-up, we found that only post scores were reliable predictors of follow-up scores. During analysis we also fitted models with fewer variables and other patients in which divorce and additional psychological treatments turned out to be significant predictors but these ceased to be significant whenever the sample size changed by only a dozen patients. Overall, we still found that about 60% of the variance in follow-up data could not be explained. To date, there are only very few studies investigating predictors and underlying mechanisms of long-term treatment success and additional therapies instead of describing the usual suspects such as demographic variables and symptom-related distress [50]. And the large heterogeneity across treatment settings makes it very hard to attribute differences to specific factors. For example, while in the United States, patients and overall quite short treatments (8 sessions on average) longer treatments were associated with better outcomes [15], no such relationship was found for Swedish patients [45] who also received much longer therapies (30 on average) or in the present study. While matching and analyzing, subsets of data may help to increase comparability between trials [16, 18], the large number of possible moderator variables is too large to make this a sustainable strategy for the future.

Nonresponse or Relapse

Since we collected data only on the 3 different measurement points, we could not clearly differentiate between patients who experienced a relapse and those who continuously suffered from their psychological disorder until the follow-up assessment. Overall, the levels of nonresponse and deterioration were a bit higher than other studies into routine care [23]. Also the factors associated with poorer outcome are comparable to other studies [25].

Since relapse and partial remission is a known problem in depressive disorders, various treatments aiming at relapse prevention have been developed and studied for depression [51]. For instance, mindfulness-based approaches [52] or rumination-focused interventions [53] are effective in reducing residual symptoms and improving long-term outcome, but a substantial number of patients nevertheless experience a relapse in depression even after receiving relapse prevention therapy. For other disorders, such as anxiety disorders, there is a shortage of treatment as augmentation or relapse prevention because relapse or residual symptoms are not considered part of the problem. The high number of newly diagnosed patients with these conditions in the present study and previous studies into anxiety disorders show that a relevant proportion of these patients do not sufficiently benefit from therapy [54] and may develop a novel anxiety disorder after remitting from another disorder or discontinuing antidepressive medication [55]. Future research should aim at identifying the underlying mechanisms and developing respective treatments for nonresponse, including trials on switching the therapeutic approach [56].

Nevertheless, compared to the effects of pharmacological treatment, which are well studied for the short-term efficacy of antidepressants in patients with depression and hover around 0.3 [57], we found overall much larger effect sizes in our sample. The superiority of psychological treatment (here routine care CBT) becomes even clearer when looking at the long-term effects. In our sample, patients even improved after treatment, while the effects of medication disappear or can worsen the patients' condition after discontinuing [55, 58].

Limitations

This study shares a number of limitations with previous studies into the effectiveness of CBT in routine care [17, 19, 20]. First, it should be noted that a university treatment center does not meet all of the criteria laid out by Shadish et al. [19], while several other studies consider university outpatient clinics as routine care [16, 18]. Second, about one-third of the former patients refused to participate. Even though we did not find any systematic differences in symptom severity between those who participated and those who did not, more studies are needed to establish generalizability of our findings. Third, the interpretation of follow-up data is complicated by the lack of a control group.

These limitations point to some general problems with long-term follow-up studies in routine care. First, it is extremely difficult to contact former patients for retrospective studies over extensive time periods. In principle countries with official registration offices such as Germany should have an advantage here, but we found that even with the help of official sources not all patients could be contacted. It may be vital to build an online panel of former patients that are regularly contacted. This would also enable a more fine-grained analysis of the time course of symptoms using survival analysis [59, 60]. Second, since more and more studies can only be realized with short-term grants, financial limitations constrain the time frame of follow-up. As most funding schemes limit project duration to 3 years, at the most 2 years remain for follow-up after data collection. The same holds for more complex posttreatment designs [56].

To our knowledge, we reported the first data on effectiveness of CBT in routine care over an extended period of several years. While the overall effects at posttreatment and follow-up were smaller compared to data from RCTs, the effects are similar to studies in routine care. It is noteworthy that these results are very heterogeneous with some patients showing further improvements and with a substantial minority of patients reporting no benefit or even harm.

Acknowledgment

We thank Helen Vollrath for proofreading the manuscript and language checking.

Statement of Ethics

All participants gave written consent to use their data both at the time of therapy and at the time of the follow-up assessment. The study was approved by the Ethics Commission of the German Psychological Association on November 21, 2014.

Disclosure Statement

The authors have no conflicts of interest to declare.

Funding Sources

The study was supported by the Alexander von Humboldt-Professorship of Jürgen Margraf.

Authors Contributions

All authors, especially J.M., contributed to the planning and designing of the study. All authors contributed to the writing of the manuscript. All authors have approved the final version of the manuscript. R.B. supervised the data collection and – together with A.B. – supervised the post graduate students who conducted the structured interviews. G.H. and R.B. conducted the statistical analysis. T.T., J.C., and J.V. provided additional supervision in case a patient reported suicidal thoughts or plans. U.W. and D.S. organized the implementation of the patients' data warehouse over 20 years ago.

References

1 Chambless DL, Ollendick TH. Empirically supported psychological interventions: controversies and evidence. Annu Rev Psychol. 2001;52(1):685–716.
2 Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognit Ther Res. 2012 Oct;36(5):427–40.
3 Lambert MJ. Bergin and Garfield's handbook of psychotherapy and behavior change. Sons New York (NY): John Wiley; 2013.
4 Cuijpers P, Smit F, Bohlmeijer E, Hollon SD, Andersson G. Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. Br J Psychiatry. 2010 Mar;196(3):173–8.
5 Steinert C, Hofmann M, Kruse J, Leichsenring F. The prospective long-term course of adult depression in general practice and the community. A systematic literature review. J Affect Disord. 2014 Jan;152–154:65–75.
6 Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev. 2006 Jan;26(1):17–31.
7 Carter FA, Jordan J, McIntosh VV, Luty SE, McKenzie JM, Frampton CM, et al. The long-term efficacy of three psychotherapies for anorexia nervosa: a randomized, controlled trial. Int J Eat Disord. 2011 Nov;44(7):647–54.
8 Dugas MJ, Ladouceur R, Léger E, Freeston MH, Langlois F, Provencher MD, et al. Group cognitive-behavioral therapy for generalized anxiety disorder: treatment outcome and long-term follow-up. J Consult Clin Psychol. 2003 Aug;71(4):821–5.
9 Durham RC, Chambers JA, MacDonald RR, Power KG, Major K. Does cognitive-behavioural therapy influence the long-term outcome of generalized anxiety disorder? An 8-14 year follow-up of two clinical trials. Psychol Med. 2003 Apr;33(3):499–509.
10 Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015 Nov;16(1):495.
11 Leichsenring F. Randomized controlled versus naturalistic studies: a new research agenda. Bull Menninger Clin. 2004;68(2):137–51.
12 Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005 Jan;365(9453):82–93.
13 Stewart RE, Chambless DL. Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. J Consult Clin Psychol. 2009 Aug;77(4):595–606.
14 Merrill KA, Tolbert VE, Wade WA. Effectiveness of cognitive therapy for depression in a community mental health center: a benchmarking study. J Consult Clin Psychol. 2003 Apr;71(2):404–9.
15 Gibbons CJ, Fournier JC, Stirman SW, DeRubeis RJ, Crits-Christoph P, Beck AT. The clinical effectiveness of cognitive therapy for depression in an outpatient clinic. J Affect Disord. 2010 Sep;125(1–3):169–76.
16 Schindler AC, Hiller W, Witthöft M. Benchmarking of cognitive-behavioral therapy for depression in efficacy and effectiveness studies–how do exclusion criteria affect treatment outcome? Psychother Res. 2011 Nov;21(6):644–57.
17 Hans E, Hiller W. Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: a meta-analysis of nonrandomized effectiveness studies. J Consult Clin Psychol. 2013 Feb;81(1):75–88.
18 Lutz W, Schiefele AK, Wucherpfennig F, Rubel J, Stulz N. Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. J Affect Disord. 2016 Jan;189:150–8.
19 Shadish WR, Matt GE, Navarro AM, Phillips G. The effects of psychological therapies under clinically representative conditions: a meta-analysis. Psychol Bull. 2000 Jul;126(4):512–29.
20 Persons JB, Bostrom A, Bertagnolli A. Results of randomized controlled trials of cognitive therapy for depression generalize to private practice. Cognit Ther Res. 1999;23(5):535–48.
21 Westbrook D, Kirk J. The clinical effectiveness of cognitive behaviour therapy: outcome for a large sample of adults treated in routine practice. Behav Res Ther. 2005 Oct;43(10):1243–61.
22 DiMauro J, Domingues J, Fernandez G, Tolin DF. Long-term effectiveness of CBT for anxiety disorders in an adult outpatient clinic sample: a follow-up study. Behav Res Ther. 2013 Feb;51(2):82–6.
23 Wittmann WW, Lutz W, Steffanowski A, Kriz D, Glahn EM, Völkle MC, et al. Qualitätsmonitoring in der ambulanten Psychotherapie: Modellprojekt der Techniker Krankenkasse–Abschlussbericht. Hamburg: Techniker Krankenkasse; 2011.
24 Munsch S, Meyer AH, Biedert E. Efficacy and predictors of long-term treatment success for cognitive-behavioral treatment and behavioral weight-loss-treatment in overweight individuals with binge eating disorder. Behav Res Ther. 2012 Dec;50(12):775–85.
25 Benjamin CL, Harrison JP, Settipani CA, Brodman DM, Kendall PC. Anxiety and related outcomes in young adults 7 to 19 years after receiving treatment for child anxiety. J Consult Clin Psychol. 2013 Oct;81(5):865–76.
26 Landheim AS, Bakken K, Vaglum P. Impact of comorbid psychiatric disorders on the outcome of substance abusers: a six year prospective follow-up in two Norwegian counties. BMC Psychiatry. 2006 Oct;6(1):44.
27 Finney JW, Moos RH. The long-term course of treated alcoholism: II. Predictors and correlates of 10-year functioning and mortality. J Stud Alcohol. 1992 Mar;53(2):142–53.
28 Margraf J, Cwik J. Mini-DIPS open access: Diagnostisches Kurzinterview bei psychischen Störungen. Bochum: Forschungs-und Behandlungszentrum für psychische Gesundheit. Ruhr-Universität Bochum; 2017.
29 Margraf J, Cwik JC, Suppiger A, Schneider S. DIPS Open Access: Diagnostisches Interview bei psychischen Störungen. Bochum: Ruhr-Universität Bochum, Forschungs-und Behandlungszentrum für psychische Gesundheit. Available at: http://dips-interviews.rub.de 2017.

30 Wittchen H-U, Zaudig M, Fydrich T. SKID Strukturiertes Klinisches Interview für DSM-IV. Göttingen: Hogrefe; 1997.
31 In-Albon T, Suppiger A, Schlup B, Wendler S, Margraf J, Schneider S. Validität des Diagnostischen Interviews bei psychischen Störungen (DIPS für DSM-IV-TR). Z Klin Psychol Psychother. 2008;37(1):33–42.
32 Derogatis LR, Melisaratos N. The brief symptom inventory: an introductory report. Psychol Med. 1983 Aug;13(3):595–605.
33 Franke GH. Erste Studien zur Güte des Brief Symptom Inventory (BSI). Z Med Psychol. 1997;6:159–66.
34 Beck AT, Ward C, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry. 1961;4:561–71.
35 Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the validity of the Beck Depression Inventory. A review. Psychopathology. 1998;31(3):160–8.
36 Byerly FC, Carlson WA. Comparison among inpatients, outpatients, and normals on three self-report depression inventories. J Clin Psychol. 1982 Oct;38(4):797–804.
37 Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. Clin Psychol Rev. 1988;8(1):77–100.
38 Carmody DP. Psychometric characteristics of the Beck Depression Inventory-II with college students of diverse ethnicity. Int J Psychiatry Clin Pract. 2005;9(1):22–8.
39 Krefetz DG, Steer RA, Gulab NA, Beck AT. Convergent validity of the Beck Depression Inventory-II with the Reynolds adolescent depression scale in psychiatric inpatients. J Pers Assess. 2002 Jun;78(3):451–60.
40 Sprinkle SD, Lurie D, Insko SL, Atkinson G, Jones GL, Logan AR, et al. Criterion validity, severity cut scores, and test-retest reliability of the Beck Depression Inventory-II in a university counseling center sample. J Couns Psychol. 2002;49(3):381–5.
41 Michalak J, Kosfelder J, Meyer F, Schulte D. Messung des Therapieerfolgs. Z Klin Psychol Psychother. 2003;32(2):94–103.
42 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991 Feb;59(1):12–9.
43 Hiller W, Schindler A. Response und Remission in der Psychotherapieforschung. Psychother Psych Med. 2011;61(3/04):170–176.
44 Seggar LB, Lambert MJ, Hansen NB. Assessing clinical significance: application to the Beck Depression Inventory. Behav Ther. 2002;33(2):253–69.
45 Werbart A, Levin L, Andersson H, Sandell R. Everyday evidence: outcomes of psychotherapies in Swedish public health services. Psychotherapy (Chic). 2013 Mar;50(1):119–30.
46 Cahill J, Barkham M, Stiles WB. Systematic review of practice-based research on psychological therapies in routine clinic settings. Br J Clin Psychol. 2010 Nov;49(Pt 4):421–53.
47 Ruhmland M, Margraf J. Effektivität psychologischer Therapien von Generalisierter Angststörung und Sozialer Phobie: Meta-analysen auf Störungsebene. Verhaltenstherapie. 2001;11(1):27–40.
48 Mausbach BT, Moore R, Roesch S, Cardenas V, Patterson TL. The relationship between homework compliance and therapy outcomes: an updated meta-analysis. Cognit Ther Res. 2010 Oct;34(5):429–38.
49 Wilson GT, Fairburn CC, Agras WS, Walsh BT, Kraemer H. Cognitive-behavioral therapy for bulimia nervosa: time course and mechanisms of change. J Consult Clin Psychol. 2002 Apr;70(2):267–74.
50 Månsson KN, Frick A, Boraxbekk CJ, Marquand AF, Williams SC, Carlbring P, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl Psychiatry. 2015 Mar;5(3):e530.
51 Bockting CL, Hollon SD, Jarrett RB, Kuyken W, Dobson K. A lifetime approach to major depressive disorder: the contributions of psychological interventions in preventing relapse and recurrence. Clin Psychol Rev. 2015 Nov;41:16–26.
52 Kuyken W, Byford S, Taylor RS, Watkins E, Holden E, White K, et al. Mindfulness-based cognitive therapy to prevent relapse in recurrent depression. J Consult Clin Psychol. 2008 Dec;76(6):966–78.
53 Teismann T, von Brachel R, Hanning S, Grillenberger M, Hebermehl L, Hornstein I, et al. A randomized controlled trial on the effectiveness of a rumination-focused group treatment for residual depression. Psychother Res. 2014;24(1):80–90.
54 Bruce SE, Yonkers KA, Otto MW, Eisen JL, Weisberg RB, Pagano M, et al. Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. Am J Psychiatry. 2005 Jun;162(6):1179–87.
55 Fava GA, Belaise C. Discontinuing antidepressant drugs: Lesson from a failed trial and extensive clinical experience. Psychother Psychosom. 2018;87(5):257–67.
56 Steinert C, Kruse J, Leichsenring F. Long-term outcome and non-response in psychotherapy: are we short-sighted. Psychother Psychosom. 2016;85(4):235–7.
57 Gibertini M, Nations KR, Whitaker JA. Obtained effect size as a function of sample size in approved antidepressants: a real-world illustration in support of better trial design. Int Clin Psychopharmacol. 2012 Mar;27(2):100–6.
58 Margraf J, Schneider S. From neuroleptics to neuroscience and from Pavlov to psychotherapy: more than just the "emperor's new treatments" for mental illnesses? EMBO Mol Med. 2016 Oct;8(10):1115–7.
59 Fava GA, Rafanelli C, Grandi S, Conti S, Ruini C, Mangelli L, et al. Long-term outcome of panic disorder with agoraphobia treated by exposure. Psychol Med. 2001 Jul;31(5):891–8.
60 Fava GA, Grandi S, Rafanelli C, Ruini C, Conti S, Belluardo P. Long-term outcome of social phobia treated by exposure. Psychol Med. 2001 Jul;31(5):899–905.