Are depression trials generalizable to clinical practice? Something clinicians always wanted to know about RCTs, but were afraid to ask.....

Lem, R. van der

Chapter 2

Efficacy versus effectiveness: A direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice

Rosalind van der Lem
Nic J.A. van der Wee
Tineke van Veen
Frans G. Zitman

Psychotherapy and Psychosomatics 2012: 81:226-234

ABSTRACT

Background: results from RCTs are considered to give the most reliable information on treatment outcome (efficacy). Yet, the generalizability of efficacy results to daily practice (effectiveness) might be diminished by the design of RCTs. The STAR*D trial approached daily practice as much as possible, but still has some properties of an RCT. In this study, we compare results from treatment of major depressive disorder (MDD) in routine clinical practice to those of RCTs and STAR*D.

Methods: Effectiveness in routine clinical practice was compared with efficacy results from 15 meta-analyses on antidepressant, psychotherapeutic and combination treatment and results from STAR*D. Data on daily practice patients and treatments was derived from a Routine Outcome Monitoring (ROM) system. Treatment outcome was defined as proportion of remitters (MADRS≤10) and within group effect size.

Results: From ROM, 598 patients suffering from a MDD according to the MINIplus were included. Remission percentages were lower in routine practice than in meta-analyses for all treatment modalities (32% vs.40–74%). Differences were less explicit for antidepressants (21% vs. 34-47%) than for individual psychotherapy (27% vs. 34–58%; effect size of 0.85 vs. 1.71) and combination therapy (21% vs. 45–63%), since only 60% of the meta-analyses for antidepressants showed significant differences with ROM, while for psychotherapy and combination treatment almost all meta-analyses showed significant differences. No differences in effectiveness were found between routine practice and STAR*D

Conclusions: effectiveness of treatment for mild to moderate MDD in daily practice is similar to STAR*D and significantly lower than efficacy results from RCTs.

INTRODUCTION

During the past decades, the selection of treatments for major depressive disorder (MDD) has shifted from an approach based on clinical expertise towards evidence based medicine. Evidence based treatment guidelines are based on the results of randomized controlled trials (RCTs) [1]. Adherence to these treatment guidelines is expected to improve effectiveness of treatment in daily practice [2]. However, the effects in clinical practice may not be comparable to those found in RCTs. RCTs are designed to maximise the internal validity of the investigated trial i.e. their aim is to look for effects that are replicable and solely attributable to the investigated treatment. Therefore, RCTs usually include patients following stringent selection criteria. Unlike clinical practice, in RCTs patients with co morbidity or risk of suicide are excluded, and a minimum symptom severity is required for inclusion. Also, much more effort is put in maximizing treatment adherence than is usual in clinical practice. Furthermore, RCTs are frequently carried out in specialised settings [3-5]. While clearly increasing the internal validity, these features limit the external validity, i.e. the generalizability, of RCTs [6,7]. Hence, it is important for clinicians to know what may be expected from evidence based treatments of MDD in routine clinical practice. Unfortunately, publications on effectiveness are very scarce. The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was one of the first studies designed to resemble the routine clinical practice of the treatment of MDD. However, in STAR*D prior non-responders to the study drugs, patients with a preference for non-pharmacological treatment before the first treatment step, and patients with a baseline severity of less than 14 on the 17-item Hamilton Rating Scale for Depression (HAMD) were excluded [8]. Furthermore, in STAR*D much effort has been put in maximizing adherence to the treatment protocol, both for participating patients and therapists. Therefore, the STAR*D-design has elements of both RCTs and routine clinical practice. Success rates in STAR*D were modest, and in daily practice treatment results for might be even worse, as many of the STAR*D strategies to enhance treatment adherence may not be feasible. However, some other factors in routine mental health care might contribute to a better treatment outcome, such as the possibility to allow for patient preferences [9]. To our knowledge, no studies on the effects of evidence based depression treatments in routine clinical practice have been published yet.

Since 2002, the Dutch Regional Mental Health Provider (RMHP) Rivierduinen assesses psychopathology and other characteristics with structured interviews and rating scales as a part of routine clinical practice. The assessments are done during the first visit and subsequently on every three to four months to monitor progress. This routine outcome monitoring (ROM) is integrated with stepped care protocols, based on evidence based treatment guidelines [10].

The ROM data allow a comparison of the effects of treatment for MDD in RCTs, in STAR*D, and in routine daily outpatient practice. In the present study, we compared the effects of

treatments with antidepressants, with psychotherapy and with combination therapy between the three different settings. As in routine clinical practice a limited set of inclusion and exclusion criteria is used and no extra effort is put to enhance treatment adherence, we expected to find the smallest effect sizes in routine clinical practice, with effect sizes in STAR*D in-between those of routine clinical practice and RCTs.

METHODS

RCTs: selection of meta-analyses

As there are many RCTs of treatments for MDD, we used the results of meta-analyses. As the topic of this study is the outcome of treatments in different settings and not the effects relative to a control group, we selected meta-analyses providing within-group results.

First, we searched PubMED and PsycInfo for meta-analyses of depression treatment in adult psychiatric outpatients. We also screened the reference lists of selected meta-analyses for other meta-analyses, searched an extensive database on psychotherapy-RCTs (http://www.psychotherapyrcts.org) and contacted two experts in psychotherapy of depression for other references of meta-analyses of psychotherapy. We finally identified 431 meta-analyses.

Inclusion criteria for meta-analyses in the present study were: 1) provide data on within-group efficacy on depression severity, total number of patients per treatment arm and number of responders or remitters or within-group effect size. 2) select secondary care outpatients with unipolar, non-psychotic MDD without co morbidity. 3) outcome defined with the Hamilton Depression Rating Scale (HAMD)[8], the Montgomery Asberg Depression Rating Scale (MADRS) [11] or the Beck Depression Inventory (BDI-II) [12]. Exclusion criteria were: 1) rate of relapse, drop-out rate or response rapidity as only efficacy measure, or focus on specific symptoms like physical complaints. 2) focus not the effect of treatment, but another like the placebo-effect. 3) focus on antidepressant drugs unavailable in the Netherlands or very infrequently used in RMHP Rivierduinen during the investigated period (2002-2006): duloxetine, escitalopram, desvenlafaxine, reboxetine, moclobemide, milnacipran, trazodone, and nefazodone.

After application of these in- and exclusion criteria, 17 of the 431 meta-analyses were included. Most meta-analyses were excluded because they did not provide within-group data. Another two meta-analyses were excluded because they overlapped with other meta-analyses [13,14]. In six of the finally 15 selected meta-analyses, outcome was defined as the proportion of responders [15-20]; in four as proportion of remitters [21-24] and in four as both [25-28]. Only one meta-analysis [29] (on psychotherapy) defined outcome as effect size on the BDI [30]. Pre-and post-treatment data within each individual trial were aggregated to estimate the individual effect size of that trial. In addition, the individual effect sizes of the trials were aggregated to estimate an overall effect size [30]. Although effect size is generally

used as an estimate of between-group efficacy, in the meta-analysis of Minami et al. on psychotherapy a within-group effect size was generated as a benchmark for future research and clinical practice.

Response was defined as a 50% reduction of severity on a depression scale. Four meta-analyses used the 17-items HAMD [20,25-27] to assess response, one meta-analysis used the 21-items HAMD [28] and five used either one or both versions of the HAMD (17 or 21 items) and/or the MADRS [15-19]. With respect to remission, cut-off scores on the post-treatment assessment of a depression severity scale were used. In two meta-analyses the cut-off was a score of ≤7 on the 17-item HAMD [26,27], in one meta-analysis the cut-off was ≤7 on the first 17 items of the 21-item version HAMD [28], another used <7 as cut-off on the 17-item HAMD [23]and yet another a score of ≤ 8 on the 17-item HAMD [25]. Two meta-analyses included trials with different definitions of remission on different scales (17 item HAMD, 21 item HAMD, MADRS, and BDI) [21,22].

STAR*D

Two publications from the STAR*D trial provided within-group results of antidepressant therapy, cognitive behavioral therapy and combination therapy [31,32]. Only remission was assessed and the cut-off was defined as a score of ≤7 on the 17-item HAMD.

Routine Clinical practice: The Dutch mental health care system and treatment steps for MDD

In the Netherlands, health insurance is obligatory for all citizens, and mental health care is not (yet) restricted by the financial means of patients. The Dutch mental health care system is organized in a stepped-care-manner and uses evidence based treatment guidelines. Patients with mood complaints visit their general practitioner (GP) first. The treatment guidelines recommend that patients with mild to moderate depression be treated with psychotherapy or pharmacotherapy, based on the patient's preferences [33]. Patients with severe depression should preferably start a pharmacotherapy. Rating of severity is based on clinical judgment. Reasons to refer patients to a RMHP are a preference for psychotherapy, more severe, recurrent or refractory depression or the presence of co morbid psychiatric or somatic disorders. After baseline assessment and a clinical interview at our RMHP, patients were offered treatment steps as recommended by the guidelines. Patients suffering from moderate to severe MDD could choose between psychotherapy and antidepressants. For severe depression antidepressants were the first choice. When patients are already on antidepressants the dose is optimized or patients are offered to switch to another antidepressant or start psychotherapy.

Routine Outcome Monitoring

ROM at the RMHP Rivierduinen is described in detail elsewhere [10]. The assessments are carried out by specially trained research nurses with the help of dedicated software. The outcomes are fed back to the therapist, and discussed with the patient. The baseline assessment comprises a standardized diagnostic interview (Mini-International Neuropsychiatric Interview Plus MINIplus [34]), collection of sociodemographic and socioeconomic data, administration of observer rated scales (including the MADRS) and self report questionnaires (including the BDI-II), and general measures of health and quality of life. Only patients with insufficient mastery of the Dutch language are not eligible for ROM.

In this study anonymized ROM data were used, in agreement with the Psychiatric Academic Registration Leiden (PAREL), which has been approved by the Medical Ethical Committee of the University Medical Hospital Leiden.

From 2002 through 2006, 1653 of the patients with ROM suffered from a MDD according to the MINI plus at intake. Of these, 879 had only one ROM assessment: 190 did not start treatment after the baseline assessment, 350 remained in treatment but had no follow-up assessments, and for 339 no information on treatment continuation after baseline assessment was available. Of 774 patients two or more assessments were available. Of these patients, 169 were excluded for the following reasons: time between baseline and follow-up too short (<4 weeks) or too long (>52 weeks) for a single treatment (n=47), inpatient treatment in the period before the second assessment (n=43), psychotic features or bipolar disorder (n=42), no treatment information available (n=28), a MADRS ≤10 at baseline (n=15), and over 65 years of age (n=1). There remained 598 patients for further analysis. Of them, 82 were treated with antidepressants only, 170 with individual psychotherapy only, 167 with the combination of both, 90 with antidepressants and supportive therapy and 89 with other treatments.

Statistical analysis

For comparison of the ROM results with the RCT and STAR*D data, the response and remission rate, and the within group effect size Δ were computed. For ROM, we defined response as a 50% symptom reduction on the MADRS. For remission, the most commonly used threshold in RCTs is a score of 7 or less on the HAMD17. Several methods to compute an equivalent MADRS score has been described, either by equations [35-37] (a score of 7 on the HAMD17 corresponds to a MADRS score between 8–10) or with the Item Response Theory ([38] (a score of 7 of the HAMD17 corresponds with a MADRS score of 8–9). In other research several definitions of remission on the MADRS were described: a threshold of 10, 11 or 12, corresponding with "borderline mentally ill" or no symptoms of illness on the CGI-S (clinical global impression-severity of illness) [39,40]. We defined research remission as MADRS score of ≤10 [39]. Recently, a cut-off of ≤5 on the MADRS has been suggested as more appropriate to define remission [40]. Therefore, we also computed proportions of remission defined as

MADRS ≤5. However, these proportions were not used in the comparison with the meta-analyses, since RCTs used a higher cut-off to define remission. Proportions of response and remission were computed for antidepressant treatment, individual psychotherapy, the combination of both and the combination of antidepressants and supportive therapy. Within-group effectsize for individual psychotherapy was defined as $\Delta = (\mu pre - \mu post) / \sigma$ in which μpre = mean pre-treatment, μpost =mean post-treatment and σ = standard deviation pre-treatment on the BDI-II [30].

The results of the meta-analyses, of STAR*D and of ROM were compared using two independent proportions in the following statistical formulas [41,42]:1) 95% Confidence interval: $se(p_1-p_2) = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$, $p_1-p_2 - 1{,}96 \times se(p_1-p_2)$ to $p_1-p_2 + 1{,}96 \times se(p_1-p_2)$, in which se= standard error, p1= proportion of responders/remitters in meta-analyses, p2=proportion of responders/remitters in ROM population, n1= number of patients in meta-analyses within the treatment modality, n2= number of patients in ROM population within the treatment modality. and 2) Hypothesis test (with continuity correction): $P = (r_1 + r_2) / (n_1 + n_2)$, $se(p_1-p_2) = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$, $z_c = |p_1-p_2| - 0{,}5\,(1/n_1 + 1/n_2) / se(p_1-p_2)$, in which P= probability given H_0 is true (no difference between p1 and p2), r_1= number of responders/remitters in meta-analyses, r_2= number of responders/remitters in ROM sample, z_c = z-score in normal distribution-two tailed areas (z -> p-value).

RESULTS

ROM patients

Characteristics of the included ROM patients. We included 598 patients with a current MDD with at least one follow-up ROM assessment. Table 1 shows the sociodemographic features of this group. Table 2 shows clinical characteristics at baseline assessment (severity of the MDD, co morbid Axis I disorders, primary clinical diagnosis) and the timeframe of assessments.

Table 1. Sociodemographic and baseline characteristics of treatment groups.

	Antidepressants (n=82)	Individual Psychotherapy (n=170)	Antidepressants + Ind. Psychotherapy (n=167)	Antidepressants + social supportive therapy (n=90)	Other (n=89)
Age * (mean, 95% CI)	42.2 (39.6-44.8)	36.0 (34.1-37.8)	38.7 (36.9-40.5)	42.1 (39.4-45.0)	39.6 (37.0-42.2)
Gender (% male)	56.1%	30.6%	24.6%	32.2%	34.8%
Marital status					
Percentage Married/cohabitating	54.5%	51.0%	54.6%	47.5%	54.2%
Educational level					
None/primary	11.7%	9.3%	9.9%	17.5%	16.7%
Intermediate low	36.4%	29.1%	31.9%	36.2%	34.7%
Intermediate high	37.7%	37.7%	44.0%	33.8%	36.1%
Academic/high	14.3%	23.8%	14.2%	12.5%	12.5%
Employment					
Percentage Currently Employed	41.6%	37.7%	34.8%	22.5%	33.3%
Baseline severity					
MADRS (mean)	26.9	23.2	26.9	27.9	24.3
(95% CI; ≈**HDRS-17**)	(25.2-28.5; **18-22**)	(22.2-24.3; **16-19**)	(25.9-27.9; **18-22**)	(26.4-29.3; **19-23**)	(22.8-25.8; **16-20**)
BDI-II (mean)	30.8	28.0	32.8	32.7	30.9
(95% CI)	(28.4-33.3)	(26.3-29.6)	(31.3-34.3)	(30.5-34.9)	(28.5-33.3)
Co morbidity					
Percentage Anxiety and/or Somatoform disorders	26.4%	47.1%	46.1%	44.4%	43.8%
Percentage Other Axis I disorders	9.8%	12.9%	11.4%	11.1%	14.6%
Percentage Alcohol/Drugs Abuse	8.5%	7.6%	6.6%	11.1%	12.4%
Using Antidepressants prior to referral	46%	12%	57%	50%	20%
Time between baseline assessment and treatment start in weeks (mean + 95% CI)	0.7 (-0.4-1.8)	3.5 (2.1-4.9)	0.53 (-0.5-1.6)	-0.3 (-1.5-0.9)	1.7 (0.1-3.3)
Time between start treatment and follow up assessment in weeks (mean + 95% CI)	20.8 (18.7-22.9)	20.1 (18.5-21.6)	21.5 (20.0-23.1)	21.6 (19.9-23.3)	19.1 (17.1-21.1)

*Sums do not always equal N due to missing values. Percentages are based on available data. CI = confidence interval.

Comparison with lost-to-follow-up group. To assess a possible selection bias, we compared the 598 included patients to the 879 patients who were not included. We classified the 879 patients who were lost to follow-up in three categories: patients who dropped out of treatment after baseline assessment (n=190), patients who remained in treatment but had no follow-up assessments (n=350), and patients on whom no information on treatment continuation after baseline assessment was available (n=339). We compared these groups and the included sample on the following variables: age, gender, ethnicity, marital status, employment status, educational level, baseline severity of the depression (MADRS, BDI-II), recurrence of the depressive disorder, co morbid Axis I disorders and suspected personality pathology as assesses with DAPP-SF. On the majority of these variables, no differences were found. The included patients differed from those who dropped out of treatment immediately after baseline assessment with respect to co morbidity and socioeconomic status. The latter were younger (mean age 36 vs. 39, difference 3 years, 95% CI 1.0–4.7, p=0.002), more often single (46% vs. 31%, p=0.003) and there was a trend towards a lower educational level (56% vs. 44 % had less than secondary school diploma, p=0.05). They also had a higher score on the BDI-II (33.2 vs. 30.6, difference 2.6, 95% CI -4.2– -0.6, p=0.003) and higher scores on the DAPP-SF (52% vs. 38%, p=0.006). Finally, they suffered more often from posttraumatic stress disorder (21% vs. 15%, p=0.05) and alcohol/drug abuse (16% vs. 9%, p=0.003).

Effectiveness in the ROM sample. Response percentages for the different modalities in ROM varied between 29% and 32%, remission percentages between 17% and 27%. Response percentages in the ROM sample were very close to the remission percentages due to the low baseline severity. The mean baseline severity of the different treatment modalities varied between 23.3 and 27.9 on the MADRS, which means that a response (50% reduction of symptoms) had to be a MADRS score ≤11.7-14.0; while remission was defined as a MADRS ≤10. When remission was defined as MADRS ≤5, remission percentages in the ROM sample were between 7% and 10%. The within group effect sizes of the treatment modalities varied between 0.68 and 0.97.

Comparison of outcomes

We compared proportions of remitters in the ROM sample and in meta-analyses or STAR*D. We also compared proportions of responders, but these results were similar, as most patients in ROM suffered from mild to moderate depression (data not shown).

Comparison of the remission percentages in ROM and meta-analyses showed that effectiveness (27%) was lower than efficacy (34–47%). Differences in remission between meta-analyses and daily practice were less explicit for pharmacotherapy than for individual psychotherapy or combination therapy, since for only three of the five meta-analyses on antidepressants the differences were significant (table 2). However, for individual psychotherapy and combination therapy, daily practice did significantly worse than RCTs: two out of three meta-analyses of individual psychotherapy showed significantly better

results on remission (34–58% versus 27%) and all three meta-analyses of combination therapy also showed significantly better results for combination therapy than in daily practice (45–63% versus 21%). We found no differences between the proportion of remitters in routine clinical practice and the proportions of remitters on the different treatment steps in STAR*D.

Within-group effect sizes in ROM could be compared with the results of one meta-analysis on psychotherapy and this showed that the outcome of individual psychotherapy was significantly less favorable in ROM (effect size 0.85) than in RCTs (effect size 1.71, CI 1.60–1.82). This difference was statistically significant, since the ROM effect size is smaller than the 95% confidence interval in the meta-analysis ($p<0.05$).

Table 2. Comparison of remission percentages of meta-analyses (efficacy) and daily practice (effectiveness).

Reference Number	Investigated treatment in meta-analysis	Number of studies (and total of patients in treatment arm) included in meta-analysis	Mean baseline severity in meta-analysis	*Efficacy:* Proportion remission in meta-analysis	*Effectiveness:* Proportion remission in our daily practice data defined as MADRS≤10 (and defined as MADRS≤5)	Difference in remission efficacy-effectiveness (+ 95% CI interval)[3]	P-value
25	ssri	39 (1769)	23.9[1]	38%	27% (9%)	11% (1-21%)	0.06
25	tca	39 (1680)		39%	27% (9%)	12% (2-22%)	**0.04**
21	ssri	8 (748)	26[1] 30.7[2]	35%	27% (9%)	8% (-2-18%)	0.18
21	snri	8 (851)		45%	27% (9%)	18% (8-28%)	**0.003**
26	snri	5 (199)	25.4[1]	39%	27% (9%)	12% (0-24%)	0.08
26	tca	5 (206)		38%	27% (9%)	11% (-1-23%)	0.10
27	ssri	3 (245)	22.0[1]	32%	27% (9%)	5% (-6-16%)	0.48
28	ssri	7 (731)	22.3[1]	47%	27% (9%)	20% (10-30%)	**<0.001**
22	Individual psychotherapy	7 (459)	21.4[1]	34%	27% (10%)	7% (-1-15%)	0.11
22	Combination	7 (444)		45%	21% (7%)	25% (17-33%)	**<0.001**

Reference Number	Investigated treatment in meta-analysis	Number of studies (and total of patients in treatment arm) included in meta-analysis	Mean baseline severity in meta-analysis	*Efficacy:* Proportion remission in meta-analysis	*Effectiveness:* Proportion remission in our daily practice data defined as MADRS≤10 (and defined as MADRS≤5)	Difference in remission efficacy-effectiveness (+ 95% CI interval)[3]	P-value
23	Individual Psychotherapy	6 (595)	n.a.	37%	27% (10%)	10% (2-18%)	**0.02**
23	Combination	6 (595)		48%	21% (7%)	27% (19-35%)	**<0.001**
24	Individual Psychotherapy	7 (288)	n.a.	58%	27% (10%)	31% (22-40%)	**<0.001**
24	Combination	7 (72)		63%	21% (7%)	42% (28-56%)	**<0.001**
31 **STAR*D**	ssri	- (2867)	21.8[1]	28%	27% (9%)	1% (-8-10%)	0.91
32 **STAR*D**	switch to individual psychotherapy	-(36)	-	25%	27% (10%)	-2% (-18-14%)	0.96
32 **STAR*D**	augmentation of ssri with psychotherapy	-(65)	-	23%	21% (7%)	2% (-10-14%)	0.94

n.a.: not available in publication

[1] HDRS17

[2] MADRS

[3] comparison efficacy-effectiveness based on MADRS≤ 10 in daily practice sample

DISCUSSION

We compared the outcome of evidence based treatments for MDD in a Dutch daily practice sample with the outcomes reported in meta-analyses of RCTs and in the STAR*D trial. As expected, effectiveness results were less favorable than efficacy results reported in RCTs for antidepressants, individual psychotherapy and combination treatment, and more comparable to those of the STAR*D trial [31,32]. The differences were smaller for pharmacotherapy than for individual psychotherapy and combination treatment.

Our findings support a frequently heard criticism of clinicians who claim that achieving success with "real world" patients is more difficult than RCT-results suggest.

Several explanations for the observed differences may be considered. First, there may be differences in patient characteristics, as RCTs use stringent exclusion criteria, for obvious ethical and methodological reasons, which may jeopardize the generalizability of the results [43] [3-6]. The STAR*D group found that patients who were eligible for RCTs had better treatment outcome than non-eligible patients [44]. Contrary to this finding, in a previous study on our routine practice sample, the influence of eligibility for RCT on outcome was very small [45]. However, we did confirm that milder depression very frequently occurs in routine practice, and that exclusion of these patients from the analysis led to a more favorable treatment outcome. In RCTs, patients with less severe depression are usually excluded [46].

Participants of RCTs probably also differ from daily-practice patients on other features potentially related to positive treatment outcome. Improvement of treatment outcome due to participation in a research setting, the Hawthorne effect, is well known [47]. Further, participants of RCTs may be highly motivated and hence have good adherence to treatment [2] [48]. Participants in RCTs typically accept randomization to different therapies, whereas in daily practice many patients specifically ask for a certain type of therapy. There might be an intrinsic difference between these groups of patients. The fact that in some trials patients are rewarded for participation might also influence outcome. Furthermore, the treatment provided in trials might be of higher quality than in daily practice as in trials special efforts are made to increase adherence and improve quality of treatment.

It remains unclear why differences between effectiveness and efficacy are more profound for individual psychotherapy and combination therapy than for pharmacotherapy. The side-effects of antidepressants that resemble symptoms of depression may contribute to a lower proportion of remission in antidepressant efficacy trials. Furthermore, there may be relevant differences between participants in antidepressant trials and in psychotherapy/combination therapy trials. Finally, there are some methodological differences between antidepressant trials and psychotherapy trials, for instance in the definition of the placebo treatment.

Besides the finding that treatment outcome for depression in RCTs is better than in daily practice, there are two other remarkable findings in our study. First, contrary to our

hypothesis, we found no differences between outcome in our ROM population and the results of the STAR*D trial. This is somewhat contradictory to previous reports that stated that STAR*D exaggerated the effectiveness of antidepressant treatment [9]. Second, there was a discrepancy between the low baseline severity on the MADRS and the relatively high score on the BDI in our population. One of the explanations is that in our population many patients might suffer from so-called "stress-related" depressions, rather than so called "somatic/biological" depression. In previous research, a discrepancy was found between observer rated (MADRS) and self reporting (BDI) scales in stress-related depression [49].

We consider the generalizability to "real life" psychiatric outpatient populations and the large number of well-documented, routinely monitored patients to be major strengths of our study. To our knowledge, no previous research has reported on treatment outcome for MDD with data from daily routine clinical practice.

There are also limitations to consider. It should be noted that we relied on meta-analyses, which might have overestimated the efficacy of treatments for depression because of publication bias. We could only include a limited number of meta-analyses. Meta-analyses of psychotherapy that reported within-group results were scarce.

There was a considerable loss-to-follow up in our naturalistic sample. However, the loss-to follow-up-analysis showed that our patient selection was fairly representative for daily-practice-patients who receive treatment. Nevertheless, there was a small under-representation of employed patients and patients with higher baseline severity of depression.

Due to the loss-to-follow-up, the subgroups for each treatment modality were rather small. To assess possible power problems, we computed the differences between efficacy and effectiveness for a situation in which the number of patients would have been ten times larger. In this scenario, we found that still all but one meta-analysis reported significantly better outcomes than our results. For STAR*D, the differences remained non-significant. We therefore conclude that the relatively small sample-size did not importantly influence our main findings.

Although we believe our sample to be representative of an out-patient population with MDD, this might not be completely the case for the setting. The fact that patients were monitored and patients and therapists received feedback may have influenced treatment outcome. Previous research has shown improvement of treatment outcome by monitoring [50,51]. Finally, to allow comparison with meta-analyses we had to use "classical" measures of outcome like percentages remission and response. The validity of these definitions of treatment outcome for daily practice has been questioned [52-55]. In previous research [40] a cutoff of a MADRS score ≤5 (equivalent to "completely recovered" on the CGI, in our ROM population only 9.5% of the patients) has been suggested as a more valid definition of remission. The use of lower thresholds for remission or other definitions of treatment outcome, together with more advanced techniques of statistical modelling, might yield

more useful information on outcome in daily practice, but may also diminish the possibilities for comparison with previous scientific literature.

In conclusion, our results indicate that the outcomes of treatments for MDD in routine clinical practice, which is predominantly of mild to moderate severity, are indeed less favorable than the outcomes reported in meta-analyses of RCTs of different treatments for MDD. Further research into factors that influence outcome in routine clinical care is needed to optimize treatment for patients with MDD.

REFERENCE LIST

1. Fava GA, Tomba E: New modalities of assessment and treatment planning in depression: the sequential approach. *CNS Drugs* 2010, 24: 453-465.
2. IJff MA, Huijbregts KM, van Marwijk HW, Beekman AT, Hakkaart-van Roijen L, Rutten FF *et al*.: Cost-effectiveness of collaborative care including PST and an antidepressant treatment algorithm for the treatment of major depressive disorder in primary care; a randomized clinical trial. *BMC Health Serv Res* 2007, 7: 34.
3. Tunis SR, Stryer DB, Clancy CM: Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003, 290: 1624-1632.
4. Mulder RT, Frampton C, Joyce PR, Porter R: Randomized controlled trials in psychiatry. Part II: their relationship to clinical practice. *Aust N Z J Psychiatry* 2003, 37: 265-269.
5. Licht RW, Gouliaev G, Vestergaard P, Frydenberg M: Generalizability of results from randomized drug trials. A trial on antimanic treatment. *Br J Psychiatry* 1997, 170:264-7.: 264-267.
6. Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *Am J Psychiatry* 1999, 156: 5-10.
7. Rothwell PM: External validity of randomized controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005, 365: 82-93.
8. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967, 6: 278-296.
9. Pigott HE, Leventhal AM, Alter GS, Boren JJ: Efficacy and effectiveness of antidepressants: current status of research. *Psychother Psychosom* 2010, 79: 267-279.
10. de Beurs E., den Hollander-Gijsman ME, van Rood YR, van der Wee NJ, Giltay EJ, van Noorden MS *et al*.: Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clin Psychol Psychother* 2011, 18: 1-12.
11. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
12. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J: An inventory for measuring depression. *Arch Gen Psychiatry* 1961, 4:561-71.: 561-571.
13. Entsuah AR, Huang H, Thase ME: Response and remission rates in different subpopulations with major depressive disorder administered venlafaxine, selective serotonin reuptake inhibitors, or placebo. *J Clin Psychiatry* 2001, 62: 869-877.
14. Davidson JR, Meoni P, Haudiquet V, Cantillon M, Hackett D: Achieving remission with venlafaxine and fluoxetine in major depression: its relationship to anxiety symptoms. *Depress Anxiety* 2002, 16: 4-13.
15. Stahl SM, Entsuah R, Rudolph RL: Comparative efficacy between venlafaxine and SSRIs: a pooled analysis of patients with depression. *Biol Psychiatry* 2002, 52: 1166-1174.
16. Einarson TR, Arikian SR, Casciano J, Doyle JJ: Comparison of extended-release venlafaxine, selective serotonin reuptake inhibitors, and tricyclic antidepressants in the treatment of depression: a meta-analysis of randomized controlled trials. *Clin Ther* 1999, 21: 296-308.
17. Storosum JG, Elferink AJ, van Zwieten BJ, van den BW, Gersons BP, van Strik R *et al*.: Short-term efficacy of tricyclic antidepressants revisited: a meta-analytic study. *Eur Neuropsychopharmacol* 2001, 11: 173-180.
18. Nelson JC: A review of the efficacy of serotonergic and noradrenergic reuptake inhibitors for treatment of major depression. *Biol Psychiatry* 1999, 46: 1301-1308.
19. Steffens DC, Krishnan KR, Helms MJ: Are SSRIs better than TCAs? Comparison of SSRIs and TCAs: a meta-analysis. *Depress Anxiety* 1997, 6: 10-18.
20. Bech P, Cialdella P, Haugh MC, Birkett MA, Hours A, Boissel JP *et al*.: Meta-analysis of randomized controlled trials of fluoxetine v. placebo and tricyclic antidepressants in the short-term treatment of major depression. *Br J Psychiatry* 2000, 176:421-8: 421-428.

21. Thase ME, Entsuah AR, Rudolph RL: Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *Br J Psychiatry* 2001, 178:234-41: 234-241.
22. de Maat SM, Dekker J, Schoevers RA, de Jonghe F: Relative efficacy of psychotherapy and combined therapy in the treatment of depression: a meta-analysis. *Eur Psychiatry* 2007, 22: 1-8.
23. Thase ME, Greenhouse JB, Frank E, Reynolds CF, III, Pilkonis PA, Hurley K *et al*.: Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Arch Gen Psychiatry* 1997, 54: 1009-1015.
24. Wexler BE, Cicchetti DV: The outpatient treatment of depression. Implications of outcome research for clinical practice. *J Nerv Ment Dis* 1992, 180: 277-286.
25. Montgomery SA: A meta-analysis of the efficacy and tolerability of paroxetine versus tricyclic antidepressants in the treatment of major depression. *Int Clin Psychopharmacol* 2001, 16: 169-178.
26. Kasper S, Zivkov M, Roes KC, Pols AG: Pharmacological treatment of severely depressed patients: a meta-analysis comparing efficacy of mirtazapine and amitriptyline. *Eur Neuropsychopharmacol* 1997, 7: 115-124.
27. Beasley CM, Jr., Nilsson ME, Koke SC, Gonzales JS: Efficacy, adverse events, and treatment discontinuations in fluoxetine clinical studies of major depression: a meta-analysis of the 20-mg/day dose. *J Clin Psychiatry* 2000, 61: 722-728.
28. Thase ME, Haight BR, Richard N, Rockett CB, Mitton M, Modell JG *et al*.: Remission rates following antidepressant therapy with bupropion or selective serotonin reuptake inhibitors: a meta-analysis of original data from 7 randomized controlled trials. *J Clin Psychiatry* 2005, 66: 974-981.
29. Minami T, Wampold BE, Serlin RC, Kircher JC, Brown GS: Benchmarks for psychotherapy efficacy in adult major depression. *J Consult Clin Psychol* 2007, 75: 232-243.
30. BJ Becker: Synthesizing standardized mean-change measures.S. *British Journal of Mathematical and Statistical Psychology* 1988, 41: 257-278.
31. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L *et al*.: Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry* 2006, 163: 28-40.
32. Thase ME, Friedman ES, Biggs MM, Wisniewski SR, Trivedi MH, Luther JF *et al*.: Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR*D report. *Am J Psychiatry* 2007, 164: 739-752.
33. National Taskforce Guideline. Multidisciplinaire Richtlijn voor diagnostiek en behandeling van volwassen cliënten met een depressie, herziene versie. 1-1-2005. Stuurgroep Richtlijnen/ Trimbos Instituut.
34. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E *et al*.: The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998, 59 Suppl 20: 22-33.
35. Hawley CJ: Depression rating scales can be related to each other by simple equations. 1998.
36. Mittmann N, Mitter S, Borden EK, Herrmann N, Naranjo CA, Shear NH: Montgomery-Asberg severity gradations. *Am J Psychiatry* 1997, 154: 1320-1321.
37. Zimmerman M, Posternak MA, Chelminski I: Defining remission on the Montgomery-Asberg depression rating scale. *J Clin Psychiatry* 2004, 65: 163-168.
38. Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D *et al*.: The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol* 2006, 16: 601-611.
39. Hawley CJ, Gale TM, Sivakumaran T: Defining remission by cut off score on the MADRS: selecting the optimal value. *J Affect Disord* 2002, 72: 177-184.
40. Bandelow B, Baldwin DS, Dolberg OT, Andersen HF, Stein DJ: What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? *J Clin Psychiatry* 2006, 67: 1428-1434.
41. Altman DG: Practical Statistics for Medical Research. 1991:229-276.

42. Altman DG: Practical Statistics for Medical Research. 1991:179-228.
43. Stewart JW, McGrath PJ, Quitkin FM: Can mildly depressed outpatients with atypical depression benefit from antidepressants? *Am J Psychiatry* 1992, 149: 615-619.
44. Wisniewski SR, Rush AJ, Nierenberg AA, Gaynes BN, Warden D, Luther JF *et al*.: Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report. *Am J Psychiatry* 2009, 166: 599-607.
45. Lem Rvd, Wee Nvd, Veen Tv, Zitman FG: The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychological Medicine* 2010.
46. Zimmerman M, Posternak MA, Chelminski I: Symptom severity and exclusion from antidepressant efficacy trials. *J Clin Psychopharmacol* 2002, 22: 610-614.
47. Leonard KL: Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the Hawthorne effect. *J Health Econ* 2008, 27: 444-459.
48. Demyttenaere K, Adelin A, Patrick M, Walthere D, Katrien dB, Michele S: Six-month compliance with antidepressant medication in the treatment of major depressive disorder. *Int Clin Psychopharmacol* 2008, 23: 36-42.
49. Bech P: Struggle for subtypes in primary and secondary depression and their mode-specific treatment or healing. *Psychother Psychosom* 2010, 79: 331-338.
50. McKenzie N, Marks I: Routine monitoring of outcome over 11 years in a residential behavioural psychotherapy unit. *Psychother Psychosom* 2003, 72: 223-227.
51. McKay R, McDonald R: Expensive detour or a way forward? The experience of routine outcome measurement in an aged care psychiatry service. *Australas Psychiatry* 2008, 16: 428-432.
52. Jacobson NS, Truax P: Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991, 59: 12-19.
53. Schmitz N, Hartkamp N, Franke GH: Assessing clinically significant change: application to the SCL-90-R. *Psychol Rep* 2000, 86: 263-274.
54. Barkham M, Stiles WB, Connell J, Twigg E, Leach C, Lucock M *et al*.: Effects of psychological therapies in randomized trials and practice-based studies. *Br J Clin Psychol* 2008, 47: 397-415.
55. Moleiro C, Beutler LE: Clinically significant change in psychotherapy for depressive disorders. *J Affect Disord* 2009, 115: 220-224.