Are depression trials generalizable to clinical practice?

Something clinicians always wanted to know about RCTs, but were afraid to ask……

Rosalind van der Lem

Are depression trials generalizable to clinical practice?

Something clinicians always wanted to know about RCTs, but were afraid to ask……

Rosalind van der Lem

Lay-out: Legatron Electronic Publishing, Rotterdam
Printing: Ipskamp Drukkers BV, Enschede

ISBN/EAN: 9789461917218

Are depression trials generalizable to clinical practice?

Something clinicians always wanted to know about RCTs,

but were afraid to ask……

Proefschrift

Ter verkrijging van de graad van Doctor

aan de Universiteit Leiden,

op gezag van de Rector Magnificus Prof. Mr. C.J.J.M. Stolker,

volgens het besluit van het College voor Promoties

te verdedigen op woensdag 12 juni 2013

klokke 11:15 uur

door

Rosalind van der Lem

Geboren te 's Gravenhage in 1974

PROMOTIECOMMISSIE:

Promotor:	Prof. Dr. F.G. Zitman
Copromotores:	Dr. N.J.A. van der Wee
	Dr. T. van Veen
Overige leden:	Prof. Dr. P. Spinhoven
	Prof. Dr. R.A. Schoevers (UMCG)
	Prof. Dr. A.M. van Hemert
	Prof. Dr. R.R.J.M. Vermeiren

This research project was financially supported by the independent research fund from The Dutch Government ZonMW (grant number 100-002-026OOG) and Rivierduinen GGZ Leiden in collaboration with the Leiden University Medical Center.

Voor mijn beste vriend Brian Comanne, psychiater in opleiding.

Mijn beste vriend is vorig jaar overleden, hij had er graag bij willen zijn.

CONTENTS

Chapter 1

General Introduction

Dr. X is a well respected psychiatrist who is 60 years old and works in a large psychiatric outpatient clinic. Every day, he sees many patients with a wide range of psychopathology. Often, younger colleagues refer complex patients to him because of his extensive experience. During his career, he has witnessed many developments in psychiatry: new types of medication, the anti-psychiatry movement, empowerment of patients, the diminishing popularity of psychoanalytic therapy, the upcoming of protocollized therapies, and the progress of molecular and genetic insights in psychiatric disorders. In his outpatient clinic, like in many others, the national guidelines for the treatment of psychiatric disorders have been embraced and implemented. Like many of his colleagues, dr. X was interested, yet sceptical, and worried that guidelines would make all creativity in his profession disappear. Nevertheless, dr. X committed to the treatment algorithms used in his institution. He kept up with the scientific publications on medication and psychotherapy, especially on major depressive disorder (MDD), since most of his patients suffered from depression. He read the promising results of randomized clinical trials (RCTs) on different drugs and new methods of psychotherapy. Meanwhile, in his clinical practice, the results of medication or psychotherapy were often disappointing and patients kept struggling with their depression. Dr. X got the impression that treatment for MDD in RCTs is a lot more successful than in "real life". He started to wonder: do my patients even look like those in RCTs? How should I interpret the results from RCTs? Do RCTs tell us anything about "real life"? Is it right to base treatment guidelines for daily practice on results from RCTs that might be so far away from daily practice?
This thesis is about Dr. X's questions.

Not so long ago, the treatment of psychiatric disorders was based on the personal expertise and interests of individual psychiatrists. Nowadays, evidence based medicine has become the 'gold standard' for clinical practice. In this respect, modern psychiatry does not differ from other medical specialties. Treatments proven to be effective in randomized clinical trials (RCTs) are transformed into clinical practice guidelines, which are implemented in routine clinical practice. Treatments (yet) without evidence are left aside. But how "golden" is this modern medical standard? Are therapies that have been proven effective in the strict research setting of RCTs as effective in routine clinical care? Clinical practice guidelines are based on results from RCTs. But are results from clinical trials generalizable to daily psychiatric practice?

In this thesis, we aim to establish to what extent results from RCTs are applicable to daily practice for patients suffering from major depressive disorder (MDD), one of the most common psychiatric disorders. Next, we explore factors that may influence the generalizability of results from clinical trials in MDD to daily practice.

How are results from randomized clinical trials used in daily practice?

Like dr. X, most psychiatrists in the Western world now follow evidence based guidelines on the treatment of MDD. Often, guidelines are presented as or implemented in treatment algorithms that are used in daily practice. In Western psychiatry, the guidelines of the American Psychiatric Association (APA) and the National Institute of Clinical Excellence (NICE) guidelines in the UK are well known. Most other countries have developed similar guidelines for the treatment of MDD based on scientific evidence. In the Netherlands, the guidelines are developed by a national task-force for guideline development: Landelijke Stuurgroep Multidisciplinaire Richtlijnontwikkeling in de GGZ and are published by the Netherlands Institute on Mental Health and Addiction (Trimbos Instituut).

In every guideline a clear description of the way it was constructed is given. They all rely heavily on evidence from RCTs, and in most guidelines, the reliability of evidence from scientific research has been ranked (weighted). Below are the descriptions that two well-known professional organisations give of their methodology. We also describe the methods used to weigh the evidence for the multidisciplinary guidelines for depression in the Netherlands.

Guideline of the American Psychiatric Association

(APA, United States of America, http://www.psych.org)

"This guideline strives to be as free as possible of bias toward any theoretical posture, and it aims to represent a practical approach to treatment. Studies were identified through an extensive review of the literature by using MEDLARS for the period 1971–1999. Major review articles and standard psychiatric texts were consulted. The Agency for Healthcare Policy Research Evidence Report on Treatment of MDD-Newer Pharmacotherapies [14] was reviewed in its entirety. Review articles and relevant clinical trials were reviewed in their entirety; other studies were selected for review on the basis of their relevance to the particular issues discussed in this guideline. Definitive standards are difficult to achieve, except in narrow circumstances in which multiple replicated studies and wide clinical opinion dictate certain forms of treatment. In other areas, the specific choice among two or more treatment options is left to the clinical judgment of the clinician. The recommendations are based on the best available data and clinical consensus with regard to the particular clinical decision. The summary of treatment recommendations is keyed according to the level of confidence with which each recommendation is made."

Guideline of the National Institute of Clinical Excellence

(NICE, United Kingdom, http://www.nice.org.uk)

"The systematic identification of evidence is an essential step in clinical guideline development. Systematic literature searches undertaken to identify evidence of clinical and cost effectiveness should be thorough, transparent and reproducible. These searches will

also minimize 'dissemination biases' [15], such as publication bias and database bias, that may affect the results of reviews."

Guideline from the Netherlands Institute of Mental Health and Addiction
(Trimbos Institute, the Netherlands, http://www.ggzrichtlijnen.nl)
"A guideline is based on results from scientific research and additional opinions by professionals and patients, and aims to specifically describe good medical practice. In this partial revision of the guideline, the EBRO method of evidence based guideline development is used and the assumptions of the Landelijke Stuurgroep Multidisciplinaire Richtlijnontwikkeling in de GGZ are followed. Subsequently the Appraisal of Guidelines for Research & Evaluation (AGREE) instrument has been used. AGREE is a European instrument to assess the quality of guidelines. Finally, the Health Technology Assessment was used in the substantiating of the recommendations." (translated from Dutch). Scientific evidence is evaluated as follows in the Dutch Guidelines (in order of methodological rigour):

A1. Systematic review of at least two independent research projects of A2 level.
A2. Randomized, double-blind, clinical trials of good quality and large enough sample size.
Research comparing a method to a reference test (gold standard) with beforehand defined outcome and independently judged results in a large enough sample size of patients who had both the investigated method and the reference test.
Prospective cohort study with large enough sample size, controlled for confounding and selective follow-up.
B. Clinical trials, without the methodological rigour mentioned in A2.
Research comparing a method to a reference test (gold standard) without the methodological rigour mentioned in A2.
Prospective cohort study without the methodological rigour mentioned in A2
C. Non-comparative research.
D. Expert opinion.

What is Major Depressive Disorder?

Major depressive disorder (MDD) is one of the most common psychiatric disorders. Patients with MDD suffer from a depressed mood and/or loss of interest or pleasure, often accompanied by loss of weight, disturbed sleep, psychomotor agitation or retardation, loss of energy, feeling of worthlessness, loss of concentration and recurrent thoughts of death [2]. According to the World Health Organisation (https://www.who.int/en), MDD is the leading cause of disability as measured by Years Lived with Disability (YLDs), and the fourth leading contributor to the global burden of disease (Disability Adjusted Life Years, DALY). DALY measures the total number of days lived with disability of a population. By the year 2020, MDD is projected to reach second place of the ranking of DALYs regardless of age and gender. Today, MDD is already the second cause of DALYs in the age category 15–44 years for both sexes combined. MDD occurs in persons of both genders, all ages, and regardless of ethnic and social backgrounds and affects about 121 million people worldwide. In the Netherlands, the lifetime prevalence for MDD is 10.9% for men and 20.1% for women. The 12-month prevalences are 4.1% and 7.5%, respectively [9]. The number of DALY's in the Netherlands for MDD is 158.000 per year. Besides the unmistakable suffering of individual patients and their loved ones, MDD has substantial economic consequences for society: patients suffering from MDD use more health care and social security, and MDD causes a loss in production due to absence. The costs of treatment for MDD in the Netherlands amount to 660 million euros per year. Besides these costs, 953 million euros are lost due to absence from employment. In total, the costs for MDD are 1.1% of the total healthcare costs in the Netherlands [11] (https://www.trimbos.nl).

First step-treatments for MDD according to the guidelines

In the guidelines mentioned above, the use of either pharmacotherapy or psychotherapy is recommended as first treatment step for moderate MDD in psychiatric outpatient practice [16,17]. Both therapies have been proven to be effective for patients who are suffering from MDD for longer than three months. Pharmacotherapy and psychotherapy are equally effective in patients suffering from moderate-to-severe MDD [18]. For patients suffering from (very) severe MDD, the guidelines recommend antidepressant medication as first treatment step. In the past, different types of antidepressants, selective serotonin reuptake inhibitors (SSRIs), tricyclic antidepressants (TCAs), venlafaxine and mirtazapine have been shown to be equally effective. However recent studies indicate differences in efficacy and tolerability [19]. Regarding psychotherapeutic treatment for MDD, cognitive behavioral therapy (CBT), behavioral therapy (BT) and interpersonal therapy (IPT) are recommended. All have been proven to be effective, and so far, few differences have been found in the efficacy of CBT, BT and IPT [20]. Currently, studies on CBT do outnumber IPT trials though.

For dr. X, the guidelines provide a clear algorithm of the subsequent evidence based steps that he has to take when treating patients suffering from MDD. He is however still puzzled by questions on the generalizability of the results from RCTs, which are conducted in a strict research setting, to his daily practice. In the next paragraph, we describe several methodological aspects of RCTs that are related to the generalizability of RCT results.

What clinicians always wanted to know about the methodology of RCTs, but were afraid to ask...

In order to answer the question: "Are results from RCTs generalizable to daily practice?" we first have to know how RCTs are designed. RCTs are also called **efficacy**-trials: their results describe the impact of treatment on the disease (e.g. MDD) in a defined population (e.g. patients suffering from MDD without co morbid disorders within a certain age-range). To obtain the most reliable results in efficacy trials, much effort is put in optimization of the **internal validity** of the trial: the extent to which a result reflects the real causal relationship between a compound (investigated treatment) and change in disease status. Results from efficacy trials need to be replicable and solely contributable to the investigated treatment. For example, when in an RCT, CBT in MDD has proven to be **effective**, this means that for a group of patients usually between 18–55 years old, suffering from MDD without co morbid disorders, CBT applied according to the protocol has been proven to be more effective than a placebo treatment or treatment as usual within a defined period of time. But dr. X does not treat groups of patients, does not apply patient selection, and does not treat patients with a placebo treatment. He simply treats individual patients, each with their own specific features, and he merely wants to know: "what do I tell my patients about the chances of recovery when I apply a treatment that is in my guidelines?" **Efficacy** might be a useful way to describe the influence of a specific treatment on disease status, but for daily practice, the concept of **effectiveness** is more appropriate. Effectiveness is a broader concept than efficacy. It may comprise a number of outcomes (e.g. efficacy, tolerability, costs of treatment, outcome in social functioning or quality of life). It can be defined as the impact of the treatment on the disease in a general population. For MDD and most other psychiatric disorders, it is still unknown how efficacy and effectiveness relate to one another. Is outcome the same in a trial setting and daily practice, when the same treatments are applied? Both researchers and clinicians would intuitively state: "Probably not!". But how large is the difference between efficacy and effectiveness?

Whereas **internal** validity is essential for the evidence of efficacy of treatments, **external validity** is equally important or even more important for the evidence for effectiveness of a treatment in a daily practice. External validity is defined as the extent to which a result can be generalized to a larger (real world) population with more heterogeneous characteristics. External validity is equivalent to **generalizability**. Internal and external validity are sometimes the two ends of the same balance: if one improves the internal validity, one decreases the

external validity. Internal and external validity might seem methodological concepts that are important for researchers and methodologists, but not for clinicians like dr. X. However, in order to judge whether results from RCTs can be generalized to daily practice, it is also of clinical relevance to understand the relationship between these two concepts.

The following methods are used in efficacy trials to optimize internal validity: randomization, blinding, sample size calculation (power estimation), and the strict use of eligibility criteria. Some of these methods might jeopardize the generalizability of the outcome of the trial to routine clinical practice, while others do not. In the frames below we will give an overview of the methods used to improve the internal validity of RCTs and their influence on external validity [21]. Furthermore, we will describe the effect of these methods on the difference between efficacy and effectiveness [7,22-30].

Methods in RCTs to improve internal validity that do not jeopardize external validity

Randomization is used to ensure that unknown factors that could influence the result (e.g. age, gender, baseline severity of the disease, co morbid disorders) will be equally distributed in both the treatment and the control group. By randomization possible confounding (confounders are factors that influence treatment outcome if they are unequally distributed between treatment and control group) is neutralized. Randomization does not influence external validity.

Blinding is an attempt to prevent investigators and/or participants from influencing the identification of relevant events during a trial. In other words, if the participant and/or the researcher do not know whether the participant receives the active drug or placebo, they are not biased by this knowledge in observing the effect. Blinding is used to optimize internal validity by ruling out placebo-effect as much as possible. It is often difficult to guarantee complete blinding in antidepressant-trials, since antidepressants cause specific side-effects that are impossible to mimic in placebo-pills. For psychotherapy, blinding is even more difficult and requires creative procedures [3-5]. As an alternative to complete double-blinding, independent (blind) outcome-rating personnel is often used in trials. Blinding does not influence external validity.

Sample size calculation: the sample size defines the robustness of the result of the efficacy trial. A larger sample size provides more accurate findings (and narrower confidence intervals and smaller p-values). Sample size calculations (power estimation) are performed in advance of the start of the efficacy trials. Sample size calculation does not influence external validity and has no effect on the difference between efficacy and effectiveness. Nevertheless, it is important for clinicians to take into account that, while interpreting results from efficacy-trials for daily practice, p-values and confidence intervals are influenced by two factors: by the magnitude of the found difference between an investigated treatment and control-condition and by the sample size. For example, a difference in proportion of remitters of 5% between investigated treatment and control-condition can be highly significant if the sample size is large, but does not tell you what the clinical relevance of the found difference is. It is up to the clinician to judge the clinical relevance of efficacy results for daily practice. Statistical analysis is no more and no less than an estimation of the magnitude of a result and an estimation of the probability of finding these results. Clinicians who are not very familiar with statistical analysis might easily be impressed by very small p-values when reading reports on clinical trials.

Randomization and *Blinding* can contribute to possible differences between efficacy and effectiveness: in daily practice the clinician judges whether a certain treatment is more appropriate, or more likely to be successful for his individual patient based on several features of this patient (e.g. age, gender, co morbid disorders). Also, the patient can express his preference for a certain treatment. In a (double) blind trial, the patient's preference for a specific treatment is not taken into account. Some patients might refuse participation in RCTs if the treatment of their choice is not investigated in the trial. Furthermore, the possibility by itself of clinicians and patients to choose a treatment might be associated with a better treatment outcome [12,13]. Thus, if randomization and blinding were the only differences between a trial setting and daily practice, one would probably expect to find better results in daily practice.

Methods in RCTs to improve internal validity that might jeopardize external validity

The use of eligibility criteria: in efficacy trials stringent in- and exclusion criteria are used. The use of these criteria is vital for methodological reasons: only in a homogenous (e.g. without co morbid disorders, with sufficient severity etc.) patient population the difference found in outcome can be solely attributed to the investigated treatment. Furthermore, the use of exclusion criteria might be inevitable for ethical reasons: e.g. risk of suicide, risk of teratogenic effects of the investigated drug in pregnancy; risk of dangerous or intolerable side effects of certain drugs to specific patient groups etc. The use of strict eligibility criteria facilitates analysis and detection of differences in treatment outcome between groups. Therefore, the use of strict eligibility criteria in efficacy trials is essential during the development of a new compound. However, the generalizability of the results to routine care is usually poor, since the results are only applicable to a small, selected part of the patient population in clinical practice. It is not clear to what extent clinical practice may benefit from results of RCTs when the generalizability of these results is poor [7]. This topic will be addressed in detail in the paragraph "The influence of (un) intended patients selection on treatment outcome in RCTs" in the Introduction Chapter of this thesis.

Other aspects of RCTS that may hamper the external validity

The trial setting: the circumstances under which the trial takes place might differ in many aspects from the routine clinical practice of the clinician who wants to find evidence for a treatment. The trial can be conducted in another country than that of the clinician, where they use other methods of diagnosis and management, where the susceptibility to the disease in the population is different, or where the health care at the location of the trial is organised in a different way (length of waiting lists, access to health care, financial limits). Furthermore, trials are often conducted in very specialized centres and by highly trained and motivated clinicians with ample time for the protocollized treatment of every individual trial participant without the daily time pressure so common in routine clinical practice.

The selection of patients before or beyond consideration of eligibility criteria: due to recruitment procedures, unintended patient selection might occur. This topic is addressed in detail in the paragraph "The influence of (un)intended patient selection on treatment outcome in RCTs" in the Introduction Chapter of this thesis.

Other aspects of RCTS that may hamper the external validity *(Continued)*

The use of run-in periods and/or enrichment strategies: run-in periods of medication are used to exclude patients who are poorly compliant or who suffer from unacceptable side effects. Exclusion of these patients does probably jeopardize external validity. Likewise, the use of enrichment-strategies in which patients who are likely to respond well are actively recruited might jeopardize external validity.

Pre-trial treatment or non-trial management: patients who need medication for other medical conditions (non-trial management) are often excluded from participation in efficacy trials. Exclusion of these patients might jeopardize the generalizability of the results. Furthermore, in some RCTs, specific preparation of participants for the trial is conducted (pre-trial treatment), which probably influences treatment outcome in RCTs and therefore might contribute to the efficacy-effectiveness difference.

Treatment in the control group of efficacy trials: the control condition in the trial sometimes differs very much from daily practice, which hampers generalizability.

The definition of outcome and duration of follow-up period: sometimes in trials, outcome measures that are not clinically relevant are used and the follow-up period is usually short. Therefore the generalizability to daily practice might be poor.

How is successful treatment outcome in MDD defined in RCTs?

In RCTs, results of treatment of MDD have been defined in many different ways. Different instruments have been used to assess treatment progression and final results. The most common method to evaluate treatment is the use of questionnaires. These questionnaires can be generic, which means that they measure improvement in general terms of "well being", or "quality of life". They can also measure the severity of symptoms of a specific disorder. For MDD, outcome can be rated on symptoms like anhedonia, loss of sleep, and depressed mood. Furthermore, questionnaires can either be self report instruments, which means that the patients fill in the questionnaires by themselves, or observer rated, which means that the severity of symptoms is assessed by an observer, usually a clinically trained person. In antidepressant efficacy trials (AETs), the most commonly used instruments to define primary treatment outcome are the Hamilton Rating Scale for Depression (HAMD), 17-item or 21-item version [31] or the Montgomery-Asberg Depression Rating Scale (MADRS) [32]. These instruments are both symptom-specific, observer rated instruments. In psychotherapy efficacy trials (PETs), the most commonly used instrument to define primary treatment outcome is the Beck Depression Inventory (BDI) [33], a symptom-specific self report questionnaire. Often other instruments are used in efficacy trials to assess secondary

outcome (quality of life, social functioning, additional disease specific instruments, instruments that measure tolerance for medication or treatment adherence, etc.). AETs and PETs typically use different definitions of treatment outcome:

In general, AETs use response and remission percentages to define outcome.

> Response percentage: proportion of MDD patients who reach a reduction of symptoms of 50% or more.
>
> Remission percentage: proportion of MDD patients who reach a symptom level below a defined cut-off score.

PETS generally use effect size, written in abbreviated form as Δ, [34] as the definition of outcome.

> Effect size Δ: difference in symptom level pre-and post treatment, controlled for sample size
>
> $\Delta = (\mu pre - \mu post) / \sigma$
>
> μpre = mean pre-treatment
>
> μpost = mean post-treatment
>
> σ = standard deviation pre-treatment

During the last decade, several researchers have introduced other definitions for treatment outcome in research and in daily practice, e.g. clinical significant change. These recent definitions of treatment success might have more clinical relevance [35-38]. However, they have not been used in RCTs, yet. So to compare the available body of RCTs and daily practice, we have to use the same definitions as used in RCTs.

Treatment outcome for MDD in daily practice: how can we assess success?

In daily practice, systematic evaluation of treatment progress is needed to provide insight in the course of individual therapies, or of groups of patients suffering from the same psychiatric disorder such as MDD. Routine Outcome Monitoring (ROM) comprises the systematic assessment of patients in daily practice. ROM provides data on treatment effects in daily practice that allows clinicians to evaluate treatment progress. It also allows researchers to explore treatment success in routine clinical practice in general, and factors associated with success. Through ROM, many clinical research questions can be addressed scientifically. The data on routine clinical practice used in this thesis are derived from the ROM system of Rivierduinen, which is described in detail in the frame below.

In spring 2002, the Regional Mental Health Provider (RMHP) 'Rivierduinen' (an institute serving a region with more than 1 million inhabitants) and the Department of Psychiatry of the Leiden University Medical Center (LUMC) started collaboration for routine assessment of the DSM-IV diagnosis as well as the symptom severity, well-being and health status at time of the first interview of outpatients referred to the RMHP Rivierduinen.

At the start, ROM was restricted to patients referred for treatment of mood, anxiety, and somatoform (MAS) disorders. These patients form a relatively homogenous group with substantial mutual co morbidity [1] and they mainly receive outpatient care. To be eligible, patients had to have sufficient mastery of the Dutch language and had to be able to complete self report instruments. Patients who are considered (by their clinician) to be too ill to complete questionnaires or who refuse to be assessed are excluded from ROM assessment.

All patients are assessed by an independent psychiatric research nurse at the start, and during follow up at intervals of three to four months, at the beginning of a new treatment step and at the end of the treatment.

During the first session, a standardized diagnostic interview is administered and observer- and self reported ratings are determined. At baseline the Axis-I diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) is established using the Mini-International Neuropsychiatric Interview-plus [6]. The interviews are performed by psychiatric research nurses who have been extensively trained and supervised. The Dimensional Assessment of Personality Pathology (DAPP-SF) is administered to assess maladaptive personality traits [8].

Subsequently, a number of symptom severity rating scales is administered at baseline, and is also completed at each re-assessment to allow for the evaluation of treatment outcome. Together, these instruments cover change in three areas of functioning: symptom reduction, increased wellbeing, and improvement in general life functioning [10]. They are commonly used in treatment outcome research and have good psychometric properties as evidenced by national and international publications (an overview of instruments used is available at http://www.lumc.nl/psychiatry/ROM-instruments). Outcome is assessed by patients' self report and by an independent assessor, and includes both generic and disorder-specific measures. Clinicians receive a report on the results of the baseline assessments as well as follow-up reporting on treatment outcome in the above mentioned domains. Results of the assessments are provided in detail by the research nurses as well as in a summarized form. The summaries facilitate clinicians to discuss the results with their patients and use them as a tool to evaluate the treatment. Results are also used, in an anonymous form, for scientific purposes. Since ROM-data are primarily being used by clinicians and patients to monitor treatment progress, no specific informed consent is needed. The use of anonymized data for research purposes has been approved by the Medical Ethical Committee of the LUMC.

The influence of (un) intended patient selection on treatment outcome

As mentioned above, results from efficacy trials in MDD might not be applicable to daily practice because of the use of stringent in- and exclusion criteria for patient selection in these trials. A fairly recent solution to this problem is the so called "pragmatic trial". Pragmatic trials are designed to optimize the generalizability of the results, and therefore use broader inclusion criteria. For instance, for participation in pragmatic trials some co morbid Axis I and II disorders [39] are allowed. Pragmatic trials provide more information for daily practice, but the certainty of a causal relationship between investigated treatments and outcome and the reproducibility is less clear. As mentioned before, the efficacy of first step MDD treatments (antidepressants, cognitive behavioral therapy and interpersonal therapy) is investigated in AETs and PETs. In this thesis we aim to explore the differences between efficacy and effectiveness of antidepressant treatment as well as individual psychotherapy for MDD. Both AETs and PETs use exclusion criteria for their selection of patients. The consistency of exclusion criteria across AETs has been explored in previous research and a set of consistently used exclusion criteria was identified (see below) [40,41]. Remarkably, which eligibility criteria are consistently used in psychotherapy efficacy trials (PETs) for MDD was not studied in previous research. In this thesis we explore for the first time the consistency of eligibility criteria across PETs.

It has been demonstrated that the use of exclusion criteria in AETs leads to exclusion of many MDD patients [30,42,43]. The use of exclusion criteria might also influence the outcome of AETs, i.e. patients not meeting an exclusion criterion might do better. Limited data on the influence of eligibility criteria on outcome in AETs are available. A study found that patients who would be eligible for AETs had a more favorable outcome in clinical practice, but this has not been explored further [44]. As mentioned above, previous research identified a set of consistently used exclusion criteria across AETs. The following criteria were found to be consistently used in AETs: history of DSMIV manic or hypomanic episodes; presence of psychotic features in current depression; significant risk of suicide, alcohol/drug abuse or dependency; mild MDD (not meeting a baseline severity of 18 on the Hamilton Rating Scale for Depression, 17 item version, HAMD17 [31]); presence of underlying dysthymic disorder; presence of non-depressive, non-substance use co morbid Axis I disorders; presence of borderline personality disorder.

Clearly, the use of these exclusion criteria might hamper the generalizability of the results of AETs. Exclusion of patients suffering from bipolar depression or from MDD with psychotic features, which is very often done in AETs, will limit the generalizability of results from MDD trials to bipolar patients and patients suffering from MDD with psychotic features. However, as bipolar disorder and MDD with psychotic features are considered to be separate entities of MDD that are covered in trials especially designed for those target populations, the use of these exclusion criteria does not hamper clinicians in their evaluation of the usefulness of RCTs for their patients. Exclusion of suicidal patients, patients with co morbid substance

abuse disorders, patients with other Axis I or Axis II disorders and patients suffering from milder MDD, however, will hinder the generalizability of results from RCTs to daily practice. This is because many "real world" MDD patients suffer from suicidal ideations, substance abuse or other co morbid disorders and/or personality pathology. In addition, while in RCTs a minimum depression severity is required, in daily practice many more patients suffer from mild-to-moderate MDD than from severe MDD. These exclusion criteria might also influence the outcome in clinical practice. Substance abuse disorders are associated with poorer treatment outcome [45], presence of other Axis I disorders also seem to be associated with poorer treatment outcome, although results are not unambiguous [45-48]. Suicidality seems to be associated with treatment resistance [47] and the presence of personality pathology seems to be associated with poorer or different treatment outcome, but also in this field the results are not unambiguous [47,49-51]. Milder MDD is sometimes associated with better treatment outcome [47], but is also associated with poorer response due to a larger effect of regression to the mean in more severe MDD. Outcome research in MDD with co morbid disorders as well as in mild-to-moderate MDD is scarce and the results are contradictory.

In brief, the following exclusion criteria that are consistently used in AETs are relevant for the generalizability of the results of AETs: co morbid Axis I and II disorders, suicidality and mild MDD. In this thesis, we explore the occurrence of these criteria in daily practice and subsequently investigate the eligibility of daily practice patients for MDD trials and the influence of the mentioned features on treatment outcome.

The use of inclusion and exclusion criteria leads to explicit selection, but selection bias in the research population might also occur beforehand as the sample from which participants will be selected may differ from clinician to clinician. For instance, they may already differ with respect to age, sex, race, severity of disease, educational status, social class, and place of residence [21]. Other aspects of recruitment may also contribute to unintended selection bias in sociodemographic and socioeconomic features. For instance, participation in RCTs is usually without costs for the participants. In countries where there is no extensive social security system and patients have to pay themselves for mental healthcare, participation in trials might be the only way to obtain treatment for patients with limited financial means. As a result recruited patients might have lower socioeconomic status than the average patient in daily practice when this is not controlled for. The area in which patients are recruited, and the recruitment strategy (ads in newspapers, certain magazines, internet, through clinicians), may also contribute to unintended selection bias. Finally, as participants will have to agree with the possibility of receiving placebo treatment, this might also introduce selection bias. Together with the use of exclusion criteria, unintended selection bias amounts to an exclusion rate of 73% of the initial patient population available for RCTs. Most of the selection takes place before or beyond consideration of the exclusion criteria [52].

In this thesis, we explore sociodemographic and socioeconomic differences between RCT participants and "real life" MDD patients. Subsequently, we explore the influence of

sociodemographic and socioeconomic features on treatment outcome for MDD in clinical practice.

AIMS AND RESEARCH QUESTIONS

Until recently the choice of therapies for psychiatric disorders was based on personal experience, knowledge and preferences (experience based medicine). Nowadays, it has become common practice to gain evidence for the efficacy of treatments in RCTs, incorporate treatments that are proven to be effective in RCTs in guidelines, and implement these guidelines in routine psychiatric care. However, important questions about the generalizability of results from RCTs to daily psychiatric practice have not been addressed:

1. To what extent does outcome of treatments for MDD in trial settings (efficacy) and in routine clinical practice (effectiveness) differ?
2. Which proportion of "real life" patients would be eligible for participation in MDD trials and what is the influence of exclusion criteria on treatment outcome for MDD in daily practice?
3. Do participants of RCTs on MDD differ from daily practice patients in sociodemographic and socioeconomic features and do sociodemographic and socioeconomic features influence treatment outcome in MDD in daily practice?

CONTENTS OF THE THESIS

In chapter 2, we addressed the first research question: To what extent does outcome of treatments for MDD in trial settings (efficacy) and in routine clinical practice (effectiveness) differ? We examined treatment outcome of antidepressant treatment; individual psychotherapy; and a combination of both. We derived the efficacy results from a large sample of selected meta-analyses. These meta-analyses all provided an aggregated estimate of the within group efficacy of antidepressants, individual psychotherapy and/or combination treatment. We also compared the outcome results from ROM to a large so-called "pragmatic" trial, STAR*D [39], which was designed to be as comparable to daily practice as possible. Outcome of treatments for MDD in routine clinical practice was explored in data derived from ROM. We compared effectiveness results from ROM with the efficacy results of these therapies when investigated in RCTs. We hypothesized that outcome in daily practice would be less favorable than efficacy results from RCTs and closer to the results from STAR*D.

In chapter 3, we addressed the second research question: Which proportion of "real life" patients would be eligible for participation in MDD trials and what is the influence of these eligibility criteria on treatment outcome for MDD in daily practice for AETs? For this purpose,

we used the model of Zimmerman and colleagues [30] on consistency in the use of exclusion criteria in antidepressant trials. Furthermore, we investigated the influence of eligibility, both for the individual exclusion criteria as well as "being eligible" in general, on treatment outcome. We explored how many patients of a large group of ROM patients suffering from MDD would be eligible for AETs. In line with previous research, we hypothesized that only a minority of patients in daily practice will be eligible for participation in AETs. In line with a previous report on the STAR*D trial [44], we also expected patients who are eligible for AETs to have better treatment outcome than patients who are not. If the generalizability of results from AETs would turn out to be hindered by the use of eligibility criteria, this might be an explanation for differences between efficacy and effectiveness.

In chapter 4, we addressed the second research question, but now for PETs. We explored the consistency of exclusion criteria in trials on CBT and IPT. We aimed to create a set of consistently used exclusion criteria, in line with the model of Zimmerman and colleagues [40]. Furthermore, we estimated the influence of commonly used exclusion criteria in PETs on treatment outcome in our ROM population. We hypothesized that patients who meet the exclusion criteria would have better treatment results than patients who do not. If generalizability of results from PETs would turn out to be hindered by the use of exclusion criteria, this might be an explanation for differences between efficacy and effectiveness.

In chapter 5 and 6 we addressed the third research question: Do participants in RCTs on MDD differ from daily practice patients in terms of sociodemographic features and socioeconomic status and do these sociodemographic and socioeconomic features influence MDD treatment outcome in daily practice? In chapter 5, we explored the reporting of several sociodemographic and socioeconomic features of participants in a large number of AETs and PETs. We summarized the sociodemographic and socioeconomic features of RCT participants. In this way, clinicians will be able to judge whether the results of RCTs are generalizable to their own patient population or individual patients. In chapter 6, we used the results of this study to compare the sociodemographic and socioeconomic characteristics (age, gender, marital status, ethnicity and employment status) of participants in AETs and PETs with those of ROM patients. We subsequently assessed the influence of sociodemographic and socioeconomic status as seen in AETs and PETs on treatment outcome in ROM. We hypothesized that daily practice patients differ from trial participants and expected that sociodemographic/economic differences between RCT participants and daily practice would influence treatment outcome. If generalizability of results from AETs and PETs would turn out to be hindered by this form of selection bias, it might be an explanation for differences between efficacy and effectiveness.

In chapter 7, we summarized and critically reviewed the main findings of our studies. We addressed the difficulties and pitfalls in comparing treatment outcome in daily practice to efficacy estimates from RCTs. We discussed the limitations of scientific research on data from Routine Outcome Monitoring. Finally, we discussed the implications of our findings for clinical practice as well as several suggestions for future research.

REFERENCE LIST

1. Kessler RC, Nelson CB, McGonagle KA, Liu J, Swartz M, Blazer DG: Co morbidity of DSM-III-R major depressive disorder in the general population: results from the US National Co morbidity Survey. *Br J Psychiatry Suppl* 1996, 17-30.
2. American Psychiatric Association: Diagnostic and statistical manual of mental disorders. 1994.
3. Double D: Blinding trials. *Br J Psychiatry* 1991, 158:573-4.
4. Carroll KM, Rounsaville BJ, Nich C: Blind man's bluff: effectiveness and significance of psychotherapy and pharmacotherapy blinding procedures in a clinical trial. *J Consult Clin Psychol* 1994, 62: 276-280.
5. Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, Ravaud P: Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS Med* 2007, 4: e61.
6. 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.
7. Leucht S: Translating research into clinical practice: critical interpretation of clinical trials in schizophrenia. *Int Clin Psychopharmacol* 2006, 21 Suppl 2:S1-10.
8. van Kampen D, de Beurs E, Andrea H: A short form of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ): the DAPP-SF. *Psychiatry Res* 2008, 160: 115-128.
9. Bijl RV, Ravelli A, van Zessen G: Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc Psychiatry Psychiatr Epidemiol* 1998, 33: 587-595.
10. Sperry L., Brill PL., Howard KI.Grisson GR.: *Treatment Outcome in Psychotherapy and Psychiatric Interventions*. New York: Brunner/Mazel Inc.; 1996.
11. Romijn GARMSF. Meer effect met depressiepreventie? Strategieën voor publieksvoorlichting, vroegherkenning en terugvalpreventie. 1-1-2008. Utrecht: Trimbos-instituut .
12. McPherson K: Do patients' preferences matter? *BMJ* 2008, 337:a2034. doi: 10.1136/bmj.a2034.: a2034.
13. Livesley WJ, Jackson DN, Schroeder M.L.: Dimensions of Personality Pathology. *Canadian Journal of Psychiatry* 1991, 557-562.
14. San Antonio Evidence based Practice Center. Agency for Healthcare Policy Research: Evidence Report on Treatment of Depression - Newer Pharmacotherapies. 1999. Washington DC, AHCPR.
15. Song F, Eastwood A, Gilbody S (Eds):Publication and related biases. In *Health Technology Assessment* 2000, 1-115.
16. National Taskforce Guideline. Multidisciplinaire Richtlijn voor diagnostiek en behandeling van volwassen cliënten met een depressie, herziene versie. 1-1-2005. Stuurgroep Richtlijnen/ Trimbos Instituut.
17. Landelijke Stuurgroep Richtlijn Ontwikkeling in de GGZ. Richtlijnherziening van de Multidisciplinaire Richtlijn Depressie. 2010. Netherlands Institute of Mental Health and Addiction (Trimbos Instituut).
18. Cuijpers P, van SA, van OP, Andersson G: Are psychological and pharmacologic interventions equally effective in the treatment of adult depressive disorders? A meta-analysis of comparative studies. *J Clin Psychiatry* 2008, 69: 1675-1685.
19. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R *et al*.: Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009, 373: 746-758.
20. Cuijpers P, Geraedts AS, van OP, Andersson G, Markowitz JC, van SA: Interpersonal psychotherapy for depression: a meta-analysis. *Am J Psychiatry* 2011, 168: 581-592.
21. Rothwell PM: External validity of randomized controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005, 365: 82-93.
22. Rittenhouse BE: The relevance of searching for effects under a clinical-trial lamppost: a key issue. *Med Decis Making* 1995, 15: 348-357.

23. Freemantle N, Mason J, Eccles M: Deriving treatment recommendations from evidence within randomized trials. The role and limitation of meta-analysis. *Int J Technol Assess Health Care* 1999, 15: 304-315.
24. Goodwin PJ, Pritchard KI, Spiegel D: The Fox guarding the clinical trial: internal vs. external validity in randomized studies. *Psychooncology* 1999, 8: 275.
25. TenHave TR, Coyne J, Salzer M, Katz I: Research to improve the quality of care for depression: alternatives to the simple randomized clinical trial. *Gen Hosp Psychiatry* 2003, 25: 115-123.
26. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R *et al.*: Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol* 2003, 3:28.
27. Leichsenring F: Randomized controlled versus naturalistic studies: a new research agenda. *Bull Menninger Clin* 2004, 68: 137-151.
28. Persaud N, Mamdani MM: External validity: the neglected dimension in evidence ranking. *J Eval Clin Pract* 2006, 12: 450-453.
29. 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.
30. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
31. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967, 6: 278-296.
32. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
33. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J: An inventory for measuring depression. *Arch Gen Psychiatry* 1961, 4:561-71.
34. Becker BJ: Synthesizing standardized mean-change measures.S. *British Journal of Mathematical and Statistical Psychology* 1988, 41: 257-278.
35. Jacobson NS, Truax P: Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991, 59: 12-19.
36. Schmitz N, Hartkamp N, Franke GH: Assessing clinically significant change: application to the SCL-90-R. *Psychol Rep* 2000, 86: 263-274.
37. 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.
38. Moleiro C, Beutler LE: Clinically significant change in psychotherapy for depressive disorders. *J Affect Disord* 2009, 115: 220-224.
39. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA *et al.*: Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials* 2004, 25: 119-142.
40. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
41. Zimmerman M, Chelminski I, Posternak MA: Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *J Nerv Ment Dis* 2004, 192: 87-94.
42. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. *J Clin Psychiatry* 1996, 57: 572-575.
43. Zetin M, Hoepner CT: Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol* 2007, 27: 295-301.
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. Howland RH, Rush AJ, Wisniewski SR, Trivedi MH, Warden D, Fava M *et al*.: Concurrent anxiety and substance use disorders among outpatients with major depression: clinical features and effect on treatment outcome. *Drug Alcohol Depend* 2009, 99: 248-260.
46. 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.
47. Souery D, Oswald P, Massat I, Bailer U, Bollen J, Demyttenaere K *et al*.: Clinical factors associated with treatment resistance in major depressive disorder: results from a European multicenter study. *J Clin Psychiatry* 2007, 68: 1062-1070.
48. Petersen T, Andreotti CF, Chelminski I, Young D, Zimmerman M: Do co morbid anxiety disorders impact treatment planning for outpatients with major depressive disorder? *Psychiatry Res* 2009, 169: 7-11.
49. Kool S, Schoevers R, de Maat S, Van R, Molenaar P, Vink A *et al*.: Efficacy of pharmacotherapy in depressed patients with and without personality disorders: a systematic review and meta-analysis. *J Affect Disord* 2005, 88: 269-278.
50. Newton-Howes G, Tyrer P, Johnson T: Personality disorder and the outcome of depression: meta-analysis of published studies. *Br J Psychiatry* 2006, 188:13-20.
51. Fournier JC, Derubeis RJ, Shelton RC, Gallop R, Amsterdam JD, Hollon SD: Antidepressant medications v. cognitive therapy in people with depression with or without personality disorder. *Br J Psychiatry* 2008, 192: 124-129.
52. Charleson ME, Horwitz RI: Applying results of randomized trials to clinical practice: impact of losses before randomisation. *BMJ* 1984, 289: 1281-1284.

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.

Chapter 3

The generalizability of antidepressant efficacy trials to routine psychiatric outpatient practice

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

Psychological Medicine 2011:41:1353-1363

ABSTRACT

Background: Generalizability of antidepressants efficacy trials (AETs) to daily practice is questioned because of their very stringent patient selection. This study aims to determine eligibility for AETs of outpatients suffering from major depression in a routine outpatient-setting and investigates influence of eligibility on treatment outcome.

Methods: Data collection (n=1653) through routine outcome monitoring by independent trained research nurses. MINIplus and DAPP-SF were used for diagnostic assessment and personality pathology screening. MADRS was used for assessment of baseline severity and treatment outcome. Eligibility was assessed by stepwise application of commonly used exclusion criteria. Influence of eligibility on treatment outcome was investigated in a subsample of the 1653 patients who had at least one follow up assessment (n=626). Eligible and non-eligible patients were compared on proportion of response (50% reduction) and remission on MADRS (MADRS≤10).

Results: 17–25% of the patients were eligible for AETs. The most common reasons for exclusion would be "not meeting minimum baseline severity" and "presence of co morbid Axis I disorder". Eligible and non-eligible patients did not differ in treatment outcome. Only "meeting the minimum baseline severity" is associated with remission.

Conclusion: The majority of "real life" outpatients is not eligible for AETs. However, the influence of eligibility on treatment outcome seems to be small. This suggests that stringent patient selection by eligibility criteria is not the major reason for lack of generalizability of AETs. Exclusion of less severely depressed patients from the analyses resulted in better treatment outcome. Milder depression is highly prevalent in daily practice and more research into treatment effectiveness in milder depression is warranted.

Key words: major depression; routine outcome monitoring; generalizability; antidepressant efficacy trials; eligibility

INTRODUCTION

During the past decades, depression treatment has shifted from an approach based on clinical expertise towards an evidence based approach using results from randomized clinical trials (RCTs) on antidepressants and/or psychotherapy [1,2]. However, for methodological and ethical reasons, antidepressant efficacy trials (AETs) will always need strict, randomized and placebo-controlled conditions, and use stringent inclusion and exclusion criteria for patient selection. In this way, internal validity is optimized. However, by optimizing internal validity, external validity (generalizability) might be compromised. Hence, the generalizability of the results from AETs to clinical practice can be questioned [3-7]. Three studies in the United States examined the eligibility of depressive patients for inclusion in AETs [8-10]. These reported that only 12–34% of these patients were eligible for AETs. However, these investigations regarded only fee-for-service settings, which may not be generalizable to the European healthcare system. In a European study, the eligibility of volunteers for AETs was also found to be limited: 34% of the patients who volunteered for an AET finally entered the trial [11]. The majority of the volunteers was excluded because of co morbid disorders. Investigators of the STAR*D trial [12] used less stringent inclusion criteria in order to obtain more generalizable results. In their study 22% of the included patients would have been eligible for AETs and had better treatment outcome than non-eligible patients [13]. However, the generalizability of STAR*D to routine clinical practice may still be limited, due to exclusion of prior non-responders to the study drugs, and the use of a minimum baseline severity. In addition, the generalizability of STAR*D to non-US health settings is unclear [14]. In the present study, we investigated in routine outpatient-care to what proportion of depressive patients the results of AETs would apply. We chose to limit the AETs to classical RCTs, since in national and international treatment guidelines [15-17], classical RCTs are considered the most robust evidence for efficacy. We do, however, expect that in the near future the results from more pragmatic trials like STAR*D and GENDEP [12,18] will influence guidelines. We applied the most frequently used inclusion and exclusion criteria for classical AETS to a large consecutive series of patients. Comprehensive data on patients' characteristics were available through the extensive Routine Outcome Monitoring (ROM) system. In addition, we investigated whether eligible patients differ from non-eligible patients in treatment outcome.

METHODS

The Dutch mental health care system and treatment steps for major depression

In the Netherlands, health insurance is obligatory for all citizens and regulated by the government. Mental health care is easily accessible and not restricted by the financial means of individual 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 recommended that patients with mild to moderate depression should be treated with psychotherapy of pharmacotherapy, based on the patient's preferences [17]. Patients with severe depression should preferably start a pharmacotherapy. Rating of severity is based on clinical judgment. Reasons to refer patients to a regional mental health provider (RMHP) are a preference of patients for psychotherapy (not provided by GPs), 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 major depression could choose between psychotherapy and antidepressants. For severe depression antidepressants were the first choice. When patients were already on antidepressants the dose was optimized or patients were offered to switch to another antidepressant or start psychotherapy.

Routine Outcome Monitoring

In 2002, the RMHP Rivierduinen (service area with 1.1 million inhabitants), in collaboration with the University Medical Hospital Leiden, implemented ROM and evidence based, stepped care protocols. In ROM, all patients referred to the RMHP for treatment of a mood, anxiety or somatoform disorder have an extensive baseline assessment. Treatment progress is then assessed at three to four monthly intervals and before starting a new treatment step. The baseline assessment comprises a standardized diagnostic interview (Mini-International Neuropsychiatric Interview Plus [19]), the collection of sociodemographic and socioeconomic data, the administration of disease specific severity-scales, and general measures of health. For a more extensive description of ROM we refer to the design paper [20].

Patients

To examine the eligibility of depressive outpatients for AETS, we included all outpatients with a DSM-IV diagnosis of a current major depressive disorder as established by the Mini-International Neuropsychiatric Interview (MINIplus) [19], who sought treatment at the RMHP Rivierduinen from January 2002 until January 2007. The MINIplus does not yield a hierarchy in primary disorder and co morbid disorders. We included all patients with a major depressive disorder, regardless of the fact whether depression was the primary diagnosis

as determined by the treating clinician or a so-called co morbid disorder. We decided to do so since primary clinical diagnosis is a concept often used in clinical practice but not well defined in literature and is depending heavily on the personal expertise of the individual clinicians. Including patients based on primary clinical diagnosis of depressive disorder only would have led to selection bias and results in a less well defined sample. Since the presence of a primary clinical diagnosis of depression might influence treatment outcome, we controlled for it in the analyses on the influence of eligibility on treatment outcome.

In order to examine the influence of eligibility to AETs on treatment outcome, we selected all patients in our sample with at least one follow up assessment in ROM (follow-up group). The treatment outcome of the first treatment step was explored in this project. We examined possible selection bias by comparing the patient characteristics of the follow-up group and the lost-to-follow-up group. We conducted an extensive chart review in the follow-up group in order to obtain information on primary clinical diagnosis and treatment modality. In order to allow comparison with classical AETs, we defined treatment outcome in the same dichotomous variables used in AETs:

1. Proportion of responders: 50% reduction of the baseline score on the Montgomery Asberg Depression Rating Scale (MADRS) [21].
2. Proportion of remitters: MADRS≤10.

Commonly used exclusion criteria for Antidepressant Efficacy Trials

In an extensive review of the literature Zimmerman and co-workers identified exclusion criteria that where consistently used in AETs published between 1994 and 1999 in the top-five Impact Ranking journals in the US [9,22]. We expanded this search by inclusion of AETs published between 1994 and 2007, not only in the aforementioned journals, but also in the remaining journals of the top-ten Impact Ranking psychiatric journals of 2005. With our expanded search, we obtained 17 additional articles on AETs [1,23-36,37,38]. No additional exclusion criteria for AETs were identified. The commonly used exclusion criteria, identified by Zimmerman and co-workers [9] are listed below, together with the operationalisations for our sample.

1. ***History of DSM-IV manic or hypomanic episodes***
   At least one (hypo) manic episode on the MINIplus.
2. ***Experiencing psychotic features during the current episode of depression***
   Diagnosis of a current depression with psychotic features on the MINIplus.
3. ***Significant risk of suicide***
   In our sample, suicidality was assessed with the corresponding item on the MADRS item 8. Patients with a score of 3 or higher were considered to meet this exclusion-criterion.

The item is a Likert-scale from 0-6:

0 Enjoys life, takes it as it is.

2 Tired of life, only transient suicidal thoughts.

4 Probably better off dead. Suicidal thoughts often occur and suicide is considered to be a possible solution. No specific plans.

6 Explicit plans to commit suicide. Active preparation of suicide.

4. Alcohol or drug abuse or dependency within the last six months

Diagnosis of current abuse or dependence on drugs or alcohol on the MINIplus.

5. Mild depression, as determined by low baseline score on the Hamilton depression-scale

The most commonly used threshold for inclusion in an efficacy trial is a minimum score of 18 (HAMD 17 items) or a minimum score of 20 (HAMD 21 items) on the Hamilton depression scale [10,39]. Because in our setting the MADRS is used to asses depression severity, an equivalent of the HAMD score was computed using three previously developed regression equations based on three trials that compared the MADRS and the HAMD17 in outpatients Mittmann et al. [40] (A); Hawley et al., [41] (B) and Zimmerman et al. (C) [42]. Since the Item Response Theory (IRT) has recently been proven to be a probably more reliable method of conversion of the MADRS into the HAMD17 as well [43,44], we also used the IRT method to compute proportions of patients not meeting the criterion of minimum baseline severity.

1. (A) MADRS = 1.23 x HAMD - 0.30 (cutoff MADRS = 21.8)
2. (B) MADRS = 1.30 x HAMD + 0.7 (cutoff MADRS = 24.1)
3. (C) MADRS = 1.43 x HAMD + 0.87 (cutoff MADRS = 26.6)

6. Presence of underlying dysthymic disorder

Diagnosis of dysthymic disorder on the MINIplus.

7. Illness duration of less than 4 weeks or more than 2 years

Duration of less than 4 weeks or more than 2 years of the current episode is an exclusion criterion for antidepressant efficacy trials. Unfortunately, in our sample no reliable information on the duration of the current episode of the major depression was available. Therefore, we could not use this exclusion criterion in our analysis.

8. Presence of co morbid non-depressive, non-substance use Axis I disorders

Diagnosis of anxiety disorder, somatoform disorder, eating disorder, or attention deficit hyperactivity disorder on the MINIplus.

9. Presence of borderline personality disorder

In our setting, the Dimensional Assessment of Personality Pathology, short Dutch version DAPP-SF [45,46] was used as a screening instrument for personality-pathology. Stringent and less stringent cut-off scores were used to identify patients with a possible personality disorder within a population suffering from mood-, anxiety-, and somatoform disorders [46]. Quartiles (low score-intermediate low-intermediate high-high score) were computed for the patients in our sample on (weighted) scores for the dimensions Emotional Dysregulation, Dissocial Behavior and Inhibition. The scores were weighted by the factor loadings derived from research on psychometrics of the DAPP-SF [46]. In our sample, patients with a cut-off of 3.7 and a "high score" on all three dimensions were considered to meet the exclusion-criterion of borderline personality according to "*stringent criteria*". Patients with a cut-off of 2.6 and a "high score" on all three dimensions were considered to meet the exclusion-criterion of borderline personality disorder according to "*less stringent criteria*".

Statistical analysis

For each exclusion criterion, we determined the percentage of patients that met the criterion. For the DAPP-SF quartiles were computed for (weighted) scores. The scores were weighted by the factor loadings derived from research on psychometrics of the DAPP-SF [20]. In our sample, there were missing values for the MADRS (n=103) and the DAPP-SF (n= 415). Comparison of complete cases and cases with missing data showed differences on many variables. Therefore it is likely that the missing data were not missing-completely-at-random (MCAR). Complete case analysis is likely to yield biased estimates [47]. Therefore, the MICE (multivariate imputation by chained equations [48]) method was used to estimate missing values for MADRS. With these imputed data, we computed the percentage of patients meeting the exclusion criteria of *Mild Depression* and *Significant Risk of Suicide*. We did not impute missing values for the DAPP-SF, as this instrument consists of dimensional components that we considered too complex to predict. If the score for the DAPP-SF was missing for a patient, we considered the patient as *not meeting* the exclusion criterion of *Presence of Borderline Personality Disorder*. Comparison of proportion of responders and remitters in the eligible and non-eligible patient-groups were performed by Chi-square tests. The influence of the exclusion criteria and "eligibility for RCTs" on treatment outcome was computed by logistic regression after MICE. Odds-ratios (OR) and their confidence intervals were computed by using the robust standard error. Statistical analyses were performed with SPSS 16.0 and STATA10.0.

RESULTS

Patients

4157 outpatients were assessed at baseline between January 2002 and January 2007. Of these patients, 1653 suffered from a current major depressive disorder according to the MINIplus. The demographic features of the 1653 patients are described in table 1.

Table 1. Demographics.

N=1653	Percentage	Mean (+SD)
Age in years		38.19 (SD 11.68)
Gender	33.3% male; 66.7% female	
Marital situation Children living at home	37.9% married/living together 13.2% divorced/widowed 25.5 single 23.4% unknown 31.9% yes 43.4% no 23.7% unknown	
Professional situation	16.5% unemployed 27.8% employed 0.7% retired 26.8% sickness/disability benefit 28.2% unknown	
Education	9.2% primary school or less 25.4% secondary school, lower level 29.5% secondary school intermediate/high level 12.1% academic or higher professional education 23.8% unknown	
Ethnicity	64.5% born in the Netherlands 4.1 % born in Morocco/Turkey 2.1% born in Suriname/Antilles 5.6% born elsewhere 23.4% unknown	
Ethnicity II	60.0% parent(s) born in the Netherlands 4.6% parent(s) born in Morocco/Turkey 2.3% parent(s) born in Suriname/Antilles 8.8% parent(s) born elsewhere 23.4% unknown	
MADRS at baseline		26.76 (SD 7.52)

SD = standard deviation

Application of commonly used exclusion criteria for AETs

Bipolarity and Psychotic Features

A total of 25 of the 1653 patients (1.5%) had at least one (hypo) manic episode (current or history). 31 patients suffered from a depression with psychotic features (1.9%). There was no overlap between these two groups. Following the approach by Zimmerman and colleagues [9,49], we excluded these 56 patients (3.4%) from further analysis. The other exclusion criteria were examined on the remaining 1597 patients.

Suicidality

Of the 1597 patients 241 patients (15.1%) would have been excluded from AETs because of suicide risk.

Alcohol or drug abuse/dependence

142 of the 1597 patients (8.9%) met the exclusion-criterion of current abuse/dependence on drugs or alcohol.

Severity of the depression at baseline

According to the first regression equation (A), 435 of the 1597 patients (27.2%) did not meet the cut-off score of 18 on the HAMD17. The second and the third regression (B, C) equations yielded identical scores, and 664 of the 1597 patients (41.6%) had a score lower than 18 on the HAMD17. The IRT yielded almost identical proportions: 38.7 % (cut-off MADRS 24) – 44.5% (cut-off MADRS 25) of the patients had a score lower than 18 on the HAMD17.

Co morbid Dysthymic Disorder

136 of the 1597 patients (8.5%) met the exclusion-criterion for a co morbid dysthymic disorder.

Other co morbid Axis I disorders

1003 of the 1597 patients (62.8 %) had co morbid diagnoses on Axis I according to the MINIplus. 730 patients (45.7%) had at least one anxiety disorder.180 patients (11.3%) had at least one somatoform disorder. Another 32 patients (2.0%) had other co morbid disorders.

Personality Pathology

31.6–61.6% of the 1597 patients in our sample may have had some form of personality pathology according to the DAPP-SF. Within this group, the estimated percentage of patients suffering from a borderline personality disorder ranges from 3 patients (0.2%, stringent criteria) to 112 patients (7.0%, less stringent criteria).

Percentage of patients eligible for Antidepressant Efficacy Trials and comparison with previous research

Finally, the sample of 1653 depressed outpatients was filtered by stepwise application of the exclusion criteria. Only 17.0%–24.5% of our patients would have been eligible for AETs. Stepwise application of the exclusion criteria is described in figure 1. Comparison of the incidence of the individual exclusion criteria in our sample with previous research [9] is described in table 2.

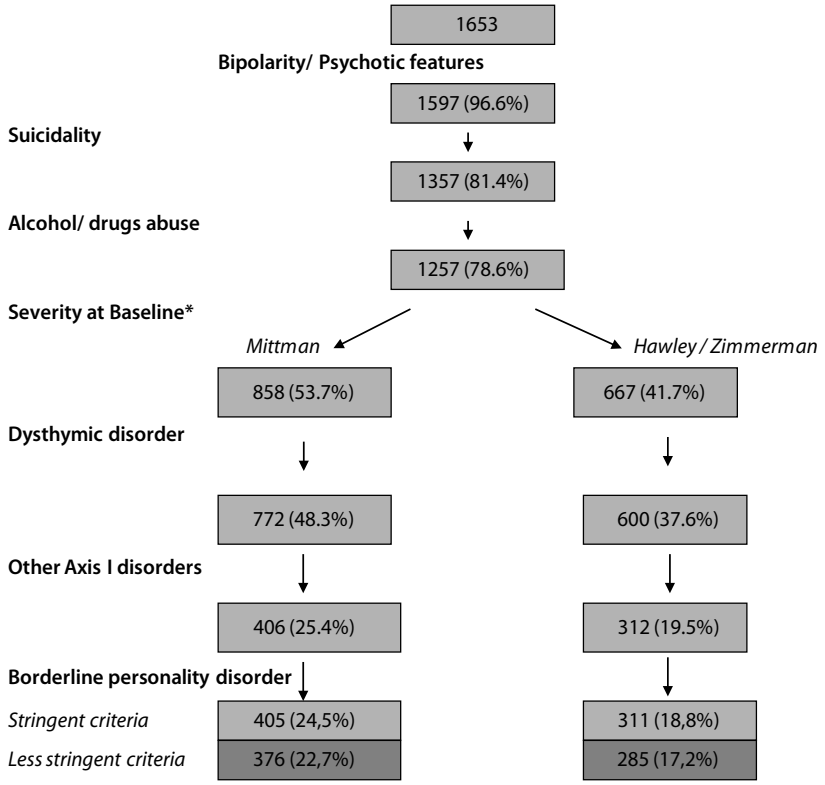

Figure 1. Stepwise Application of Commonly Used Exclusion Criteria and the Resulting Percentages of Patients Eligible for Antidepressant Efficacy Trials.

* Severity at baseline was assessed with the Montgomery Asberg Rating Scale for Depression. Equivalent Hamilton rating Scale for Depression scores (17 items version) were calculated using three previously developed regression equations.

Table 2. Comparison of incidences of exclusion criteria (%) between our sample and Zimmerman's sample [9].

	Current research Percentages of excluded patients	**Previous research Percentages of excluded patients**
Bipolarity/ psychotic features	3.4	15.3
Suicidality	15.2	19.8
Alcohol/drugs	8.6	7.8
Severity at baseline	27.2–41.6	54.3
Dysthymic disorder	8.5	8.9
Other Axis I disorders	62.8	68.3
Borderline personality pathology	0.2–7.0	11.9

Follow-up group

From the 1653 patients suffering from major depression, 46% (n=774) had a follow-up assessment. Extensive chart-review was done for those 774 patients. 148 patients had to be excluded from further follow-up analysis due to suspected bipolarity/psychotic features, admission to an inpatient-clinic during follow-up, remission on the MADRS at baseline or a time-span between baseline and follow-up assessment which we considered either to be too short or too long to provide reliable information. Finally, 626 patients were selected for follow-up analysis. In 4% of the 626 patients, information on primary clinical diagnosis for was missing. Patient selection is described in figure 2.

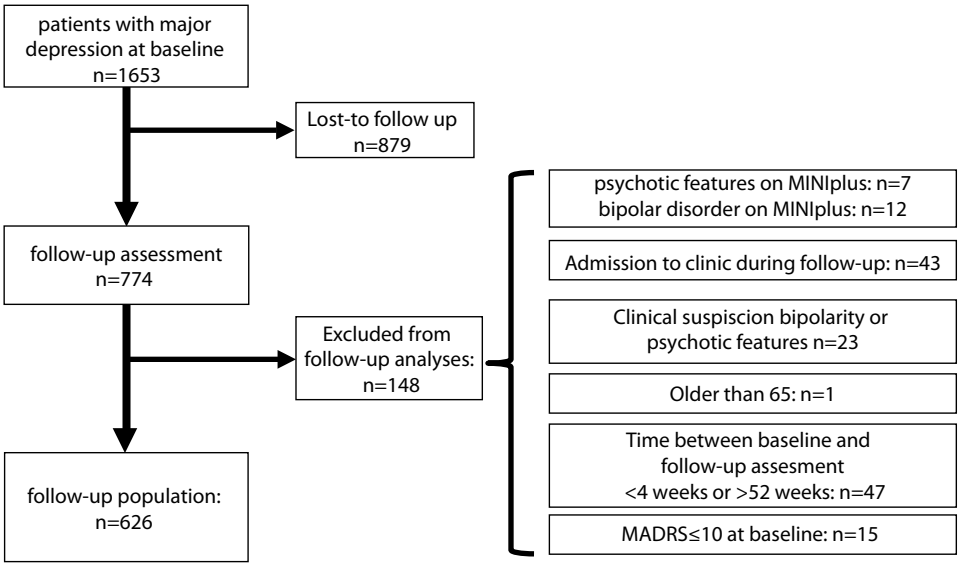

Figure 2. Selection of the follow-up group.

In chart review, we identified that 54% of the selected patients in the follow-up group received antidepressants, either as solo treatment or in combination with other treatment modalities. Five treatment modalities were identified: "antidepressants (AD)" (13%), "individual psychotherapy (IP)" (27%, mostly cognitive behavioral therapy or interpersonal therapy), "combination of antidepressants and individual psychotherapy (AD+IP)" (27%), "antidepressants and social supportive therapy (AD+SST)" (14%) and "other treatment/ insufficient information" (19%). The mean time-span between start of treatment and follow-up assessment was as follows: AD 20.8 weeks (CI 18.7–22.9); IP 20.1 weeks (CI 18.5–21.6); AD+IP 21.5 weeks (CI 20.0–23.1); AD+SST 21.6 weeks (CI 19.9-23.3); other 19.1 weeks

(CI 17.1–21.1). In 113 patients treatment was primarily started for a clinical diagnosis other than major depression, of whom 88 patients received psychotherapeutic treatment focussed specifically on anxiety/somatoform disorders.

Lost to follow-up analysis

The follow-up group and the lost-to-follow-up group did not differ on most exclusion criteria. The follow-up group only differed from the lost-to-follow-up group in larger proportions of patients with a generalized anxiety disorder (7.3% vs. 4.6%, X^2=5.08, df1, p=0.02) and depression with psychotic features (0.9% vs. 2.7%, X^2=7.42, df1, p=0.01). Based on these results, selection bias was considered to be very small.

Influence of eligibility on treatment outcome

In the follow-up group, 28% of the patients met the criteria for response and 21% of the patients met the criteria for remission. There were no significant differences in response-percentages between the patients who would have been eligible for AETs (25%) and those who were not (28%), X^2=0.26, df1, p=0.61. Remission percentages did not differ either: 16% (eligible patients) vs. 23 % (non-eligible patients), X^2=1.80, df1, p=0.18. The influence of patient features commonly used as exclusion criteria on response and remission was examined in multivariate regression models. The following variables were entered as covariates in a multivariate regression model: risk of suicide; minimum baseline severity of depression; co morbid substance dependency/abuse; co morbid dysthymia, co morbid anxiety disorder, co morbid somatoform disorder, other co morbid Axis I disorders. "Primary clinical diagnosis" and "treatment modality" were entered in the model as possible confounders. Overall, the explained variance (R-square) was very low for remission (4.1%) and response (1.4%). Only "the criterion of minimum baseline severity" contributed to remission (OR 2.0, CI 1.3–3.1). None of the exclusion criteria contributed significantly to response. The influence of "eligibility for AETs", which we defined as "not meeting any of the exclusion criteria", was investigated in a separate model and did not contribute significantly to response (OR 0.90, CI 0.5–1.8) nor remission (OR 1.0, CI 0.5–2.0).

DISCUSSION

We evaluated the eligibility for inclusion in AETs in 1653 outpatients with a major depressive disorder in a Dutch general psychiatric outpatient setting. We followed a model developed for the consistency of exclusion criteria used in AETs [9,22]. We found that the majority of patients in our sample (75%) did not meet the inclusion criteria. The most common criteria for inclusion that would not have been met were "minimum baseline severity of 18 on the Hamilton rating scale" and "no co morbid Axis I disorder". In addition, we examined the

influence of eligibility on treatment outcome. The influence of exclusion criteria on response and remission appears to be small. Only the exclusion of mild depression contributed to improvement of treatment outcome in our sample. Exclusion of less severely depressed patients from the analyses resulted in better treatment outcome. Milder depression is highly prevalent in daily practice and more research into treatment effectiveness in milder depression is warranted.

Comparison with previous research: eligibility for AETs

Our findings are in line with those of previous research [8,11,13,49]. The percentage of patients eligible for participation in AETs in our study was higher than in earlier research but similar to the latest report on eligibility in the STAR*D trial [13]. An explanation for the larger proportion of eligible patients in our sample might be the fact that the percentage of patients meeting the criterion of minimum baseline severity was larger in our sample. This might be due to the way in which the Dutch health care system is organized. First, there is no (financially) limited access to mental health care. Poor socioeconomic status has been shown to be associated with more severe pathology and co morbidity [50]. Therefore, we expected a higher percentage of patients with more severe depression in our sample. We also expected higher percentages of co morbid Axis I disorders, but the prevalence of co morbidity was similar to previous research. Another explanation for the higher percentage of patients that met the criteria for baseline severity is the role of the GP as 'gate keeper' in Dutch health care. Still, a considerable part of our patients did not meet the criteria for minimum baseline severity (27–42%).We found lower percentages of bipolarity / psychotic features and borderline personality disorders in our sample. In our RMHP those patients are often directly (preceding ROM baseline assessment) referred to specialized teams, which might explain the low prevalence in our sample.

Comparison with previous research: influence of eligibility on treatment outcome

In contrast to the recent STAR*D report [13], we found no differences in treatment outcome between eligible and non-eligible patients. Together with the marginal explained variance that we found in our model, this suggests that other patient features are more associated with treatment outcome than eligibility for AETs. Many patients, either eligible or not, would not be willing or able to participate in AETs. Participants might differ from non-participants in: sociodemographic/socioeconomic status, motivation/adherence to treatment and the interaction between clinician and patients. This might also partially explain the differences between our results and the ones found in the STAR*D report [13]. In the STAR*D trial, much effort has been undertaken to improve adherence to treatment and to motivate the participating patients and clinicians [12]. It is possible that by "controlling" for these aspects an association between eligibility and treatment outcome can be detected. Unfortunately,

the magnitude of the influence of eligibility on treatment outcome is not reported in the STAR*D report and therefore not available for comparison.
Treatment outcome in our study was less favorable than the outcomes typically found in classical AETs and also less favorable than the outcome in STAR*D. A thorough comparison and exploration of differences between the outcomes in RCTs, in more pragmatic trials like STAR*D, and in our ROM project will be important for daily practice. We are currently performing such a comparison and exploration.

Strengths

The use of Routine Outcome Monitoring in daily mental health care provided comprehensive data on a large number of patients As the only restriction for participation is sufficient language competence and ability to complete computerized or written questionnaires, the results of this type of data collection are very representative of and generalizable to 'real-life' psychiatric practice. Furthermore, we consider the fact that the Dutch health care system provides unrestricted access to mental health care as a strong quality of this research. It diminishes the possibility of selection bias even further.

Limitations

There was a considerable loss to follow-up in our study. In 22% of the lost to follow-up, patients dropped out of treatment directly after baseline assessment and in 38% of the lost to follow-up, patients stayed in treatment, but we had no information on their treatment course. The major reasons for drop-out are unclear; patients might have recovered, were perhaps unsatisfied with the offered treatment or treatment results, or had poor compliance. As 38% of the lost-to-follow-up patients remained in treatment, loss to follow-up may also have resulted from factors hampering the ROM follow-up assessments, such as administrative issues or a reduced adherence of clinicians to the ROM methodology. A large loss to follow up might be a problem in all studies with a more naturalistic design. For example, STAR*D had reached a loss-to-follow-up of 48% in step II of the study. Of the 4790 patients who were screened at baseline, 12% was not willing to participate; 3% did not meet inclusion criteria; 8% had an HAMD <14 or no data on the HAMD; and 25% left the study [12]. Although we had a considerable loss to follow-up, the follow-up group was very similar to the lost-to-follow-up group with respect to criteria for eligibility. We therefore expect the influence of the loss to follow-up on our results to be small.

The absence of information on illness duration is another limitation of this study. Although we expect not to have included patients with illness duration shorter than four weeks as most patients are seen several times by their GP before referral, it is however possible that patients were depressed for more than 2 years. This might have lead to an overestimation of the amount of eligible patients. Furthermore, a possible suboptimal diagnostic assessment of borderline personality disorder the fact that we had no information on physical health

problems (not included in Zimmerman's model on exclusion criteria, but still often used as an exclusion criterion) might also have led to an overestimation of the amount of eligible patients. On the other hand, there might be some underestimation of the eligibility in our sample, due to the fact that no data were available on patients who were too ill to complete questionnaires. Not all the patients in our sample were treated with antidepressants. A considerable proportion received other treatment (i.e. psychotherapy) for their depression. However, the percentage of eligible patients turned out to be equal in the antidepressants-group and the other-treatment-group. For comparability with former research, we used the model of Zimmerman et al. which does not take differences between AETs, like active versus placebo controlled, into account. Differences in AET architecture will probably influence eligibility, but were not investigated in the present study. Finally, to optimize comparability in treatment outcome with classical RCTs, we used the same definitions of outcome as RCTs: response and remission, determined by a cut-off score. This dichotomization of scales might lead to loss of information compared to continuous outcomes [18] .

REFERENCE LIST

1. Fava GA, Ruini C, Rafanelli C: Sequential treatment of mood and anxiety disorders. *J Clin Psychiatry* 2005, 66: 1392-1400.
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. Stewart JW, McGrath PJ, Quitkin FM: Can mildly depressed outpatients with atypical depression benefit from antidepressants? *Am J Psychiatry* 1992, 149: 615-619.
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. 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.
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. 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.
8. Zetin M, Hoepner CT: Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol* 2007, 27: 295-301.
9. Zimmerman M, Chelminski I, Posternak MA: Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *J Nerv Ment Dis* 2004, 192: 87-94.
10. Zimmerman M, Posternak MA, Chelminski I: Symptom severity and exclusion from antidepressant efficacy trials. *J Clin Psychopharmacol* 2002, 22: 610-614.
11. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. *J Clin Psychiatry* 1996, 57: 572-575.
12. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA *et al*.: Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials* 2004, 25: 119-142.
13. 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.
14. Hatcher S: The STAR*D trial: the 300 lb gorilla is in the room, but does it block all the light? *Evid Based Ment Health* 2008, 11: 97-99.
15. Karasu T, Gelenberg AJ, Merriam A, Wang P. Practice guidelines for the treatment of patients with major depressive disorder. Second Edition. 2000. American Psychiatric Association.
16. Anderson I, Pilling S, Barnes A, Bayliss L, Bird V. The NICE guideline on the treatment and management of depression in adults. Edited by National Collaborating Centre for Mental Healt, National Institute for Health and Clinical Excellence. Updated version 2010. 1-1-2009. London:The British Psychological Society & The Royal College of Psychiatrists.
17. Trimbos Institute. Richtlijnherziening van de Multidisciplinaire richtlijn Depressie (eerste revisie) . 2009.
18. Uher R, Maier W, Hauser J, Marusic A, Schmael C, Mors O *et al*.: Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. *Br J Psychiatry* 2009, 194: 252-259.
19. 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.

20. de BE, 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* 2010.
21. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
22. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
23. Benkert O, Szegedi A, Philipp M, Kohnen R, Heinrich C, Heukels A *et al*.: Mirtazapine orally disintegrating tablets versus venlafaxine extended release: a double-blind, randomized multicenter trial comparing the onset of antidepressant response in patients with major depressive disorder. *J Clin Psychopharmacol* 2006, 26: 75-78.
24. Bielski RJ, Ventura D, Chang CC: A double-blind comparison of escitalopram and venlafaxine extended release in the treatment of major depressive disorder. *J Clin Psychiatry* 2004, 65: 1190-1196.
25. DeMartinis NA, Schweizer E, Rickels K: An open-label trial of nefazodone in high co morbidity panic disorder. *J Clin Psychiatry* 1996, 57: 245-248.
26. Derubeis RJ, Hollon SD, Amsterdam JD, Shelton RC, Young PR, Salomon RM *et al*.: Cognitive therapy vs medications in the treatment of moderate to severe depression. *Arch Gen Psychiatry* 2005, 62: 409-416.
27. Detke MJ, Lu Y, Goldstein DJ, Hayes JR, Demitrack MA: Duloxetine, 60 mg once daily, for major depressive disorder: a randomized double-blind placebo-controlled trial. *J Clin Psychiatry* 2002, 63: 308-315.
28. Dinan TG: Efficacy and safety of weekly treatment with enteric-coated fluoxetine in patients with major depressive disorder. *J Clin Psychiatry* 2001, 62 Suppl 22: 48-52.
29. Golden RN, Nemeroff CB, McSorley P, Pitts CD, Dube EM: Efficacy and tolerability of controlled-release and immediate-release paroxetine in the treatment of depression. *J Clin Psychiatry* 2002, 63: 577-584.
30. Goldstein DJ, Mallinckrodt C, Lu Y, Demitrack MA: Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial. *J Clin Psychiatry* 2002, 63: 225-231.
31. Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA: Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. *J Clin Psychopharmacol* 2004, 24: 389-399.
32. Langworth S, Bodlund O, Agren H: Efficacy and tolerability of reboxetine compared with citalopram: a double-blind study in patients with major depressive disorder. *J Clin Psychopharmacol* 2006, 26: 121-127.
33. Mulder RT, Joyce PR, Frampton CM, Luty SE, Sullivan PF: Six months of treatment for depression: outcome and predictors of the course of illness. *Am J Psychiatry* 2006, 163: 95-100.
34. Shelton RC, Haman KL, Rapaport MH, Kiev A, Smith WT, Hirschfeld RM *et al*.: A randomized, double-blind, active-control study of sertraline versus venlafaxine XR in major depressive disorder. *J Clin Psychiatry* 2006, 67: 1674-1681.
35. Sir A, D'Souza RF, Uguz S, George T, Vahip S, Hopwood M *et al*.: Randomized trial of sertraline versus venlafaxine XR in major depression: efficacy and discontinuation symptoms. *J Clin Psychiatry* 2005, 66: 1312-1320.
36. Trivedi MH, Pigotti TA, Perera P, Dillingham KE, Carfagno ML, Pitts CD: Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder. *J Clin Psychiatry* 2004, 65: 1356-1364.
37. Fabre LF, Abuzzahab FS, Amin M, Claghorn JL, Mendels J, Petrie WM *et al*.: Sertraline safety and efficacy in major depression: a double-blind fixed-dose comparison with placebo. *Biol Psychiatry* 1995, 38: 592-602.
38. Stahl SM: Placebo-controlled comparison of the selective serotonin reuptake inhibitors citalopram and sertraline. *Biol Psychiatry* 2000, 48: 894-901.

39. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967, 6: 278-296.
40. Mittmann N, Mitter S, Borden EK, Herrmann N, Naranjo CA, Shear NH: Montgomery-Asberg severity gradations. *Am J Psychiatry* 1997, 154: 1320-1321.
41. Hawley CJ: Depression rating scales can be related to each other by simple equations. 1998.
42. Zimmerman M, Posternak MA, Chelminski I: Derivation of a definition of remission on the Montgomery-Asberg depression rating scale corresponding to the definition of remission on the Hamilton rating scale for depression. *J Psychiatr Res* 2004, 38: 577-582.
43. 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.
44. Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A *et al*.: Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychol Med* 2008, 38: 289-300.
45. Livesley WJ: The Dimensional Assessment of Personality Pathology (DAPP) Approach to Personality Disorder. 2006.
46. van Kampen D, de Beurs E, Andrea H: A short form of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ): the DAPP-SF. *Psychiatry Res* 2008, 160: 115-128.
47. Donders AR, van der Heijden GJ, Stijnen T, Moons KG: Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006, 59: 1087-1091.
48. Royston P: Multiple imputation of missing values: update. *Stata Journal* 2005, 5: 188-201.
49. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
50. Lesser IM, Leuchter AF, Trivedi MH, Davis LL, Wisniewski SR, Balasubramani GK *et al*.: Characteristics of insured and noninsured outpatients with depression in STAR(*)D. *Psychiatr Serv* 2005, 56: 995-1004.

Chapter 4

The generalizability of psychotherapy efficacy trials in major depressive disorder: An analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice

Rosalind van der Lem
Wouter W. de Wever
Nic J.A. van der Wee
Tineke van Veen
Pim Cuijpers
Frans G. Zitman

BMC Psychiatry. 2012, 12:192. DOI: 10.1186/1471-244X-12-192
URL: http://www.biomedcentral.com/1471-244X/12/192

ABSTRACT

Background: treatment guidelines for major depressive disorder (MDD) are based on results from randomized clinical trials, among others in psychotherapy efficacy trials. However, patients in these trials differ from routine practice patients since trials use stringent criteria for patient selection. It is unknown whether the exclusion criteria used in psychotherapy efficacy trials (PETs) influence symptom outcome in clinical practice. We first explored which exclusion criteria are used in PETs. Second, we investigated the influence of commonly used exclusion criteria on symptom outcome in routine clinical practice.

Methods: We performed an extensive literature search in PubMed, PsycInfo and additional databases for PETs for MDD. From these, we identified commonly used exclusion criteria. We investigated the influence of exclusion criteria on symptom outcome by multivariate regression models in a sample of patients suffering from MDD according to the MINIplus from a routine clinical practice setting (n=598). Data on routine clinical practice patients were gathered through Routine Outcome Monitoring.

Results: We selected 20 PETs and identified the following commonly used exclusion criteria: 'a baseline severity threshold of HAMD≤14', 'current or past abuse or dependence of alcohol and/or drugs' and "Previous use of medication or ECT". In our routine clinical practice sample of patients suffering from MDD (n=598), presence of 'current or past abuse of or dependence on alcohol and/or drugs' had no significant influence on outcome. 'Meeting a baseline severity threshold of HAMD≤14' and "Previous use of medication or ECT" were associated with better outcome, but the explained variance of the models was very small (R^2=2-11%).

Conclusions: the most consistently used exclusion criteria are not a major threat to the generalizability of results found in PETs. However, PETs do somewhat improve their results by exclusion of patients with minor depression and patients who used antidepressants prior to psychotherapy.

Key words: major depressive disorder; psychotherapy efficacy trials; exclusion criteria; generalizability; treatment outcome; symptom outcome; routine clinical practice; routine outcome monitoring

INTRODUCTION

In the development of guidelines, randomized controlled trials (RCTs) and meta-analyses thereof are considered the most reliable source of evidence. However, it is unknown to what extent the results of these RCTs are generalizable to routine clinical practice. In RCTs, much effort is put in optimising the internal validity, i.e. the possibility to determine to what extent the observed efficacy is reproducible and attributable to the investigated treatment. The internal validity of trials is improved by the use of strict criteria for patient selection. While this is very important for methodological and ethical reasons, it has been demonstrated that the use of eligibility criteria may well hamper the generalizability (external validity) of the results [1-6]. In trials of antidepressant treatment of major depression (MDD), a fairly consistent set of exclusion criteria is used [2]. Based on this set of criteria, we and others found that only 12–34% of the patients who received treatment for MDD in routine outpatient psychiatric care settings and fee-for-service private practice were eligible for participation in an antidepressant efficacy trial (AET) [1,3,7]. Some studies showed that eligible patients had a better treatment outcome than non-eligible patients in routine outpatient care [8]. In contrast, we found that only exclusion of minor depression was associated with better treatment outcome [9]. Thus, the AET exclusion criteria had a limited influence on treatment outcome.

Whereas the influence of exclusion criteria on treatment outcome is a topic in research on AETs, this is not the case for research on psychotherapy efficacy trials (PETs). To our best knowledge, only one study reported on the eligibility of "real life" patients for PETs. A total of 95% of patients with several common psychiatric disorders were eligible for at least one PET and 75% for two or more [10,11]. However, the authors did not investigate the comparability of the exclusion criteria used in the PETs. Lack of consistency in this respect may diminish the unequivocality of the results of PETs and thereby the generalizability of the results to "real life" patients.

In this paper, we present the effects of the most used exclusion criteria of PETs on eligibility of "real life" patients. First, we identified the exclusion criteria used in PETs. Subsequently, we examined the proportion of patients with unipolar depression eligible for PETs, applying the most used exclusion criteria, to a sample of "real life" patients with major depressive disorder (MDD) from the Leiden Routine Outcome Monitoring Study [12]. Finally, we investigated the influence of eligibility for PET on symptom outcome from the first treatment step, in this sample.

METHODS

Identification of exclusion criteria in PETs

In line with previous research on the consistency in the use of exclusion criteria in AETs [2], we performed a search in PubMed and PsycInfo for publications in English on PETs for adult patients suffering from MDD. Furthermore, we checked the reference lists of the included publications for relevant studies. We also consulted: http://www.psychotherapyrcts.org. This website is composed by a group of researchers from the VU University Amsterdam, the Netherlands, and contains a database of RCTs and comparative studies of the effect of psychotherapy on adult depression. We selected PETs in which outpatient treatment was investigated and in which one of the comparison groups was treated with either only individual cognitive behavioral therapy (CBT) or individual interpersonal therapy (IPT) as these two treatments are usually incorporated in treatment guidelines. For all the studies that met our inclusion criteria, we retrieved eligibility criteria from their Methods sections.

The Dutch mental health care system and treatment steps for MDD

The Dutch mental health care system is organized in a stepped-care-manner and uses treatment guidelines which are based on evidence from AETs and PETs. Patients with mood complaints visit their general practitioner (GP) first. GPs will refer patients with a first episode of a mild depression either to counseling sessions or prescribe antidepressants. The Dutch and many other guidelines recommend that patients with moderate depression should be treated with CBT or IPT or pharmacotherapy, based on the patient's preferences [13-15]. Reasons to refer patients to a regional mental health provider (RMHP) are a preference of patients for psychotherapy (only provided by psychotherapists), severity or recurrence of depression, and non-response to the GP's treatment. After baseline assessment and a clinical interview at our RMHP, patients are offered treatment steps as recommended by the guidelines. If patients are not too severely ill and have sufficient mastery of the Dutch language, they are eligible for psychotherapy when this is their preferred treatment.

Patients

Data on "real life" patients were drawn from the Leiden Routine Outcome Monitoring Study [12]. In 2002, the RMHP Rivierduinen (service area with 1.1 million inhabitants), in collaboration with the University Medical Hospital Leiden, implemented ROM and evidence based, stepped care protocols. In ROM, all patients referred to the RMHP for treatment of a mood, anxiety or somatoform disorder have an extensive baseline assessment. Treatment progress is then assessed at three to four monthly intervals and before starting a new treatment step. The baseline assessment comprises, besides a clinical interview, a standardized diagnostic interview (Mini-International Neuropsychiatric Interview Plus [16]), the collection of sociodemographic and socioeconomic data, the administration of

disease specific severity-scales, and general measures of health. All ROM instruments are administered by independent and specially trained research nurses. For a more extensive description of ROM, we refer to the design paper [12]. Patients were between 18–65 years of age, referred for treatment between January 2002 and January 2007 to the RMHP Rivierduinen, and had at least one follow-up assessment.

Since the goal of this research was to evaluate the generalizability of the results of psychotherapy trials, which generally use symptom reduction or remission on an observer rated instrument as primary outcome, we used the data collected with equivalent instruments in our ROM system. In ROM, MDD was diagnosed with the Dutch version of the MINI-Plus and depression severity was assessed with the Montgomery Asberg Depression Rating Scale (MADRS, [17]). To explore putative selection bias, we performed a lost to follow up analysis by comparison of patients only assessed at baseline with those included in our study. We investigated the eligibility and the effects of eligibility on outcome in all MDD patients referred for treatment irrespective of the treatment they received (antidepressants or psychotherapy). Since the type of treatment that patients receive might influence outcome, we adjusted for "treatment modality" in these analyses. To examine the effects of eligibility to PETs on treatment results of psychotherapy specifically, we also conducted the analyses in patients who were actually treated with CBT or IPT.

Effects of exclusion criteria on symptom outcome in daily practice

In line with previous research on exclusion criteria in AETs [1-3,18,19], we explored the influence on outcome of exclusion criteria used in >75% of the PETs. In line with the methodology of PETs, we defined outcome in our daily practice population as the extent of improvement on the MADRS (difference between baseline and post treatment), and in line with the methodology of both AETs and PETs also as proportion of responders (50% reduction of symptoms), and as proportion of remitters (MADRS ≤10) [20] after the first step treatment for MDD.

Statistical analysis

The effects of the exclusion criteria on outcome were computed by univariate and multivariate linear and logistic regression analyses. In the multivariate (adjusted) analyses on each individual exclusion criterion, the effects of the exclusion criterion on outcome were adjusted for age, gender and all the other exclusion criteria. In the analysis on all MDD patients we also adjusted for "treatment modality" (type of treatment that the patients received: antidepressants, psychotherapy or a combination of both). For the lost to follow-up analyses, independent sample t-tests and Chi-square analyses were carried out. The statistical software package SPSS 16.0 was used.

RESULTS

Identification of exclusion criteria in PETs

Our PubMed search yielded 3931 potentially relevant titles of studies. Another 203 potentially relevant studies were retrieved from reference lists of manuscripts and from the database of the VU University Amsterdam. The majority of these studies were carried out in specific subgroups, such as elderly, ethnic minorities or patients with specific somatic co morbidity (n=4085). Therefore, these studies were excluded. Another 22 manuscripts were excluded because they were duplicates between the three databases. Of the remaining 27 PETs, seven were excluded for the following reasons: in one PET the psychotherapeutic intervention appeared to include a prominent role for the spouse of the patients [21]; in another, the use of in- and exclusion criteria was mentioned but not made explicit [22]; five PETs were excluded as they used the same datasets as other studies already part of our review [23-27]. Finally, 20 PETs could be included [28-42]; [43-47]. In 18 studies (90%), individual CBT was one of the intervention arms and in 5 studies (25%) individual IPT was. In 12 PETs (60%), antidepressants (most frequently tricyclic antidepressants) were used as comparison treatment. No PETs used treatment as usual or a waiting list group as control group.

From the PETs, we identified 38 exclusion criteria, which we grouped into the following 15 categories (+ number of studies that reported the use of this criterion): 1) bipolar disorder or a history of a (hypo-)manic episode (19 studies); 2) history of schizophrenia or psychosis or psychotic features (18 studies); 3) current or past abuse of or dependence on alcohol and/or drugs (17 studies); 4) not meeting a minimum severity threshold (16 studies); 5) previous use of medication or electro convulsive therapy (ECT) (14 studies); 6) co morbid personality disorder (12 studies); 7) cognitive disorders (11 studies); 8) somatic concerns (11 studies); 9) receiving other treatment at the start of the trial (10 studies); 10) anxiety disorder as a primary diagnosis (9 studies); 11) contra indication for the use of medication (9 studies); 12) suicidality (8 studies); 13) previous psychotherapy (8 studies); 14) co morbid Axis I disorders (5 studies) and 15) crisis situation (4 studies). In line with the model of Zimmerman and colleagues on commonly used exclusion criteria in AETs [2], we planned to examine the criteria that were used in more than 75% of all PETs, which were: 1) bipolar disorder or a history of a (hypo-) manic episode (95%); 2) schizophrenia, a history of psychosis or psychotic features (90%); 3) current or past abuse of or dependence on alcohol and/or drugs (85%) and 4) not meeting a minimum severity threshold (80%; most common: cut-off score of 14 on the Hamilton Rating Scale for Depression [48] HAMD17). "Previous use of medication or ECT" was used in only 70% of the PETs, but we included this criterion in our further analyses as we hypothesized that it may have a large impact on eligibility of "real life" patients. Bipolar disorder and psychosis are considered to be different entities from MDD. Not only in PETs, but also in clinical practice, patients are treated differently if they have bipolar disorder or

a history of a (hypo-) manic episode, or a history of schizophrenia or psychosis or psychotic features. Therefore, these exclusion criteria are not likely to jeopardize the generalizability of the results of PETs for MDD to daily practice. Furthermore, we included the frequently used criteria "current or past abuse or dependence on alcohol and/or drugs" and "not meeting a minimum severity threshold" in our analyses. Co morbid substance abuse and relatively mild depression often occur in daily practice. Therefore, the frequently used exclusion criteria, "current or past abuse or dependence on alcohol and/or drugs" and "not meeting a minimum severity threshold" are likely to jeopardize the generalizability of the results of PETs to daily practice. Since in clinical practice alcohol abuse might be more common than drug abuse, we studied the effects of "current or past abuse or dependence on alcohol" and "current or past abuse or dependence on drugs" separately. Table 1 shows the exclusion criteria, the 15 summarized categories and their frequencies as identified in PETs.

Table 1. (Categories of) exclusion criteria found in psychotherapy efficacy trials.

Categorical exclusion criterion	Subtypes of exclusion criteria included in category*	Proportion of trials using the criteria
Bipolar disorder or history of (hypo-) manic episode		95%
Schizophrenia, a history of psychosis or psychotic features		90%
Current or past abuse or dependence on alcohol and/or drugs	– Alcohol abuse or dependence – Drug abuse or dependence	85%
Not meeting a minimum severity threshold		80%
Previous use of medication or ECT	– ECT less than 6 months before start of trial – History of use of a tricyclic antidepressant – Use of amitriptyline less than 3 months prior to trial – Use of imipramine less than 3 months prior to trial – Use of paroxetine less than 1 year prior to trial – Use of any antidepressant less than 2 months prior to trial – Use of any antidepressant less than 1 month prior to trial – Use of any antidepressant less than 2 weeks prior to trial – Current use of an antidepressant	70%
Co morbid personality disorder	– Borderline personality disorder – Antisocial personality disorder – Schizotypical personality disorder	60%

Categorical exclusion criterion	Subtypes of exclusion criteria included in category*	Proportion of trials using the criteria
Cognitive disorders	– Cognitive disorders in general – Organic brain syndrome – Delirium or dementia – Mental retardation	55%
Somatic concerns	– Somatic co morbidity in general – Co morbid somatisation disorder	55%
Receiving other treatment at start of trial		50%
Anxiety disorder as primary diagnosis	– Generalized anxiety disorder – Specific phobia – Obsessive-compulsive disorder – Panic disorder	45%
Contra indication for the use of medication in general		45%
Suicidal ideation		40%
Previous psychotherapy, with or without success	– History of psychotherapy – Psychotherapy less than 5 years prior to trial – Psychotherapy less than 2 years prior to trial – Psychotherapy less than 1 year prior to trial – Psychotherapy less than 2 months prior to trial – Current psychotherapy	40%
Psychiatric co morbidity in general, including eating disorders		25%
Crisis	– Need for immediate intervention – Indication for admission	20%

*If no subtypes are mentioned, the categorical exclusion criterion was reported in the same way in all trials.

Patients

Between January 2002 and January 2007, 1653 outpatients seeking treatment at RMHP Rivierduinen suffered from MDD according to the MINIplus. 774 patients (46%) had at least one follow-up assessment. Extensive chart review was done for those 774 patients. As we confined our study to patients with unipolar depression, we excluded 42 patients who were suspected to have a bipolar disorder or psychotic features. Furthermore, 132 patients had to be excluded from further follow-up analysis due to missing information on treatment, admission to an inpatient-clinic during follow-up, remission on the MADRS at baseline or a time-span between baseline and follow-up assessment which we considered either to be too short (less than four weeks) or too long (more than 52 weeks) to provide reliable information. Finally, 598 patients were selected for follow-up analysis. Of these 598 patients, 80 patients only received individual psychotherapy (CBT or IPT) for MDD; 82 patients received only antidepressants; 90 patients received psychotherapy for a co morbid disorder other than MDD or the focus of psychotherapy could not be extracted from chart review;

167 patients received a combination of psychotherapy for MDD and antidepressants; 90 patients received antidepressants and social supportive counseling; 89 patients received other forms of treatment, i.e. mood stabilizers, group therapy, training courses. Clinical and demographical characteristics of the whole sample as well as the 80 patients who received psychotherapy only are reported in table 2. In an earlier study on this sample we examined selection bias, due to loss to follow up of patients. We showed that the patients of this sample were very similar to the patients who were lost to follow up [7]. In table 2, we present the baseline features and symptom outcome in ROM patients suffering from MDD.

Table 2. Baseline features and treatment outcome in ROM patients suffering from MDD.

	All MDD patients (n=598)	**Patients who received psychotherapy only (n=80)**
Age (in years)	39.3 (SD 11.3)	36.2 (SD 10.8)
Gender (% female)	66.7% (n=399)	73.8% (n=59)
MADRS pre treatment	25.9 (SD 6.5)	24.1 (SD 6.0)
MADRS post treatment	18.2 (SD 9.4)	16.5 (SD 9.1)
Treatment outcome		
Effectsize[1]	1.16	1.28
Proportion of responders[2]	29.1%	35.0%
Proportion of remitters[3]	22.6%	27.5%
Ethnicity		
Netherlands	84.8%	76.4%
Turkey/Morocco	5.1%	5.5%
Suriname/Antilles	3.1%	4.2%
Other	7.0%	13.9%
Marital Status		
Married/cohabitating	52.8%	44.4%
Divorced/widowed	16.6%	22.2%
Single/LAT	30.0%	33.3%
Employment Status		
Employed	34.3%	34.7%
Not Employed	26.1%	34.7%
Sickness Benefit	39.0%	30.6%
Retired	0.6%	0%
Educational level		
Low	12.3%	9.7%
Intermediate low	33.1%	23.6%
Intermediate high	38.4%	40.3%
High	16.2%	26.4%

[1] Effectsize is a definition of treatment outcome often used in PETs and defined as: the extent of improvement (Δ MADRS pre- and post treatment) adjusted for the standard deviation pre treatment.
[2] Response is defined as a 50% reduction of symptoms on the MADRS.
[3] Remission is defined as MADRS≤10.

Effects of exclusion criteria on symptom outcome

As we confined our study to unipolar depression, we excluded patients with a "bipolar disorder or a history of a (hypo-) manic episode" and patients with a "history of schizophrenia or psychosis or psychotic features" from our daily practice sample. Hence, we did not explore the effects of these two frequently used exclusion criteria in PETs. We did analyze the effects of the exclusion criteria "current or past abuse or dependence on alcohol and/or drugs", "not meeting a minimum severity threshold" and "Previous use of medication or ECT" on outcome. In the literature, the baseline severity threshold (a cut-off score of 14 on the HAMD17 for PETs) is usually defined as a score on the HAMD17. In our routine clinical practice (ROM), depression severity is assessed with the MADRS. To enable comparison, we converted the scores MADRS of the ROM patients into HAMD17 scores with the equation proposed by Zimmerman [49] : MADRS = 1.43 X HAMD + 0.87. Recently, the Item Response Theory (IRT) was suggested to be a more reliable method to convert MADRS scores into HAMD17 scores. As a sensitivity analysis, we also used the IRT method [50] procedures yielded similar results for the conversion of the MADRS scores into HAMD17 scores.

Table 3 shows the proportions of patients meeting the exclusion criteria for all 598 patients with MDD, as well as for the 80 patients treated with psychotherapy. In the group of all MDD patients, the criterion "previous use of medication or ECT" had the largest effect on proportion of eligible patients. In the 80 psychotherapy patients, the criterion "not meeting baseline severity threshold" had the strongest effect.

Table 3. Exclusion criteria in ROM patients suffering from MDD.

Exclusion Criterion	All MDD patients (n=598)	Patients who received psychotherapy only (n=80)
Current or past abuse or dependence of drugs	2.3% (n=14)	5.0% (n=4)
Current or past abuse or dependence of alcohol	5.0% (n=30)	2.5% (n=2)
Not Meeting Baseline Severity Threshold	21.9% (n=131)	30.8% (n=24)
Previous use of medication or ECT (all patients received antidepressants, none of the patients received ECT prior to psychotherapy)	44.1% (n=230) Missing data: n=77	13.8% (n=11) Missing data: n=0

Table 4 shows the joint effects of the exclusion criteria on symptom outcome. In the group of all 598 depressed unipolar patients the criterion 'current or past abuse of or dependence on alcohol and/or drugs' had no significant influence. In the 80 psychotherapy patients, patients that met this criterion were too few in number for analysis of the effect. In the group

of all 598 depressed patients, patients with a baseline severity ≥14 on the HAMD17 had 7.23 points (95% CI 5.31–9.14 p<0.001) more improvement on the MADRS than patients meeting the exclusion criterion of "not meeting minimum severity threshold". The exclusion criterion "not meeting a minimum severity threshold" had no effect on the proportion of responders, but decreased the proportion that reached remission (OR 0.53, CI 0.33–0.84, p=0.01). For the subsample of psychotherapy patients, the joint analysis of exclusion criteria showed no associations with the exclusion criterion 'not meeting minimum severity threshold'.

For all 598 patients with MDD, exclusion of patients meeting the criterion "previous use of medication or ECT" was associated with a more favorable proportion of responders and remitters in the remaining sample (OR 1.53, CI 1.00–2.34, p=0.05, unadjusted). Among the 80 psychotherapy patients, those who met the criterion "previous use of medication or ECT" had 7.2 point less improvement on the MADRS than others (95% CI 1.94–13.30, p<0.01, unadjusted). However, in the joint analysis with the other exclusion criteria, the associations were no longer significant.

The explained variance (R^2) of the joint influence of the eligibility criteria respectively for all patients and psychotherapy patients was very small (adjusted for age, gender and type of treatment): 9 and 11% for the improvement on the MADRS; 2 and 7% for the proportion of patients who responded to therapy (50% reduction of symptoms); 4 and 7% for proportion of patients who reached remission (MADRS ≤10).

Table 4: Effects of the exclusion criteria on treatment outcome in ROM patients suffering from MDD.

	Definition of outcome	Influence on outcome in all patients	Influence on outcome in all patients adjusted[1]	Influence on outcome in psychotherapy patients	Influence on outcome in psychotherapy patients adjusted[1]
Current or past abuse or dependence on alcohol					
	ΔMADRS	B = 2.77 95% CI -0.85-6.39 p=0.13	B=2.09 95% CI: -1.50 -5.68 p=0.25	-	-
	Proportion of responders	OR = 2.12 95% CI 0.80-5.63 p=0.13	OR 2.36 95% CI 0.80-7.03 p=0.13	-	-
	Proportion of remitters	OR = 1.95 95% CI 0.67-5.68 p=0.22	OR 2.10 95% CI 0.61-7.27 p=0.24	-	-
Current or past abuse or dependence on drugs					
	ΔMADRS	B = - 0.17 95% CI -5.35-5.02 p=0.95	B= -1.12 95% CI: -6.04 -3.79 p=0.65	B=3.37 95% CI -1.04 – 7.78 p=0.13	B=3.50 95%CI -8.51-7.85 p=0.11
	Proportion of responders	OR = 0.73 95% CI 0.24-2.22 p=0.58	OR 0.70 95% CI 0.22-2.15 p=0.53	OR 1.72 95% CI 0.17-17.40 p=0.65	OR 1.52 95% CI 0.13-17.78 p=0.74
	Proportion of remitters	OR = 0.72 95%CI 0.22-2.34 p=0.59	OR 0.77 95% CI 0.23-2.54 p=0.66	OR 1.19 95%CI 0.12-12.08 p=0.88	OR 0.88 95% CI 0.07-10.39 p=0.92
Not meeting a minimum baseline severity threshold					
	ΔMADRS	**B = 6.39*** **95%CI 4.56-8.21 p<0.001**	**B=7.23*** **95%CI 5.31-9.14 p<0.001**	B=3.37 95% CI -1.04–7.78 p=0.13	B=3.50 95% CI -8.51-7.85 p=0.11
	Proportion of responders	OR = 1.28 95% CI 0.83-2.00 p=0.27	OR 1.53 95% CI 0.94-2.47 p=0.09	OR 0.90 95% CI 0.33-2.45 p=0.84	OR 0.87 95% CI 0.31-2.50 p=0.80
	Proportion of remitters	**OR = 0.46*** **95%CI 0.30-0.70 p<0.001**	**OR 0.53*** **95% CI 0.33-0.84 p=0.01**	OR 0.53 95% CI 0.19-1.49 p=0.23	OR 0.49 95% CI 0.17-1.45, p=0.20
Previous use of medication or ECT					
	ΔMADRS	B = 1.26 95%CI -0.42-2.95 p=0.14	B=1.19 95%CI -0.45 – 2.83 p=0.15	**B=7.62*** **95%CI 1.94-13.30 p<0.01**	B=5.49 95% CI -0.67-11.65 p=0.08
	Proportion of responders	OR = 1.47 95% CI 1.0-2.17 p=0.05	OR 1.39 95% CI 0.93-2.08 p=0.11	OR 6.75 95% CI 0.82-55.83 p=0.08	OR 5.46 95% CI 0.61-48.68 p=0.13
	Proportion of remitters	OR = 1.53 95% CI 1.00-2.34 p=0.05	OR 1.37 95% CI 0.88-2.13 p=0.16	OR 4.57 95% CI 0.55-38.01 p=0.16	OR 4.04 95% CI 0.44-37.26 p=0.22

* exclusion of patients who meet this criterion contributes significantly to treatment outcome. [1] Adjusted: adjusted for age, gender and treatment modality (only in all MDD patients) and for all other exclusion criteria in the model. B: regression coefficient: amount of additional improvement on the MADRS when patients who meet this exclusion criterion are excluded. OR: odds ratio, the chance of response or remission when patients who meet this exclusion criterion are excluded in relation to the chance of response or remission when these patients are not excluded. 95% CI: 95% confidence interval

DISCUSSION

We evaluated the criteria for patient selection in PETs in 598 outpatients with a unipolar major depressive disorder in a Dutch general psychiatric outpatient setting. We tried to follow the model developed for the consistency of exclusion criteria used in AETs [1,18]. However, we found a lack of consistency in the use of exclusion criteria in PETs. Only four criteria were used in at least 75% of the studies: "bipolar disorder or a history of a (hypo-) manic episode"; "schizophrenia, a history of psychosis or psychotic features"; "current or past abuse of or dependence on alcohol and/or drugs" and "not meeting a minimum severity threshold" (most common: cut-off score 14 on the HAMD17). The criterion "previous use of medication or ECT", was used in 70% of the studies and would lead to exclusion of the largest percentage (44.1%) of patients from our sample. For patients receiving psychotherapy only, the largest percentage (30.8%) would be excluded because of the criterion 'not meeting minimum severity'. In addition, we examined the influence of exclusion criteria for PETs on symptom outcome in our sample. The influence of exclusion criteria on improvement, response and remission was small, suggesting that the most consistently used exclusion criteria are not a major threat to the generalizability of the efficacy results found in PETs.

Comparison of exclusion criteria used in PETs to those used in AETs

To our knowledge there are no other studies on the effects of the exclusion criteria used in PETs on the generalizability to routine clinical practice. When we compared our results to those obtained in studies on the generalizability of AETs [2,18], there were some notable differences. First, PETs are less consistent in the use of exclusion criteria than AETs. The exclusion criteria "previous use of medication or ECT", "cognitive disorders" and "somatic co morbidity" were only found in PETs. Furthermore, PETs use a lower minimum severity threshold than AETs (14 versus 18 on the HAMD17) and exclude cluster B personality pathology more often (57% versus 21%). However, they less often use psychiatric co morbidity and suicide risk (resp. 24% versus 59% and 43% versus 75%) as exclusion criteria. Differences between PETs and AETs may have to do with the conduct of many AETs by pharmaceutical companies, especially for drug registration purposes. These AETs consequently have to adhere to standard exclusion criteria formulated by the authorities. Furthermore, pharmaceutical companies may want to maximize the likelihood to find an effect by selection of patients who are more severely ill. They may also minimize the risk of having their drug associated with suicide by exclusion of suicidal patients. Although not reported in PETs, this fear may also have led to patient exclusion in PETs.

Comparison with previous research on effects of exclusion criteria on symptom outcome

We found that the exclusion of patients who are "not meeting the baseline severity threshold of HAMD ≤14" is associated with a smaller proportion of patients who reach remission (OR 0.53), while in our previous research in the same sample we found a positive association between exclusion of patients with a baseline severity of HAMD≤17 (used in AETs) and probability of remission (OR 2.0) [7]**.** This finding may be explained by the fact that there were many patients in our sample who had a baseline severity between HAMD 14 and 17 (n=107, 18% of our study sample) who did not reach remission (78% of these 107 patients). We are currently investigating the characteristics of this specific group of patients with mild depressive symptomatology who seem to be at risk for a more chronic course of their depressive disorder. Furthermore, the treatment success in our sample was rather modest, yet in line with other research done in daily practice [51]. We commented on the differences between treatment outcome in daily practice and RCTs in previous research [52]. Interestingly, the within-group effect size of MDD treatment in our ROM population was relatively high compared to the modest remission and response percentages. An explanation for this discrepancy may be that we computed all symptom outcomes for ROM reported in table 2, including effect sizes, on the MADRS. However, in PETs, remission and response are often measured on the MADRS or HAMD, but effect sizes are usually computed on the BDI-II [53]. In our previous report, we investigated the effect sizes for MDD treatment on the BDI-II in our ROM population [52] and found indeed smaller effect sizes (0.85 for individual psychotherapy) than the ones based on the MADRS reported in the present study. Another explanation is that the standard deviation on the MADRS at baseline is relatively small in our ROM population, perhaps as a result of the assessment by specially trained independent research nurses.

We found that patients who used medication prior to psychotherapeutic treatment seem to benefit less from psychotherapy. Probably, these patients are non-responders or partial responders in a first treatment step for MDD and may form a more treatment resistant group. Hence, it is possible that PETs efficacy results were increased by exclusion of these patients. However, in routine clinical practice, many patients have used or are on medication before they start psychotherapy.

In line with our research on the influence of exclusion criteria of AETs on treatment outcome [7], we found an explained variance that was very small. This suggests that although many "real life" patients are not eligible for RCTs on MDD [1,3,6,7], the use of eligibility criteria might not jeopardize the generalizability of the results in "real life" settings. In previous research was found that patients who were eligible for AETs had a favorable treatment outcome [8], but the explained variance was not explored.

Most likely many other factors, besides eligibility, contribute to differences in outcome between RCTs and daily practice, like the Hawthorne effect [54], sociodemographic and

socioeconomic differences between RCT participants and "real life" patients [9] and the extent of protocol adherence of both therapist and patient, in which is highly invested in RCTs and likely not to the same extent in daily practice. We elaborated more extensively on the difference between efficacy and effectiveness in a previous report [52]. Further research on factors that contribute to differences in outcome between trials and daily practice is highly recommended.

Strengths

We used a large sample of patients with MDD from routine outpatient clinical practice (the Leiden Routine Outcome Monitoring study [12]), for which detailed data were available, enabling analysis of a subsample of patients receiving only psychotherapy. The use of ROM data provided comprehensive data that are very representative and generalizable to "real life daily practice" since there are nearly no restrictions for participation. Furthermore, we consider the fact that the Dutch healthcare system provides unrestricted access to mental healthcare as a strong quality of this research. Unrestricted access diminishes the possibility of selection bias even further.

Limitations

The large variability in which exclusion criteria are defined in PETs made loss of information unavoidable. In addition, in our patient sample, there was a considerable loss to follow-up of outcome measurement. However, the study sample follow-up group was similar to the lost-to-follow-up group for most sociodemographic and clinical features. Patients were lost to follow-up because they dropped out of treatment or, in 38% of the cases, remained in treatment without follow-up assessments. Loss to follow up is a problem in all studies with a more naturalistic design. For example, STAR*D reached a loss-to-follow-up of 48% in step II of the study [55].

In line with psychotherapy efficacy trials, we specifically chose to define outcome as symptom reduction or remission on an observer rated instrument in order to evaluate the generalizability of results from efficacy trials. For patients, other treatment goals might also be important, such as improvement of social functioning or quality of life. For therapists, other methods of defining treatment success, might be more useful such as clinically significant change [56]. Future effectiveness research, incorporating more definitions of outcome that are relevant to patients is therefore highly recommended. ROM can be a very useful methodology to support effectiveness research, and will also provide data to improve effectiveness research itself, as it enables a comparison between different types of treatment in daily practice, where one daily practice treatment can be a control treatment for the one under investigation. It will also provide data to explore the role of co morbid disorders in treatment and to improve diagnostic procedures in daily practice. Since there is a growing awareness that there is not just one type of major depressive disorder, in the future, ROM

will hopefully be helpful in the step towards personalised MDD treatment instead of "one treatment for all".

Another limitation of this study is the rather small size of the patient group receiving psychotherapy only. More patients received psychotherapy in combination with antidepressants, which in many cases were already prescribed by the referring physician. Unfortunately, the small number of patients with documented "current or past abuse or dependence of alcohol and/or drugs" in our psychotherapy sample prohibited exploration of this criterion. Finally, an extensive Routine Outcome Monitoring system including diagnostic instruments, symptom severity scales, both observer rated and self report, and generic instruments measuring quality of life and social functioning is a costly investment for psychiatric practice and criticism is often heard, especially from policy makers. However, besides the opportunities to improve the quality of treatments in daily practice and the possibilities to scientifically evaluate questions that rise from daily practice, it also might be cost-effective. Since ROM provides information on treatment progress, it might enable the clinician to move to a next treatment step in case of stagnation in an earlier stage. Since ROM is relatively young, research in the field of its cost-effectiveness has, to our knowledge, not been carried out yet. It is, however, highly recommended.

CONCLUSIONS

We found that patient selection in psychotherapy trials in MDD lacks consistency. A consistent set of exclusion criteria is recommended in order to facilitate comparison between trials and especially for daily practice to evaluate the generalizability of their results. We also found that the most consistently used exclusion criteria are not a major threat to the generalizability of results found in PETs. However, PETs do somewhat improve their results by exclusion of patients with minor depression and patients who used antidepressants prior to psychotherapy.

REFERENCE LIST

1. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
2. Zimmerman M, Chelminski I, Posternak MA: Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *J Nerv Ment Dis* 2004, 192: 87-94.
3. Zetin M, Hoepner CT: Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol* 2007, 27: 295-301.
4. 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.
5. Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *Am J Psychiatry* 1999, 156: 5-10.
6. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. *J Clin Psychiatry* 1996, 57: 572-575.
7. van der Lem R, van der Wee NJ, van VT, Zitman FG: The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychol Med* 2011, 41: 1353-1363.
8. 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.
9. Lem Rvd, Stamsnieder P, Wee Nvd, Veen Tv, Zitman FG (Eds):Sociodemographic features in randomized controlled trials for major depression: generalizability and individualization. In *Int J Person Cent Medicine* 2011, 1: 268-278.
10. Stirman SW, Derubeis RJ, Crits-Christoph P, Rothman A: Can the randomized controlled trial literature generalize to nonrandomized patients? *J Consult Clin Psychol* 2005, 73: 127-135.
11. Stirman SW, Derubeis RJ, Crits-Christoph P, Brody PE: Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *J Consult Clin Psychol* 2003, 71: 963-972.
12. 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.
13. Karasu T, Gelenberg AJ, Merriam A, Wang P. Practice guidelines for the treatment of patients with major depressive disorder Second Edition. 2000. American Psychiatric Association.
14. Anderson I, Pilling S, Barnes A, Bayliss L, Bird V. The NICE guideline on the treatment and management of depression in adults. Edited by National Collaborating Centre for Mental Healt, National Institute for Health and Clinical Excellence. Updated version 2010. 1-1-2009. London:The British Psychological Society & The Royal College of Psychiatrists.
15. National Taskforce Guideline. Multidisciplinary guidelines for diagnostics and treatment of adult patients with major depressive disorder, revised version. 1-1-2005. Stuurgroep Richtlijnen/ Trimbos Institute, the Netherlands.
16. 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.
17. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
18. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
19. Lem Rvd, Wee Nvd, Veen Tv, Zitman FG: The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychological Medicine* 2010.

20. Zimmerman M, Chelminski I, Posternak M: A review of studies of the Montgomery-Asberg Depression Rating Scale in controls: implications for the definition of remission in treatment studies of depression. *Int Clin Psychopharmacol* 2004, 19: 1-7.
21. McLean PD, Hakstian AR: Clinical depression: comparative efficacy of outpatient treatments. *J Consult Clin Psychol* 1979, 47: 818-836.
22. Gardner P, Oei TP: Depression and self-esteem: an investigation that used behavioral and cognitive approaches to the treatment of clinically depressed clients. *J Clin Psychol* 1981, 37: 128-135.
23. Garvey MJ, Hollon SD, Derubeis RJ: Do depressed patients with higher pretreatment stress levels respond better to cognitive therapy than imipramine? *J Affect Disord* 1994, 32: 45-50.
24. Kovacs M, Rush AJ, Beck AT, Hollon SD: Depressed outpatients treated with cognitive therapy or pharmacotherapy. A one-year follow-up. *Arch Gen Psychiatry* 1981, 38: 33-39.
25. Simons AD, Garfield SL, Murphy GE: The process of change in cognitive therapy and pharmacotherapy for depression. Changes in mood and cognition. *Arch Gen Psychiatry* 1984, 41: 45-51.
26. Sotsky SM, Glass DR, Shea MT, Pilkonis PA, Collins JF, Elkin I *et al*.: Patient predictors of response to psychotherapy and pharmacotherapy: findings in the NIMH Treatment of Depression Collaborative Research Program. *Am J Psychiatry* 1991, 148: 997-1008.
27. Weissman MM, Prusoff BA, DiMascio A, Neu C, Goklaney M, Klerman GL: The efficacy of drugs and psychotherapy in the treatment of acute depressive episodes. *Am J Psychiatry* 1979, 136: 555-558.
28. Beck AT, Hollon SD, Young JE, Bedrosian RC, Budenz D: Treatment of depression with cognitive therapy and amitriptyline. *Arch Gen Psychiatry* 1985, 42: 142-148.
29. Blackburn IM, Bishop S, Glen AI, Whalley LJ, Christie JE: The efficacy of cognitive therapy in depression: a treatment trial using cognitive therapy and pharmacotherapy, each alone and in combination. *Br J Psychiatry* 1981, 139: 181-189.
30. Blom MB, Jonker K, Dusseldorp E, Spinhoven P, Hoencamp E, Haffmans J *et al*.: Combination treatment for acute depression is superior only when psychotherapy is added to medication. *Psychother Psychosom* 2007, 76: 289-297.
31. Derubeis RJ, Hollon SD, Amsterdam JD, Shelton RC, Young PR, Salomon RM *et al*.: Cognitive therapy vs medications in the treatment of moderate to severe depression. *Arch Gen Psychiatry* 2005, 62: 409-416.
32. DiMascio A, Weissman MM, Prusoff BA, Neu C, Zwilling M, Klerman GL: Differential symptom reduction by drugs and psychotherapy in acute depression. *Arch Gen Psychiatry* 1979, 36: 1450-1456.
33. Dimidjian S, Hollon SD, Dobson KS, Schmaling KB, Kohlenberg RJ, Addis ME *et al*.: Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *J Consult Clin Psychol* 2006, 74: 658-670.
34. Elkin I, Shea MT, Watkins JT, Imber SD, Sotsky SM, Collins JF *et al*.: National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Arch Gen Psychiatry* 1989, 46: 971-982.
35. Hollon SD, Derubeis RJ, Evans MD, Wiemer MJ, Garvey MJ, Grove WM *et al*.: Cognitive therapy and pharmacotherapy for depression. Singly and in combination. *Arch Gen Psychiatry* 1992, 49: 774-781.
36. Luty SE, Carter JD, McKenzie JM, Rae AM, Frampton CM, Mulder RT *et al*.: Randomized controlled trial of interpersonal psychotherapy and cognitive-behavioural therapy for depression. *Br J Psychiatry* 2007, 190: 496-502.
37. McBride C, Atkinson L, Quilty LC, Bagby RM: Attachment as moderator of treatment outcome in major depression: a randomized control trial of interpersonal psychotherapy versus cognitive behavior therapy. *J Consult Clin Psychol* 2006, 74: 1041-1054.
38. Murphy GE, Simons AD, Wetzel RD, Lustman PJ: Cognitive therapy and pharmacotherapy. Singly and together in the treatment of depression. *Arch Gen Psychiatry* 1984, 41: 33-41.
39. Murphy GE, Carny RM, Knesevich MA, Wetzel RD, Whitworth P (Eds):Cognitive behavior therapy, relaxation training and tricyclic antidepressant medication in the treatment of depression. In *Psychol Rep* 1995, 403-420.

40. Rush AJ, Beck AT, Kovacs M, Hollon SD (Eds):Comparative efficacy of cognitive therapy and pharmocotherapy in the treatment of depressed outpatients. In *Cognitive Therapy and Research* 1977, 17-37.
41. Strauman TJ, Vieth AZ, Merrill KA, Kolden GG, Woods TE, Klein MH *et al*.: Self-system therapy as an intervention for self-regulatory dysfunction in depression: a randomized comparison with cognitive therapy. *J Consult Clin Psychol* 2006, 74: 367-376.
42. Teri L, Lewinsohn PM (Eds):Individual and group treatment of unipolar depression: comparison of treatment outcome and identification of predictors of succesful treatment outcome. *Behav Ther* 1986, 215-228.
43. Watson JC, Gordon LB, Stermac L, Kalogerakos F, Steckley P: Comparing the effectiveness of process-experiential with cognitive-behavioral psychotherapy in the treatment of depression. *J Consult Clin Psychol* 2003, 71: 773-781.
44. Wilson PH: Combined pharmacological and behavioural treatment of depression. *Behav Res Ther* 1982, 20: 173-184.
45. Wilson PH, Goldin JC, Charbonneaupowis M (Eds):Comparative efficacy of behavioral and cognitive treatments of depression. *Cognitive Therapy and Research* 1983, 111-124.
46. Wright JH, Wright AS, Albano AM, Basco MR, Goldsmith LJ, Raffield T *et al*.: Computer-assisted cognitive therapy for depression: maintaining efficacy while reducing therapist time. *Am J Psychiatry* 2005, 162: 1158-1164.
47. Zettle RD, Haflich JL, Reynolds RA: Responsivity to cognitive therapy as a function of treatment format and client personality dimensions. *J Clin Psychol* 1992, 48: 787-797.
48. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967, 6: 278-296.
49. Zimmerman M, Posternak MA, Chelminski I: Derivation of a definition of remission on the Montgomery-Asberg depression rating scale corresponding to the definition of remission on the Hamilton rating scale for depression. *J Psychiatr Res* 2004, 38: 577-582.
50. 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.
51. 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.
52. van der Lem R, van der Wee NJ, van VT, Zitman FG: Efficacy versus effectiveness: a direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice. *Psychother Psychosom* 2012, 81: 226-234.
53. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J: An inventory for measuring depression. *Arch Gen Psychiatry* 1961, 4: 561-571.
54. 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.
55. 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.
56. Jacobson NS, Truax P: Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991, 59: 12-19.

Chapter 5

Sociodemographic features of participants in randomized controlled trials for major depression: Generalizability and individualization

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

International Journal of Person Centred Medicine 2011; 1:268-278

ABSTRACT

Rationale, Aims and Objectives: It is important for clinicians to know to what extent the results of randomized controlled trials (RCTs) are generalizable to their psychiatric practice, since RCTs are considered to be the most reliable source of evidence for treatment guideline development. Furthermore, it is important to know whether results from individual randomized controlled trials (RCTs) can be directly compared to each other. Sociodemographic and socioeconomic (SES) features influence treatment outcome in major depressive disorder (MDD). Differences in (reporting of) SES features of participants in RCTs will hamper comparison and jeopardize the external validity (generalizability) of their results. We explored the reporting of SES features in RCTs for depression.

Methods: We selected 45 antidepressant efficacy trials (AETs) and 19 psychotherapy efficacy trials (PETs). We listed the reported sociodemographic and -economic features.

Results: Reporting on SES features was very diverse and often limited. Especially important SES features like educational level, socioeconomic status and income were reported insufficiently. The mean age of RCT participants in MDD trials was 41 years. Participants are predominantly female (62%) and white (89%). Of the participants 61% were employed and 45% of the participants were married/cohabitating.

Conclusions: Standardisation of reporting on sociodemographic and socioeconomic status is needed to adequately judge the generalizability of RCTs to daily practice and to facilitate comparisons within the body of RCTs.

INTRODUCTION

Major depression is one of the most common psychiatric disorders, affecting about 121 million people worldwide [1]. Improvement of the quality of depression treatment would be beneficial to many people [2-4]. During the past decades, the selection of treatment for patients suffering from depression has shifted from an approach based on clinical expertise towards evidence based medicine. This has resulted in guidelines based on results from randomized controlled trials (RCTs) of antidepressants and/or psychotherapy [5,6]. There are long standing concerns regarding the generalizability of the results from the strictly controlled RCTs to the treatment of patients in "real world" clinical practice [7-12]. Patients in routine clinical practice have been shown to differ from patients included in RCTs on a number of clinical features, like the severity of symptoms or the presence of co morbidity or suicidality. These clinical differences between RCT participants and daily practice patients are mainly caused by selection bias due to the use of eligibility criteria [13-17]. However, beyond the use of clinical eligibility criteria, there are other forms of (probably unintended) selection bias which might jeopardize the external validity (i.e. generalizability) of RCTs. Patients may be eligible, but still not willing to participate in RCTs for several reasons, for instance a preference for a treatment modality. Furthermore, due to recruitment and inclusion procedures, participants in RCTs might also differ importantly from "real life" patients with respect to sociodemographic and socioeconomic background [7].

Previous research in both general medicine and psychiatry has shown that socio-demographic and socioeconomic features influence the outcome of treatment. Lower socioeconomic status and increased age were associated with poorer treatment outcome and mortality in several medical conditions [18,19]. In psychiatry, several studies on the influence of age and gender on the outcome of antidepressant treatment showed a negative association with increased age and the male gender [20-26]. In three studies increased age was not associated with poorer treatment outcome of psychotherapy for depression [27-29], and in one study, male gender was associated with better treatment outcome in psychotherapy for depression [28]. In pharmacotherapy, being married and a better socioeconomic or employment status predicted better outcomes. In psychotherapy, employment had no influence [25-28,30-33]. Remarkably, level of education was predictive for outcome neither in pharmacotherapy nor in psychotherapy [20,23,27,33-37]. Furthermore, patients with different ethnic backgrounds seem to benefit equally from pharmacotherapy and psychotherapy, yet in certain ethnic minorities treatment adherence was found to be significantly worse [38-43].

As sociodemographic and socioeconomic features (SES features) may influence treatment results, clinicians should be able to compare their "real life" patients with the participants of the trials in order to assess the generalizability of the results of the trials to their own population. Therefore, the quality of the reporting of SES features in RCTs is

of importance. For this paper we reviewed the reporting of SES features in RCTs on major depression.

METHODS

Literature Review

Inclusion: We included peer reviewed publications of RCTs, published through 2007 in outpatients with a unipolar, non-psychotic depression according to DSM-III-R or DSM-IV (major depressive disorder MDD). Because we aimed to review the reporting of sociodemographic and socioeconomic features in RCTs usually selected for the development of guidelines for routine treatment, we excluded trials which a priori included only participants from specified subgroups like elderly or a specific ethnic minority. For the same reason, we also excluded augmentation trials, trials that focused on refractory depression, or trials limited to patients with a particular co morbid condition such as alcoholism, anxiety disorder, or medical illness. Furthermore, it was essential that the publication provided baseline information on sociodemographic and/or socioeconomic features. When there were several publications from the same trial, we included the report that provided the most detailed information on sociodemographic and/or socioeconomic features. When the reports on a trial provided the same information, we included the first report. We included trials written in English, since international guidelines for treatment of MDD are predominantly based on English literature.

Psychotherapy: We performed a Medline search for RCTs investigating psychotherapy (cognitive behavioral therapy and interpersonal therapy) for adult patients suffering from MDD. Furthermore, we performed an additional search in PsycInfo and checked the reference lists of included trials for other relevant studies as well as the database http://www.psychotherapyrcts.org. This website contains a database of RCTs and comparative studies examining the effect of psychotherapy on adult depression, collected by a group of researchers from the VU University in Amsterdam, the Netherlands, and Linköping University in Sweden. We selected the psychotherapy efficacy trials (PETs) in which outpatient treatment was investigated and in which either only individual cognitive behavioral therapy (CBT) or individual interpersonal therapy (IPT) was the intervention or control group, as these two treatments are usually incorporated in treatment guidelines.

Pharmacotherapy: Because of the large number of published antidepressant efficacy trials (AETs), we restricted our search to AETs published in journals from the top ten Impact Ranking psychiatric journals of 2005. By including only high impact factor journals, we expected to have a sample of trials with the most systematic manner of reporting SES features. The journals

were Archives of General Psychiatry; Molecular Psychiatry; American Journal of Psychiatry, Biological Psychiatry; Neuropsychopharmacology; Journal of Psychopharmacology; Journal of Clinical Psychiatry; Psychotherapy/Psychosomatics; the British Journal of Psychiatry and Sleep. We added Psychopharmacology Bulletin to our selection of journals, since AETs from this journal are frequently cited in literature on antidepressants. We excluded trials with experimental medication such as dexamethason or valproate.

Sociodemographic and socioeconomic features

For the included RCTs, we explored the sociodemographic and socioeconomic features of the intent-to-treat samples. If intent-to-treat data were missing we used the data of the completers. We determined the most frequently described features and their operationalisation. If the operationalisation of the sociodemographic and socioeconomic features in a study was not well defined, we tried to contact the authors for further information. We converted the reported SES features into dichotomous or trichotomous variables.

Statistics

Descriptive summary statistics (means, frequencies, percentages) were used to describe the baseline sociodemographic and socioeconomic features of the RCT patients. These procedures were performed in SPSS 16.0. As standard deviations for continuous variables (age) were often missing in trials, we corrected for sample size by dividing the sum of all "mean age x number of patients in a trial" by the total number of patients of all trials.

RESULTS

Review of sociodemographic and socioeconomic features used in RCTs

Based on our criteria and search strategy, we included 64 published RCTs; 45 AETs and 19 PETs. We found no PETs published after 2007 meeting our inclusion criteria, and therefore also limited the inclusion of AETs to those published before 2008. Table 1 shows a list of the included trials. The total number of patients who participated in these trials is 9694; 8838 patients in the AETs group and 856 patients in the PETs group. Table 2 provides an overview of the eight most frequently described sociodemographic and socioeconomic features that were used in the 64 studies. Remarkably, only three features were reported in at least half of the included trials: mean age (n=62, 96.9%), gender (n=63, 98.4%) and race or ethnicity (n=41, 64.1%). The operationalisation of sociodemographic and socioeconomic status, which varied greatly among the studies for some features, will be discussed below.

Table 1. List of included trials.

	Title of trial	First Author	Year of Publication	Journal
1	Comparative efficacy of CT and pharmacotherapy in the treatment of depressed outpatients	Rush	1977	Cognitive therapy and research
2	Differential symptom reduction by drugs and psychotherapy in acute depression	Dimascio	1979	Archives of General Psychiatry
3	The efficacy of CT in depression: a treatment using CT and pharmacotherapy, each alone and in combination	Blackburn	1981	British Journal of Psychiatry
4	Group versus individual cognitive therapy: a pilot study	Rush*	1981	Cognitive therapy and research
5	Comparative efficacy of behavioral and cognitive treatments of depression	Wilson	1983	Cognitive therapy and research
6	Cognitive therapy and pharmacotherapy: singly and together in the treatment of depression	Murphy	1984	Archives of General Psychiatry
7	Treatment of depression with cognitive therapy and amitriptyline	Beck	1985	Archives of General Psychiatry
8	Individual and group treatment of unipolar depression: comparison of treatment outcome and identification of predictors of successful treatment outcome	Teri	1986	Behavior Therapy
9	NIMH treatment of Depression Collaborative Research Program: General effectiveness of treatments	Elkin	1989	Archives of General Psychiatry
10	Cognitive therapy and pharmacotherapy for depression: singly and in combination	Hollon	1992	Archives of General Psychiatry
11	Responsivity to cognitive therapy as a function of treatment format and client personality dimensions	Zettle	1992	Journal of Clinical Psychology
12	A comparison of venlafaxine, trazodone and placebo in major depression	Cunningham	1994	Journal of Clinical Psychopharmacology
13	Dothiepin versus doxepin in major depression: results of a multicenter, placebo-controlled trial	Ferguson	1994	Journal of Clinical Psychiatry

	Title of trial	First Author	Year of Publication	Journal
14	A double-blind comparison of nefazodone, imipramine, and placebo in major depression	Fontaine	1994	Journal of Clinical Psychiatry
15	Comparison of venlafaxine and imipramine in the acute treatment of major depression in outpatients	Schweizer	1994	Journal of Clinical Psychiatry
16	Is baseline agitation a relative contraindication for a selective serotonin reuptake inhibitor: A comparative trial of fluoxetine versus imipramine	Tollefson	1994	Journal of Clinical Psychopharmacology
17	Comparison of bupropion and trazodone for the treatment of major depression	Weisler	1994	Journal of Clinical Psychopharmacology
18	A double-blind multicenter trial comparing sertraline and fluoxetine in outpatients with major depression	Bennie	1995	Journal of Clinical Psychiatry
19	A Double-blind comparison of org 3770, amitriptyline, and placebo in major depression	Bremner	1995	Journal of Clinical Psychiatry
20	Sertraline safety and efficacy in major depression: a double-blind fixed-dose comparison with placebo	Fabre	1995	Biological Psychiatry
21	Cognitive behavior therapy, relaxation training, and tricyclic antidepressant medication in the treatment of depression	Murphy	1995	Psychological reports
22	A multicenter double-blind comparison of nefazodone and paroxetine in the treatment of outpatients with moderate-to-severe depression	Baldwin	1996	Journal of Clinical Psychiatry
23	Fluvoxamine maleate in the treatment of depression: A single-center, double-blind, placebo-controlled comparison with imipramine in outpatients	Claghorn	1996	Journal of Clinical Psychopharmacology
24	Responders to antidepressant drug treatment: a study comparing nefazodone, imipramine, and placebo in patients with major depression	Cohn	1996	Journal of Clinical Psychiatry
25	An open-label trial of nefazodone in high co morbidity panic disorder	DeMartinis	1996	Journal of Clinical Psychiatry
26	Nefazodone versus sertraline in outpatients with major depression: focus on efficacy, tolerability, and effects on sexual function and satisfaction	Feiger	1996	Journal of Clinical Psychiatry

	Title of trial	First Author	Year of Publication	Journal
27	A double-blind comparison of gepirone extended release, imipramine, and placebo in the treatment of outpatient major depression	Feiger	1996	Psychopharmacology Bulletin
28	A comparison of fluvoxamine and fluoxetine in the treatment of major depression	Rapaport	1996	Journal of Clinical Psychopharmacology
29	Zalospirone in major depression: A placebo-controlled multicenter study	Rickels	1996	Journal of Clinical Psychopharmacology
30	A double-blind trial of low- and high-dose ranges of gepirone-ER compared with placebo in the treatment of depressed outpatients	Wilcox	1996	Psychopharmacology Bulletin
31	Double-blind comparison of bupropion sustained release and sertraline in depressed outpatients	Kavoussi	1997	Journal of Clinical Psychiatry
32	A double-blind comparison of fluvoxamine and paroxetine in the treatment of depressed outpatients	Kiev	1997	Journal of Clinical Psychiatry
33	A double-blind, placebo-controlled study comparing the effects of sertraline versus amitriptyline in the treatment of major depression	Lydiard	1997	Journal of Clinical Psychiatry
34	Factors that influence the outcome of placebo-controlled antidepressant clinical trials	Niklson	1997	Psychopharmacology Bulletin
35	Desipramine versus phenelzine in recurrent unipolar depression: Clinical characteristics and treatment response	Swann	1997	Journal of Clinical Psychopharmacology
36	Efficacy and tolerability of once-daily venlafaxine extended release (XR) in outpatients with major depression	Thase	1997	Journal of Clinical Psychiatry
37	Once- versus twice- daily venlafaxine therapy in major depression: a randomized, double-blind study	Amsterdam	1998	Journal of Clinical Psychiatry
38	A double-blind, randomized trial of sertraline and imipramine	Keller	1998	Journal of Clinical Psychiatry
39	A Canadian multicenter study of three fixed doses of controlled-release ipsapirone in outpatients with moderate to severe major depression	Lapierre	1998	Journal of Clinical Psychopharmacology
40	Factors that influence the outcome of placebo-controlled antidepressant clinical trials	Rudolph	1998	Psychopharmacology Bulletin

	Title of trial	First Author	Year of Publication	Journal
41	Randomized, double-blind comparison of venlafaxine and fluoxetine in outpatients with major depression	Silva	1998	Journal of Clinical Psychiatry
42	Mirtazapine: Efficacy and tolerability in comparison with fluoxetine in patients with moderate to severe major depressive disorder	Wheatley	1998	Journal of Clinical Psychiatry
43	Multicenter, placebo-controlled, fixed-dose study of citalopram in moderate-to-severe depression	Feighner	1999	Journal of Clinical Psychiatry
44	Treatment of atypical depression with cognitive therapy or phenelzine	Jarrett	1999	Archives of General Psychiatry
45	Mirtazapine compared with paroxetine in major depression	Benkert	2000	Journal of Clinical Psychiatry
46	Randomized, double-blind comparison of venlafaxine and sertraline in outpatients with major depressive disorder	Mehtonen	2000	Journal of Clinical Psychiatry
47	Placebo-controlled comparison of the selective serotonin reuptake inhibitors citalopram and sertraline	Stahl	2000	Biological Psychiatry
48	Efficacy and response time to sertraline versus fluoxetine in the treatment of unipolar major depressive disorder	Suri	2000	Journal of Clinical Psychiatry
49	Duloxetine, 60mg once daily, for major depressive disorder: a randomized double-blind placebo-controlled trial	Detke	2002	Journal of Clinical Psychiatry
50	Efficacy and tolerability of controlled-release and immediate-release paroxetine in the treatment of depression	Golden	2002	Journal of Clinical Psychiatry
51	Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial	Goldstein	2002	Journal of Clinical Psychiatry
52	Outcomes of patients completing and not completing cognitive therapy for depression.	Cahill	2003	British Journal of Clinical Psychology
53	Comparing Effectiveness of Process Experiential with CBT in the treatment of depression	Watson	2003	Journal of Consulting and Clinical Psychology
54	A double-blind comparison of escitalopram and venlafaxine extended release in the treatment of major depressive disorder	Bielski	2004	Journal of Clinical Psychiatry

	Title of trial	First Author	Year of Publication	Journal
55	Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine	Goldstein	2004	Journal of Clinical Psychopharmacology
56	Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder	Trivedi	2004	Journal of Clinical Psychiatry
57	Cognitive therapy vs medications in the treatment of moderate to severe depression	Derubeis	2005	Archives of General Psychiatry
58	Randomized trial of sertraline versus venlafaxine XR in major depression: efficacy and discontinuation symptoms	Sir	2005	Journal of Clinical Psychiatry
59	Efficacy and tolerability of reboxetine compared with citalopram: a double-blind study in patients with major depressive	Langworth	2006	Journal of Clinical Psychopharmacology
60	A randomized, double-blind, active-control study of sertraline versus venlafaxine XR in major depressive disorder	Shelton	2006	Journal of Clinical Psychiatry
61	Self-system therapy as an intervention for self-regulatory dysfunction in depression: a randomized comparison with cognitive therapy	Strauman	2006	Journal of Consulting and Clinical Psychology
62	Attachment as moderator of treatment outcome in major depression: a RCT of IPT versus CBT	McBride	2006	Journal of Consulting and Clinical Psychology
63	Combination treatment for acute depression is superior only when psychotherapy is added to medication	Blom	2007	Psychotherapy and Psychosomatics
64	RCT of interpersonal psychotherapy and cognitive-behavioural therapy for depression	Luty	2007	British Journal of Psychiatry

Table 2. Reporting of sociodemographic/socioeconomic features in RCTs.

Sociodemographic/ socioeconomic characteristic	Number of trials reporting on the feature (%)
Age (mean)	62 (96.9%)
Gender	63 (98.4%)
Race/ethnicity	41 (64.1%)
Marital status	23 (35.9%)
Employment status	12 (18.8%)
Education	17 (26.6%)
Income	3 (4.7%)
SES	3 (4.7%)

Age

Sixty-two (97%) trials reported a mean age for their study population. Of the two trials who did not report a mean age, one trial divided the population in age categories (<30, 30–39, >39 years of age). The mean age of the participants in RCTs was 41 years. The AET participants had a mean age of 41 years, the PETs participants of 37 years.

Gender

There were 63 trials (98%) that described the distribution of the population by gender. Patients were predominantly female (62% woman versus 38% man). In AETs 61% of the patients were women. In the PETs 72% of the participants were female.

Race and ethnicity

There were 41 trials (64%) that reported race or ethnicity of the study population. Of these trials, two only gave a short description of race, for example: predominantly Caucasian. The other 39 studies used 16 different ways to define race/ethnicity. The most frequently used definition of race was white/non-white. Seventeen of the 39 trials used this definition (44%). Furthermore, the following descriptions were used: European; (non) Caucasian; Hispanic or Latino; African-descent or African American or Black; Asian or Oriental; Middle Eastern; Other Ethnicity. We converted the reported information on race or ethnicity into the dichotomous variable white/non-white. We considered Hispanic as "white", since two out of three authors of the RCTs, who we contacted, responded that they had considered Hispanic as "white". Latino, European and Caucasian are also considered to be "white" [44-46]. For this analysis, we considered "non-Caucasian, African descent, African American, black, Asian, middle Eastern, Oriental and other" as "non-white". Patients in AETs and PETs were predominantly white. The percentage of patients considered "non-white" in the AETs group was 11%. In the PETs this percentage was 15%.

Marital status

Twenty-three trials (36%) reported the marital status of their patients. Fourteen different definitions were used to describe the marital status. The most frequently used definition, which was used, only four times, was: married/not-married. We dichotomised marital status into "married/cohabitating" – "not married". In the RCT population 45% of the participating patients were married/cohabitating. Of the patients participating in AETs 46% was married/cohabitating. In PETs, 43% of the participants were married/cohabitating.

Employment

Only twelve (19%) trials reported information on employment status. Seven different types of definition were used to define employment status. The two most commonly used ways of reporting were: "employed-unemployed" (25%) and "percentage employed participants" (25%). We converted all reported information on employment status into: "paid work" – "non-paid work". We considered the subcategories "unemployed", "homemaker", "house person", "housewife", "student" and "retired" as "non-paid work". One trial [47] reported categorical information on employment status, which could not be converted into the dichotomous variable "paid work"- "non-paid work." The percentage of people with paid work in AETs was 59% and in PETs 66%.

Education

Seventeen trials (27%) reported information on educational level. Approximately half of these trials described the educational level by years of education (n=9). The other half described the educational level by means of categories (n=7). One trial used both ways to describe the educational level. All seven trials describing the educational level by means of categories used different definitions. We converted the reported information on educational level of all trials into a trichotomous variable: high school or less – some college education – college graduate or more. Two trials reported information that could not be converted into a trichotomous variable. This exclusion resulted in too few trials (n=5) to reliably estimate the educational level of the RCT population.

Socioeconomic status

Only three trials (5%) reported socioeconomic status (SES). Two trials used the Hollingshead and Redlich's two-factor index of social position. This index refers a person's social class to that of his family and is determined with reference to the education and occupation of the family head plus the location of the family place of residence. Five class levels are distinguished, with level five being the lowest class and level one the highest [48]. One trial used the Blishen index [49] to describe the social economic status. This index is based on the Canadian Census and uses 514 occupational categories according to the Canadian

Classification and Dictionary of Occupations. Indicators of prevailing education and income levels are derived for each occupational category. A lower index indicates a lower SES.

Income

Only three trials (5%) reported information on income. Two trials reported income/year as a continuous variable (amount of money/year), one trial reported income as a categorical variable (<8.000, 8.000–16.000, >16.000 US Dollar per year). Too few selected trials reported on income to estimate the income of the RCT population.

The sociodemographic and socioeconomic features of the RCT participants are described in table 3.

Table 3. Sociodemographic/socioeconomic features of RCT participants.

	RCT (n=64)	**AET (n=45)**	**PET (n=19)**
Age (years)	41	41	37
Gender (% female)	62	61	72
Ethnicity (% "non-white")	11	11	15
Marital status (% married/cohabitating)	45	46	43
Employment status (% employed)	61	59	66
Educational level	Reported only in 8% of included trials	-	-
Socioeconomic status	Reported only in 5% of included trials	-	-
Income	Reported only in 5% of included trials	-	-

DISCUSSION

To our knowledge, this is the first review on the reporting and operationalisation of sociodemographic and socioeconomic features of participants in antidepressant efficacy trials (AET) and psychotherapy efficacy trials (PET) in major depression.

Remarkably, we found that in RCTs the reporting and operationalisation of sociodemographic and socioeconomic features turned out to be very diverse and for socioeconomic variables often very limited, even in the high impact factor journals. Only age, gender and race were reported in the majority of studies. All other features were reported in less than 40% of the trials and often operationalised in very different ways. The lack of standardisation in defining sociodemographic and socioeconomic variables and their insufficient reporting in RCTs may be explained by the fact that interest in the relation of social economic status and treatment outcome is relatively young. Only recently, RCTs have

been carried out in specific populations like low-income women [50] and ethnic minorities [51,52]. RCTs in specific subgroups is one way to address the influence of socioeconomic features on treatment outcome in MDD, yet more interest for SES features in "general" trials is needed, since guidelines are based on results from these trials. Furthermore, our findings suggest that there are differences between AET participants and PET participants with respect to several sociodemographic and socioeconomic features. In meta-analyses results from AETs and PETs are often directly compared, without controlling for SES features as marital status, educational level, employment status etc., since these features are not reported in trials. SES features are known to influence outcome, and therefore one risks to introduce confounders in the comparison between AETs and PETs. Several factors may explain differences in the SES features between participants in AETs and PETs, for example patients' preferences for certain types of treatment, or the use of specific eligibility criteria in AETs, like the exclusion of women who are pregnant or do not use contraceptives.
Both clinical practice and scientific research would benefit from uniform reporting of a standard set of SES features. In this way, estimation of the generalizability of results of RCTs to daily practice, comparison between RCTs and future research on the influence of SES features on outcome is facilitated.

There are some limitations to our study to consider. We performed a restricted search for AET's, which may not fully represent the available literature. However, the fact that we found significant underreporting of SES features in the AETs from the included high impact factor journals suggests that that underreporting of SES features in AETs in the whole body literature might be even worse. On the other hand, we found no association between the impact factor of the journal and the reporting of sociodemographic features.

We only included RCTs published till 2008, as we did not find PETs after 2007 that met our selection-criteria. It is possible that the reporting of sociodemographic and socioeconomic features has improved after 2007. We examined a sample of AETs published after 2007 [53-59] that met our exclusion criteria. In these studies published after 2007 we found a similar variety of reporting. Finally, it is important to note that when discussing the generalizability of results of RCTs to daily practice, one might easily overlook the fact that RCTs are explicitly designed to provide relative outcomes (differences between active treatment and placebo), rather than absolute effects of treatment. However, as treatment guidelines are based on the results from RCTs and used in daily practice, where the absolute treatment effect is far more important than the relative effect, it is very important for clinicians to know to what extent RCT participants resemble their "real life" patients.

CONCLUSIONS

Previous research has shown that SES features of patients can influence treatment outcome in depression. RCTs for treatments of depression do not adequately report on SES features. A uniform reporting of a standard set of sociodemographic and socioeconomic features is recommendable; especially on those features that are already known to be associated with treatment outcome (age, gender, marital and employment status). This would facilitate comparisons not only within the body of RCTs, but especially of RCT populations with 'real-life' populations, which would clearly benefit daily practice and guideline development.

REFERENCE LIST

1. World Health Organisation. Depression. 2010.
2. Badamgarav E, Weingarten SR, Henning JM, Knight K, Hasselblad V, Gano A, Jr. *et al*.: Effectiveness of disease management programs in depression: a systematic review. *Am J Psychiatry* 2003, 160: 2080-2090.
3. Neumeyer-Gromen A, Lampert T, Stark K, Kallischnigg G: Disease management programs for depression: a systematic review and meta-analysis of randomized controlled trials. *Med Care* 2004, 42: 1211-1221.
4. Trivedi MH, Claassen CA, Grannemann BD, Kashner TM, Carmody TJ, Daly E *et al*.: Assessing physicians' use of treatment algorithms: Project IMPACTS study design and rationale. *Contemp Clin Trials* 2007, 28: 192-212.
5. Fava GA, Ruini C, Rafanelli C: Sequential treatment of mood and anxiety disorders. *J Clin Psychiatry* 2005, 66: 1392-1400.
6. 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.
7. Rothwell PM: External validity of randomized controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005, 365: 82-93.
8. Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *Am J Psychiatry* 1999, 156: 5-10.
9. 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.
10. 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.
11. 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.
12. Stewart JW, McGrath PJ, Quitkin FM: Can mildly depressed outpatients with atypical depression benefit from antidepressants? *Am J Psychiatry* 1992, 149: 615-619.
13. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
14. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
15. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. *J Clin Psychiatry* 1996, 57: 572-575.
16. 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.
17. van der Lem R, van der Wee NJ, van VT, Zitman FG: The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychol Med* 2011, 41: 1353-1363.
18. Riall TS, Reddy DM, Nealon WH, Goodwin JS: The effect of age on short-term outcomes after pancreatic resection: a population-based study. *Ann Surg* 2008, 248: 459-467.
19. Gibson PH, Croal BL, Cuthbertson BH, Gibson G, Jeffrey RR, Buchan KG *et al*.: Socioeconomic status and early outcome from coronary artery bypass grafting. *Heart* 2009, 95: 793-798.
20. Aberg-Wistedt A, Agren H, Ekselius L, Bengtsson F, Akerblad AC: Sertraline versus paroxetine in major depression: clinical outcome after six months of continuous therapy. *J Clin Psychopharmacol* 2000, 20: 645-652.

21. Kornstein SG, Schatzberg AF, Thase ME, Yonkers KA, McCullough JP, Keitner GI *et al.*: Gender differences in treatment response to sertraline versus imipramine in chronic depression. *Am J Psychiatry* 2000, 157: 1445-1452.
22. Joyce PR, Mulder RT, Luty SE, Sullivan PF, McKenzie JM, Abbott RM *et al.*: Patterns and predictors of remission, response and recovery in major depression treated with fluoxetine or nortriptyline. *Aust N Z J Psychiatry* 2002, 36: 384-391.
23. Papakostas GI, Petersen T, Mischoulon D, Hughes ME, Spector AR, Alpert JE *et al.*: Functioning and interpersonal relationships as predictors of response in treatment-resistant depression. *Compr Psychiatry* 2003, 44: 44-50.
24. Baca E, Garcia-Garcia M, Porras-Chavarino A: Gender differences in treatment response to sertraline versus imipramine in patients with nonmelancholic depressive disorders. *Prog Neuropsychopharmacol Biol Psychiatry* 2004, 28: 57-65.
25. Lowe B, Schenkel I, Bair MJ, Gobel C: Efficacy, predictors of therapy response, and safety of sertraline in routine clinical practice: prospective, open-label, non-interventional postmarketing surveillance study in 1878 patients. *J Affect Disord* 2005, 87: 271-279.
26. 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.
27. Jarrett RB, Eaves GG, Grannemann BD, Rush AJ: Clinical, cognitive, and demographic predictors of response to cognitive therapy for depression: a preliminary report. *Psychiatry Res* 1991, 37: 245-260.
28. Sotsky SM, Glass DR, Shea MT, Pilkonis PA, Collins JF, Elkin I *et al.*: Patient predictors of response to psychotherapy and pharmacotherapy: findings in the NIMH Treatment of Depression Collaborative Research Program. *Am J Psychiatry* 1991, 148: 997-1008.
29. Thase ME, Reynolds CF, III, Frank E, Simons AD, McGeary J, Fasiczka AL *et al.*: Do depressed men and women respond similarly to cognitive behavior therapy? *Am J Psychiatry* 1994, 151: 500-505.
30. Hollon SD, Derubeis RJ, Evans MD, Wiemer MJ, Garvey MJ, Grove WM *et al.*: Cognitive therapy and pharmacotherapy for depression. Singly and in combination. *Arch Gen Psychiatry* 1992, 49: 774-781.
31. Falconnier L: Socioeconomic status in the treatment of depression. *Am J Orthopsychiatry* 2009, 79: 148-158.
32. Goekoop JG, Hoeksema T, Knoppert-Van der Klein EA, Klinkhamer RA, Van Gaalen HA, Van Londen L *et al.*: Multidimensional ordering of psychopathology. A factor-analytic study using the Comprehensive Psychopathological Rating Scale. *Acta Psychiatr Scand* 1992, 86: 306-312.
33. Van HL, Schoevers RA, Dekker J: Predicting the outcome of antidepressants and psychotherapy for depression: a qualitative, systematic review. *Harv Rev Psychiatry* 2008, 16: 225-234.
34. Croughan JL, Secunda SK, Katz MM, Robins E, Mendels J, Swann A *et al.*: Sociodemographic and prior clinical course characteristics associated with treatment response in depressed patients. *J Psychiatr Res* 1988, 22: 227-237.
35. Hoberman HM, Lewinsohn PM, Tilson M: Group treatment of depression: individual predictors of outcome. *J Consult Clin Psychol* 1988, 56: 393-398.
36. Troisi A, Pasini A, Bersani G, Grispini A, Ciani N: Ethological predictors of amitriptyline response in depressed outpatients. *J Affect Disord* 1989, 17: 129-136.
37. Blom MB, Spinhoven P, Hoffman T, Jonker K, Hoencamp E, Haffmans PM *et al.*: Severity and duration of depression, not personality factors, predict short term outcome in the treatment of major depression. *J Affect Disord* 2007, 104: 119-126.
38. Blom MB, Hoek HW, Spinhoven P, Hoencamp E, Judith Haffmans PM, van DR: Treatment of depression in patients from ethnic minority groups in the Netherlands. *Transcult Psychiatry* 2010, 47: 473-490.
39. Fortuna LR, Alegria M, Gao S: Retention in depression treatment among ethnic and racial minority groups in the United States. *Depress Anxiety* 2010, 27: 485-494.
40. Givens JL, Houston TK, Van Voorhees BW, Ford DE, Cooper LA: Ethnicity and preferences for depression treatment. *Gen Hosp Psychiatry* 2007, 29: 182-191.

41. Jones EE: Psychotherapists' impressions of treatment outcome as a function of race. *J Clin Psychol* 1982, 38: 722-731.
42. Lesser IM, Myers HF, Lin KM, Bingham MC, Joseph NT, Olmos NT *et al*.: Ethnic differences in antidepressant response: a prospective multi-site clinical trial. *Depress Anxiety* 2010, 27: 56-62.
43. Schraufnagel TJ, Wagner AW, Miranda J, Roy-Byrne PP: Treating minority patients with depression and anxiety: what does the evidence tell us? *Gen Hosp Psychiatry* 2006, 28: 27-36.
44. Stahl SM: Placebo-controlled comparison of the selective serotonin reuptake inhibitors citalopram and sertraline. *Biol Psychiatry* 2000, 48: 894-901.
45. Bielski RJ, Ventura D, Chang CC: A double-blind comparison of escitalopram and venlafaxine extended release in the treatment of major depressive disorder. *J Clin Psychiatry* 2004, 65: 1190-1196.
46. Shelton RC, Haman KL, Rapaport MH, Kiev A, Smith WT, Hirschfeld RM *et al*.: A randomized, double-blind, active-control study of sertraline versus venlafaxine XR in major depressive disorder. *J Clin Psychiatry* 2006, 67: 1674-1681.
47. Keller MB, Gelenberg AJ, Hirschfeld RM, Rush AJ, Thase ME, Kocsis JH *et al*.: The treatment of chronic depression, part 2: a double-blind, randomized trial of sertraline and imipramine. *J Clin Psychiatry* 1998, 59: 598-607.
48. Mollica RF, Milic M: Social class and psychiatric practice: a revision of the Hollingshead and Redlich model. *Am J Psychiatry* 1986, 143: 12-17.
49. Blishen BR: Socioeconomic Index for Occupations in Canada. *Canadian Review of Sociology and Anthropology* 1967, 4: 41-53.
50. Araya R, Flynn T, Rojas G, Fritsch R, Simon G: Cost-effectiveness of a primary care treatment program for depression in low-income women in Santiago, Chile. *Am J Psychiatry* 2006, 163: 1379-1387.
51. Miranda J, Cooper LA: Disparities in care for depression among primary care patients. *J Gen Intern Med* 2004, 19: 120-126.
52. Roy-Byrne PP, Perera P, Pitts CD, Christi JA: Paroxetine response and tolerability among ethnic minority patients with mood or anxiety disorders: a pooled analysis. *J Clin Psychiatry* 2005, 66: 1228-1233.
53. Blier P, Ward HE, Tremblay P, Laberge L, Hebert C, Bergeron R: Combination of antidepressant medications from treatment initiation for major depressive disorder: a double-blind randomized study. *Am J Psychiatry* 2010, 167: 281-288.
54. Cutler AJ, Montgomery SA, Feifel D, Lazarus A, Astrom M, Brecher M: Extended release quetiapine fumarate monotherapy in major depressive disorder: a placebo- and duloxetine-controlled study. *J Clin Psychiatry* 2009, 70: 526-539.
55. Kasper S, Hajak G, Wulff K, Hoogendijk WJ, Montejo AL, Smeraldi E *et al*.: Efficacy of the novel antidepressant agomelatine on the circadian rest-activity cycle and depressive and anxiety symptoms in patients with major depressive disorder: a randomized, double-blind comparison with sertraline. *J Clin Psychiatry* 2010, 71: 109-120.
56. Liebowitz MR, Yeung PP, Entsuah R: A randomized, double-blind, placebo-controlled trial of desvenlafaxine succinate in adult outpatients with major depressive disorder. *J Clin Psychiatry* 2007, 68: 1663-1672.
57. Perahia DG, Quail D, Desaiah D, Corruble E, Fava M: Switching to duloxetine from selective serotonin reuptake inhibitor antidepressants: a multicenter trial comparing 2 switching techniques. *J Clin Psychiatry* 2008, 69: 95-105.
58. Stahl SM, Fava M, Trivedi MH, Caputo A, Shah A, Post A: Agomelatine in the treatment of major depressive disorder: an 8-week, multicenter, randomized, placebo-controlled trial. *J Clin Psychiatry* 2010, 71: 616-626.
59. Whitmyer VG, Dunner DL, Kornstein SG, Meyers AL, Mallinckrodt CH, Wohlreich MM *et al*.: A comparison of initial duloxetine dosing strategies in patients with major depressive disorder. *J Clin Psychiatry* 2007, 68: 1921-1930.

Chapter 6

Influence of sociodemographic and socioeconomic features on treatment outcome in RCTs versus daily psychiatric practice

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

Social Psychiatry and Psychiatric Epidemiology. 2012, 12 DOI:10.1007/s00127-012-0624-4 (published online)

ABSTRACT

Purpose: Sociodemographic and socioeconomic characteristics of participants in antidepressant and psychotherapy efficacy trials (AETs and PETs) for major depressive disorder (MDD) may limit the generalizability of the results. We compared trial participants to daily practice patients. We subsequently assessed the influence of sociodemographic and socioeconomic status on treatment outcome in daily practice.

Methods: Data on daily practice patients were derived through Routine Outcome Monitoring (ROM). We included 626 patients with MDD according to the MINIplus. Distributions of age, gender, race, marital status and employment status were compared to participants in 63 selected AETs and PETs. Influence of these features on treatment outcome was explored through multivariate regression analysis.

Results: Trial participants were older, more often male (diff. 4%, $p=0.05$), white (diff 4%, $p<0.001$) and not married (diff 7%, $p=0.003$). Although significant, most differences were relatively small. However, the difference in employment status was striking: 34% of the ROM patients were currently working, versus 68% of the trial participants (diff. 34%, $p<0.001$). Being employed contributed to a positive treatment outcome: OR 1.8 for response (50% reduction of MADRS), OR 1.9 for remission (MADRS ≤10).

Conclusions: employment status should be taken into account while interpreting results from RCTs and as predictor of treatment success in daily practice.

Keywords: major depressive disorder, randomized controlled trial, sociodemographic status, socioeconomic status, patient selection

INTRODUCTION

International guidelines on treatment of major depressive disorder are based on evidence retrieved from scientific research. Meta-analyses and large randomized controlled trials (RCTs) are considered to give the most reliable evidence on treatment outcome for most therapies. RCTs use very strict procedures in design, patient selection, randomization, and methodology to enhance internal validity, i.e. the ability to adequately determine the efficacy of an intervention versus a control condition. While doing so, RCTs diminish their external validity, i.e. the generalizability to routine daily practice. Therefore, many clinicians and researchers have stated that the practical value of RCTs as source of evidence for treatment efficacy in daily practice is limited [1-7].

In previous research [1], many differences which might influence treatment outcome between trial methodology and daily practice have been described. One of the important domains that influence external validity is the selection of participants. The methodology of recruitment, the use of eligibility criteria, the (unintentional) use of criteria beyond eligibility (e.g. by recruitment only of patients with medical insurance, or in a certain area), the use of run-in periods and enrichment strategies, all lead to the exclusion of possible non-responders, and may lead to selection bias in the research population. For major depressive disorder (MDD), most research has focused on the influence of eligibility criteria on external validity. Several researchers [8-11] have shown that only a minority of daily practice patients would be eligible for RCTs. The influence of these eligibility criteria on treatment outcome is not yet clear: Wisniewski and colleagues [9] found that eligible patients had better treatment outcome than non-eligible patients, on the other hand, in our own research [11], we found no differences in treatment outcome between eligible and non-eligible patients. However, the use of eligibility criteria is only one aspect of patient selection. Even if eligibility criteria are met, recruited RCT participants differ from not recruited participants in terms of age, sex, race, educational status, social class and place of residence [1]. In previous research, we found a great variety in the reporting of sociodemographic and socioeconomic (SES) features in major depression trials. SES factors are known to influence treatment outcome in MDD [12-22]. Differences in SES status of participants might be one of the explanations for the differences between treatment outcome in RCTs and in pragmatic trials. However, although much effort is put in approaching daily practice as much as possible in pragmatic trials like STAR*D, they still have several properties of an RCT and selection bias by recruitment is not completely ruled out [23].

In the current project, we first compared the SES status of participants in psychotherapy efficacy trials (PETs) and antidepressants efficacy trials (AETs) to the SES status of a daily practice population of patients suffering from MDD. Subsequently, we explored the influence of the SES status on treatment outcome in daily practice to assess whether possible SES differences between RCT participants and daily practice patients contribute to differences

in treatment outcome between RCTs and routine practice. Comprehensive data on patients' characteristics, treatment modalities and treatment outcome were available through an extensive Routine Outcome Monitoring system (ROM) and extensive chart review. ROM provides outcome data on treatments without the properties of clinical trial but with strong methodological features. Therefore ROM is a qualified instrument to obtain insight in "real life daily practice" [24].

Aims of the study

We investigate the differences in sociodemographic and socioeconomic features between "real life" patients suffering from MDD and trial participants. Furthermore we explore the influence of sociodemographic and socioeconomic features on treatment outcome. This study gives insight on the extent of generalizability of results from RCTs to daily practice as well as the influence of sociodemographic and socioeconomic status on treatment outcome in daily practice.

METHODS

Selection of RCTs and definitions of SES features in RCTs

We included peer reviewed publications of RCTs, published through 2007 in outpatients with a unipolar, non-psychotic major depressive disorder according to DSM-III-R or DSM-IV. Because we aimed to review the reporting of sociodemographic and socioeconomic features in RCTs usually selected for the development of guidelines for routine treatment, we excluded trials which a priori included only participants from specified subgroups like elderly or a specific ethnic minority. Furthermore, it was essential that the publication provided baseline information on sociodemographic and/or socioeconomic features. We included trials written in English, since international guidelines for treatment of MDD are predominantly based on English literature. Because of the large number of published AETs, we restricted our search to AETs published in journals from the top ten Impact Ranking psychiatric journals of 2005. By including only high impact factor journals, we expected to have a sample of trials with the most systematic manner of reporting SES features. For PETs, which are less frequently published, we performed a Medline search for RCTs investigating psychotherapy (cognitive behavior therapy and interpersonal therapy) for adult patients suffering from MDD. We performed an additional search in PsycInfo and checked the reference lists of included trials for other relevant studies as well as the database http://www.psychotherapyrcts.org. 45 AETs and 19 PETs were selected. Detailed information on RCT-selection and exploration of SES features of RCT participants has been previously reported [25]. The selected AETs and PETs were equally distributed between Europe and the United States of America. Of all the different SES features mentioned in the selected RCTs,

we only used the SES features that were reported in at least 15% of the selected trials (n=10). In RCTs, SES features are described in various manners. In order to allow an estimation of the mean socioeconomic status of RCT participants, the reporting of the SES feature in the RCT had to be in such manner that the feature could be converted in a dichotomous variable. The following SES features were reported frequently enough (in at least 15% of the trials) and were suitable for comparison between RCTs and our ROM population: age, gender, race, marital status and employment status. Educational level was relatively often reported in RCTs but in so many different ways that dichotomization was not possible. Income and social position were reported in less than 15% of the RCTs. The selected SES features were defined as follows:

- Age: age at baseline assessment calculated in years.
- Gender: percentage females in the population.
- Race: percentage white patients in the population.
- Marital status: percentage of married/cohabitating patients.
- Employment status: percentage of patients who have a paid job.

Treatment outcome in MDD trials is usually defined as percentage of responders (50% reduction of symptoms) or as percentage of remitters (score below a certain severity cut-off).

Patient selection and definitions of SES features in the Routine Outcome Monitoring population

In the Netherlands, health insurance for all citizens is regulated by the government. Therefore, (mental) health care is easily accessible and not restricted by the financial means of individual patients. The Dutch health care system is organised in a stepped-care-manner and described in treatment guidelines. First, patients visit their general practitioner (GP). In case of depression, the treatment guidelines for GPs recommend that patients with mild depression are treated by their GP with lifestyle advices. If the depression is moderate (clinical judgment) the GP can decide to prescribe antidepressants. Reasons to refer patients to a regional mental health provider (RMHP) are: preference of patients for psychotherapy (not provided by GPs); severe depression (clinical judgment); presence of co morbid disorders; complex situation of the patient due to physical health problems; social problems etc.; duration of the depression; and lack of result of antidepressant therapy. After baseline assessment and a clinical interview at our RMHP, patients suffering from major depression are offered to choose between psychotherapy and antidepressant therapy (if the severity is not too high, judged by the interviewing clinician in combination with results from ROM). When patients are already taking antidepressants from their GP, if needed the dose is optimized, and subsequently patients are offered to switch to another antidepressant or psychotherapy is added to the antidepressant therapy.

In 2002, the Regional Mental Health Provider (RMHP) Rivierduinen (service area with 1.1 million inhabitants), in collaboration with the University Medical Hospital Leiden, implemented a comprehensive system of routine outcome monitoring (ROM) and evidence based, stepped care protocols. In ROM, all patients referred to the RMHP for treatment of a mood, anxiety or somatoform disorder have an extensive baseline assessment. Treatment progress is then assessed at intervals of three to four months and before starting a new treatment step. The baseline assessment comprises a standardized diagnostic interview (Mini-International Neuropsychiatric Interview Plus [26]), the collection of sociodemographic and socioeconomic data, the administration of observer rated scales and self report questionnaires, and general measures of health and quality of life. All patients with sufficient mastery of the Dutch language who are able to complete computerized and written questionnaires are eligible for ROM. After baseline assessment and a clinical interview, patients suffering from major depression are offered to choose between psychotherapy and antidepressant therapy. When ROM data are used for research purposes, these are provided to researchers in an anonymous form, as dictated by the Psychiatric Academic Registration Leiden (PAREL) regulation. This procedure has been approved by the Medical Ethical Committee of the University Medical Hospital Leiden. The design of Routine Outcome Monitoring has been reported previously [27].

From the ROM population who sought treatment at the RMHP Rivierduinen from January 2002 until January 2007, we included all outpatients who met the following criteria:

- Patients had a DSM-IV diagnosis of a current major depressive disorder as established by the Mini International Neuropsychiatric Interview.
- Patients had at least one follow-up assessment in ROM.
- Patients were not suspected of bipolarity/psychotic features by their clinician.
- Patients were not admitted to an inpatient clinic during follow-up.
- Patients did not have a MADRS-score ≤ 10 at baseline assessment.
- The time-span between baseline and follow-up assessment was not too short (less than four weeks) or too long (more than one year) to provide reliable information.

We conducted an extensive chart review in order to obtain information on the type of offered treatment. We examined possible selection bias due to our inclusion criteria by comparing the baseline severity of the depression and the SES features of the selected patients to the characteristics of patients who had no follow up assessments in ROM. We compared the selected patients to all patients who suffered from major depression according to the MINIplus (including patients who dropped out of treatment after baseline assessment and patients of whom it was not clear if they dropped out or not). Subsequently, we compared the patients with follow-up to patients who did receive treatment, but had no follow-up assessments in ROM. Information on how many patients did receive treatment was obtained through an anonymized database from the medical administration of RMHP Rivierduinen.

Patients in our population received different types of treatment for their depression. Most of the treatments (84%) were in line with international treatment guidelines: antidepressants (19% with additional social supportive therapy/counseling), individual psychotherapy (mostly cognitive behavioral therapy or interpersonal therapy) or combination therapy of pharmacotherapy and cognitive behavioral therapy or interpersonal therapy. Sixteen percent of the patients received treatment other than antidepressants, individual psychotherapy or combination treatment. In line with RCTs we defined treatment-response as a 50% reduction of symptoms on the Montgomery Asberg Rating Scale for Depression (MADRS, [28]) and remission as MADRS≤ 10.

In ROM extensive information on sociodemographic and socioeconomic status is gathered. Age and gender are recorded. Ethnicity is primarily assessed by the recording of patients' and their parents' countries of birth. Consistent with most RCTs, we considered patients who were born (or whose parents were born) in Suriname, Antilles, Turkey and Morocco as non-white. Two-third of the patients from other non-Dutch origins was non-white. We considered patients who were single, living-apart-together, divorced or widowed as "not married/cohabitating".

Like other member states of the European Union, the Netherlands have a social security system which provides sickness benefit to people who are temporarily not able to work and disability benefit to those who are not able to work anymore. In the USA, the provision of sickness or disability benefit depends upon the insurance policies. Other countries provide no sickness or disability benefit or use their own system. Clearly, there is an important difference in financial status between receiving a sickness or disability benefit or not. It is unknown whether in RCTs, patients who receive sickness benefit are considered as having a paid job or not. Therefore, to allow comparison with RCTs, we defined employment status in ROM with two separate variables: employment status I and II:

For employment status I patients who reported having a paid job at baseline assessment are considered as having "paid work". In this definition "paid work" means that patients are *working* at the time of baseline assessment. Patients on sickness or disability benefit, unemployed, student or housewife were considered as having "no paid work". For employment status II patients who receive sickness or disability benefit were classified as having "paid work" while unemployed patients, students and housewives were still considered as having "no paid work". In this definition having "paid work" means that patients receive an income out of a job, but that a substantial part of these patients (53% of the patients classified as having paid work) is not working but on sickness benefit.

Statistical Analysis

We first compared the included sample from our ROM population with the lost-to-follow-up in ROM patients by using Chi-square tests and independent sample t-tests. Subsequently, the SES features of RCT participants were compared to those of our ROM population by

comparison of proportions of two independent groups [29]. The selected RCTs did not report enough information on the characteristic "Age" to allow a comparison between RCT population and our population.

Finally, we examined the influence of SES features on outcome in routine clinical practice. In our sample, there were missing values for the following variables: type of offered treatment (n=28), and socioeconomic characteristics (n=82). Comparison of complete cases and cases with missing data showed differences on several variables such as age, gender and treatment outcome. In such instances, complete case analysis may yield biased estimates [30]. Therefore, the MICE (multivariate imputation by chained equations [31]) method was used to estimate missing values for type of offered treatment, and socioeconomic characteristics. The influence of the sociodemographic (age, gender, race) and socioeconomic (marital status, employment status) on treatment outcome was computed by logistic regression after MICE. Univariate regression models (unadjusted analysis) were used to explore the influence of the individual SES features on treatment outcome. A multivariate regression model (adjusted analysis) was used to explore the joint influence of all SES features on treatment outcome. In this model, type of offered treatment and educational level were entered as possible confounders. In our previous research [11] on the same patient population, we showed that typical clinical features of MDD patients in daily practice, namely co morbid Axis I disorders, substance abuse, and suicidality, which are often used as exclusion criteria in RCTs, were not associated with treatment outcome in clinical practice. A low baseline severity, however, was associated with a less favorable treatment outcome. Therefore, we entered baseline severity of the depression as another possible confounder in our multivariate regression model.

Odds-ratios (OR) and their confidence-intervals were computed by using the robust standard error. Statistical analyses were performed with SPSS 16.0 and STATA10.0.

RESULTS

Daily practice (ROM) population

4157 outpatients were assessed at baseline between January 2002 and January 2007. Of these patients, 1653 suffered from a current major depressive disorder according to the MINIplus. Since in scientific literature it is not well defined when a depression should be considered the "primary diagnosis" or when a so called "co morbid disorder", we included all patients who suffered from major depression according to the MINIplus. From the 1653 patients suffering from major depression, 46% (n=774) had at least one follow-up assessment. Extensive chart-review was done for those 774 patients. 148 patients had to be excluded from further follow-up analysis due to suspected bipolarity/psychotic features, admission to an inpatient clinic during follow-up, remission on the MADRS at baseline or an insufficient

time-span between baseline and follow-up assessment. Finally, 626 patients were eligible for follow-up analysis. Of 76% of these 626 patients, the clinician stated that depressive disorder was the primary clinical diagnosis. 24% of the patients suffered from depression, but it was considered to be a co morbid disorder. Selection of the patient population is described in figure 1. The characteristics of the patient population are described in table 1.

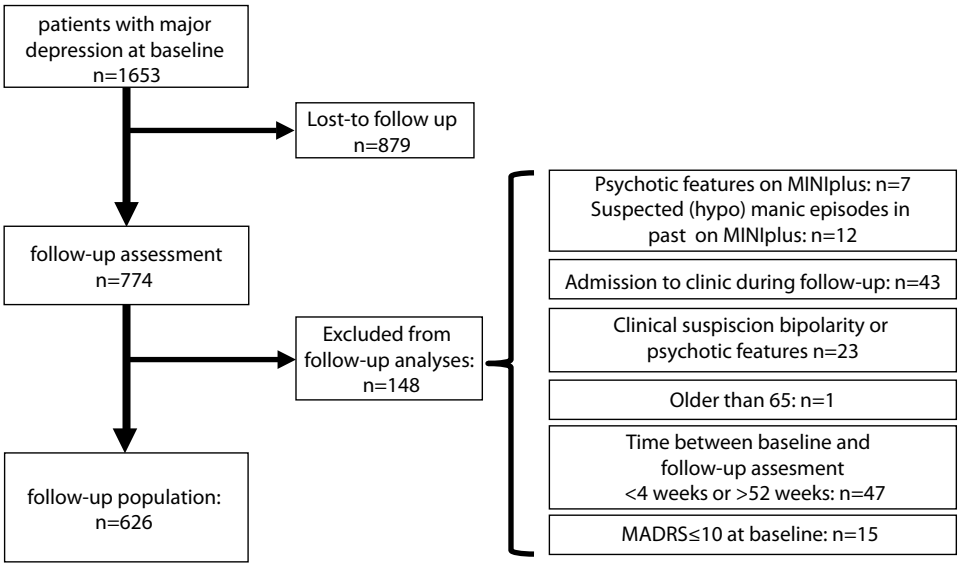

Figure 1. Selection of the follow-up group.

Table 1. Baseline characteristics of ROM patients.

N=626		
Mean age in years (standard deviation)	39.2 (11.4)	
Gender ***% Female***[1]	Male 3 Female ***66 %***	3.7% 66.3%
Presence of co morbid other Axis I disorder	69.9% (majority anxiety disorder)	
Mean MADRS at baseline	25.8 (6.5)	
Type of offered treatment[2]	AD IP AD+IP AD+SST Other	13.8% 28.4% 27.9% 15.9% 14.1%

Duration of treatment in weeks (95% CI[3])	AD IP AD+IP AD+SST Other	20.8 (18.7-22.9) 20.1 (18.5-21.6) 21.5 (20.0-23.1) 21.5 (20.0-23.1) 19.1 (17.1-21.1)
Marital status ***% Married/cohabitating[1]***	Married/cohabitating Divorced/widowed Not married ***52%***	52.2% 16.2% 31.6%
Birth Country Parental Birth Country ***Race (% whites)[1]***	Netherlands Turkey/Morocco Suriname/Antilles Other Netherlands Turkey/Morocco Suriname/Antilles Other ***84%***	84.9% 5.5% 2.9% 6.6% 79.2% 7.2% 2.9% 10.7%
Employment status ***% paid work I[1]*** ***% paid work II[1]***	Employed Not employed Sickness benefit ***34%*** ***74%***	26.1% 34.4% 39.5%

[1] Dichotomization of variable for comparison to RCT/AET/PET, see Method section.
[2] AD=antidepressants, IP=individual psychotherapy, SST=social supportive therapy.
[3] 95% CI= 95% confidence interval.

The selected patient group was similar to the not-selected patients in baseline severity of the depression and most SES features. The selected patients were slightly older (39.2 versus 37.3 years) and more often married/cohabitating (52.6% versus 46.9%) than the patients who were not selected (see table 2, supplementary material). Of the patients who were lost to follow-up, and therefore not selected, 63% received treatment in our outpatient clinics for mood-, anxiety, and somatoform disorders after baseline assessment (37% was referred to other specialized departments or did drop out of treatment). We also compared the selected patients to the patients who did receive treatment in our outpatient clinics but had no follow-up in ROM and found them to be very similar (see table 3, supplementary material). Therefore we consider our selected patient sample as fairly representative for patients who receive outpatient treatment for depressive disorders.

RCT participants versus daily practice (ROM) population

We compared the RCT population and our ROM sample on age, gender, ethnicity, marital status, and employment status. RCT participants were less often female (62% vs. 66%), more often white (89% vs. 84%), less often married/cohabitating (45% vs. 52%) and more often employed (68% vs. 34%) than daily practice (ROM) patients. Both in antidepressant and psychotherapy trials, participants were more often white, less often married/cohabitating and more often employed than daily practice (ROM) patients. Both in antidepressant and psychotherapy trials, the male-female ratio was different than in daily practice (ROM): in antidepressant trials the ratio is in favor of males (39% vs. 34% male) compared to daily practice (ROM), while in psychotherapy trials the ratio is in favor of females (73% vs. 66% female) compared to daily practice (ROM). Although the differences were statistically significant, they were sometimes relatively small and significance was probably reached due to the large number of patients included in both groups. However, the exception is the difference between RCT participants and our daily practice population in employment status (employment status I: defined as currently having "paid work", not on sickness or disability benefit); 68% of the RCT participants were currently employed, whereas only 34% of the daily practice patients had a paid job. Results from the comparison between RCT participants and our population are described in table 4.

Table 4. SES features of RCT participants versus ROM patients.

	RCT (%, n)			**ROM[2] (%, n)**	**Difference in proportion RCT versus ROM (95% CI; p-value)**		
		AET (%, n)	PET (%, n)			Difference in proportion AET versus ROM (95% CI, p-value)	Difference in proportion PET versus ROM (95% CI, p-value)
Age in Years[1]	**41**	41	37	**39 (sd 11.4)**	-	-	-
Gender (% female)	**62% (5880)**	61% (5269)	73% (611)	**66% (415)**	**-4% (-8% – 0%; 0.05)**	-5% (-9% – -11%; 0.016)	7% (2% – 11%; 0.008)
Race (% white)	**89% (5820)**	89% (5400)	85% (420)	**84% (457)**	**4% (2% – 8%; <0.001)**	4% (2 – 8%; <0.001)	-1% (-3% – -6%; NS)
Marital status (% married)	**45% (972)**	46% (728)	42% (244)	**52% (284)**	**-7% (-12% – -3%; 0.003)**	-10% (-10% – -1%; 0.017)	-18% (-24% – -12%; <0.001)
Employment I [3] (% paid work)	**68% (1107)**	68% (905)	66% (202)	**34% (187)**	**34% (29% – 38%; p<0.001)**	34% (30% – 39%; <0.001)	32% (25% – 38%; <0.001)
Employment II (% paid work or sickness benefit)	**68% (1107)**	68% (905)	66% (202)	**74% (402)**	**-6% (-10% – -1%; 0.013)**	-5% (-10% – -1%; 0.027)	-8% (-14% – -1%; 0.019)

[1] Statistical comparison not possible.

[2] For all 626 patients in ROM information on age and gender was available, for 82 patients no information was available on race, marital status and employment status.

[3] In Employment I, paid work is defined as only patients who are *currently* working, Employment II includes both patients who are currently working and who receive a sickness benefit.

AET: Antidepressant Efficacy Trial; PET: Psychotherapy Efficacy Trial; ROM: data from routine outcome monitoring; CI: confidence Interval; n: total number of patients studied in the included RCTs/ AETs/ PETs or ROM.

Influence of SES features on treatment outcome in the daily practice (ROM) population

In the univariate analyses, having a paid job (employment status I, in which patients who receive sickness benefit were defined as not having a paid job) contributed positively to remission (MADRS≤10 [28]) and response (50% reduction of the MADRS). If patients on sickness benefit are also considered to have a paid job (employment status II), the influence of employment status on treatment outcome disappeared. None of the other SES features contributed significantly to treatment outcome. In the multivariate analyses, we investigated the influence of Age, Gender, Race, Marital status, and Employment status (Employment I) on treatment outcome in one model. In this model we also adjusted for baseline severity of the depression, so that we could analyze the influence of SES features irrespective of the severity of the illness. We found that having a paid job contributed positively to remission (OR 1.85, 95%CI 1.2–2.8; RSE 0.21; p=0.003, R-square 7%), and response (OR 1.76, 95%CI 1.2–2.6; RSE 0.20; p=0.005, R-square 3%). Results from the unadjusted analysis of the influence of the individual SES features are described in table 5; Results from the analysis of the joint influence of all SES features adjusted for baseline severity of the depression, type of offered treatment and educational level are described in table 6.

Table 5. Influence of individual SES features on treatment outcome: unadjusted analysis.

	Remission (MADRS ≤ 10)	**Response (50% reduction MADRS)**
Age	OR 0.99 (95%CI 0.97–1.00; RSE 0.008; p=0.13)	OR 0.99 (95%CI 1.0v1.0; RSE 0.01; p=0.10)
Gender	OR 0.97 (95%CI 0.65–1.44; RSE 0.20; p=0.87)	OR 1.09 (95%CI 0.8–1.6; RSE 0.21; p=0.66)
Race	OR 0.48 (95%CI 0.27–1.1; RSE 0.41; p=0.08)	OR 0.51 (95%CI 0.2–1.1; RSE 0.41; p=0.08)
Marital status	OR 1.32 (95%CI 0.9–2.0; RSE 0.02; p=0.17)	OR 1.32 (95% CI 0.9–2.0; RSE 0.20; p=0.17)
Employment I	OR 2.30 (95%CI 1.6–3.4; RSE 0.20; p<0.001)	*OR 1.82* (95%CI 1.3–2.6; RSE 0.34; p=0.002)
Employment II	OR 1.12 (95%CI 0.7–1.1; RSE 0.23; p=0.61)	OR 1.18 (95%CI 0.8–1.8; RSE 0.21; p=0.42)

OR: odds ratio
CI: confidence interval
RSE: robust standard error

Table 6. Joint influence of SES features on treatment outcome: adjusted analysis. Age, gender, race, marital status and employment I status entered in one model as predictors of outcome corrected for severity of depression at baseline, type of offered treatment and educational level.

	Remission (MADRS ≤10)	**Response (50% reduction of MADRS)**
Age	OR 0.99 (95%CI 0.97–1.00; RSE 0.01; p=0.21)	OR 0.99 (95%CI 0.97–1.00; RSE 0.01; p=0.15)
Gender	OR 1.07 (95%CI 0.69–1.67; RSE 0.23; p=0.75)	OR 1.10 (95%CI 0.73–1.59; RSE 0.20; p=0.65)
Race	OR 0.67 (95%CI 0.32–1.40; RSE 0.38; p=0.29)	OR 0.63 (95%CI 0.33–1.19; RSE 0.32; p=0.15)
Marital status	OR 1.33 (95%CI 0.85–2.07; RSE 0.23; p=0.21)	OR 1.10 (95%CI 0.73–1.67; RSE 0.21; p=0.64)
Employment I	OR 1.89 (95%CI 1.24–2.81; RSE 0.21; p=0.003	OR 1.76 (95%CI 1.19–2.60; RSE 0.20; p=0.004)

OR: odds ratio
CI: confidence interval
RSE: robust standard error

DISCUSSION

In the current project, we compared sociodemographic and socioeconomic features of participants in randomized controlled trials for depression to those of patients in daily practice. Participants in RCTs for major depression differed from daily practice patients with respect to age, male-female ratio, ethnicity and marital status, but those differences were relatively small (less than 7% difference on all features). One striking difference between RCT participants and daily practice patients is their employment status. Only 34% of the daily practice patients had a paid job (patients who were students, housewives, unemployed or on sickness or disability benefit were considered as having "no paid work") at time of assessment, while 68% of the RCT participants had a paid job at time of the trial participation. In routine clinical practice, having a paid job contributed positively to treatment outcome, both on remission (OR 1.85) and response (OR 1.76) on the MADRS. Age, gender, race and marital status did not contribute to treatment outcome in routine clinical practice.

Previous research found unequivocal results on the associations between sociodemographic and socioeconomic features and treatment outcome: in general, increased age seemed to be associated with worse outcome in pharmacotherapy [13-19], but not in psychotherapy [20-22], and one study found that men had better treatment outcome in psychotherapy than women [21]. Patients who were married benefitted more from both pharmacotherapy and psychotherapy [18-21,32-35].

The difference in employment status between RCT participants and daily-practice patients was striking. In a recent study was found that the quality of mental health is associated with lesser years of unemployment [36]. In previous research, in non-routine practice settings, on the relation between employment and treatment success, it was found that employment status improved outcome for antidepressants, but not for psychotherapy [18-21,32-35]. In the univariate analyses we found that "currently working" (employment status I) is associated with better treatment outcome, while "receiving income out of a job" (employment status II) seemed not be associated with treatment success. It is, however, likely that other clinical factors, such as baseline severity, could have acted as a confounder in these univariate analyses, since patients with a more severe depression might drop out from work earlier. Therefore, in our multivariate analyses we adjusted for baseline severity of the depression. Still, having paid work (defined as currently working) almost doubled the probability of response or remission. From clinical experience, one might expect that the social status and the daily structure, routine and distraction provided by paid work will be factors that contribute to the positive response on depression treatment. Alternatively, it may be that patients who remain working despite the opportunity of sickness benefit have personality traits that increase the likelihood of positive treatment outcome, like for example optimism. Future research is needed to explore these different possibilities.

Of course, besides differences in SES features, there are many other differences between RCT participants and daily practice patients, caused by the use of stringent exclusion criteria in RCTs. For instance, the presence of co morbid Axis I disorders is used as an exclusion criterion in more than 75% of the antidepressant efficacy trials [37] and in 25–50% of the psychotherapy efficacy trials [38]. In previous research, we found that only 17–24% of our patients would be eligible for participation in antidepressant efficacy trials [11]. Most of our patients would have been excluded because of the presence of co morbid Axis I disorders and not meeting a baseline severity threshold. However, in our previous research [11], we also found that in daily practice the exclusion of patients with co morbid Axis I disorders does not influence the treatment result for depression. One can therefore argue that it is unlikely that co morbidity acted as a major confounder in our analysis of the influence of SES features on treatment outcome for depression. It is, however, very well possible that having a co morbid disorder could lead to earlier withdrawing from work. Since co morbid disorders occur frequently in depressed patients, more specific research on the association between drop-out from work and co morbidity is recommended.

Strengths

We consider the large sample of well-characterized, routinely monitored patients from routine clinical practice to be the major strength of our study. To our knowledge, no previous research has reported on the influence of sociodemographic and socioeconomic features on treatment outcome in major depression in routine daily clinical practice.

Limitations

There are also some limitations to consider. In our model of sociodemographic and socioeconomic features of RCT participants, we converted all information on these features reported in the individual RCTs into dichotomous or trichotomous variables. This conversion lead to loss of information, but was necessary for comparability with published data [25]. Another limiting factor might be that in the model AETs outnumbered PETs, both in number of included trials and number of included patients per trial. Thus, the features of the RCT-participants as a group were dominated by the features of AET participants [25]. There was a considerable loss-to-follow up in our sample. Nevertheless, we demonstrated that our patient selection was fairly representative for the daily practice patients who received treatment. It was not possible to analyze in our study which aspect of "having a paid job" contributed positively to a favorable treatment outcome. Furthermore, no information on the duration and the number of episodes of depression was available in our ROM data. Chronic depression is known to have al less favorable prognosis and it is possible that especially patients who suffer from chronic depression have to resign from work, which may have confounded our results. Finally, since sociodemographic and socioeconomic features are closely related to the culture in the country of origin in which research takes place, cultural aspects might have somewhat limited the generalizability of our results. It is unknown to what extent the Dutch health care system (and that of several other, mostly European, countries) may limit the generalizability of our results to other countries that do not have an extensive social security system. Furthermore, we have to take into account that we compared a western psychiatric population to an RCT population derived from western countries (Europe and USA). Since our aim was to compare RCT participants (who are most often from western countries) to our daily practice patients, the fact that all RCT participants were from western countries did not limit our research. However, our results are probably not generalizable to countries outside the western world. Future research on the generalizability of results from RCTs to psychiatric patients in other parts of the world is highly recommended.

CONCLUSION

In conclusion, we found that RCT participants and daily practice patients only differed slightly on most sociodemographic and socioeconomic features, with the exception of having a paid job. Having a paid job contributed significantly to treatment success in daily practice and should be taken into account both while interpreting results from RCTs as well as in depression treatment in daily practice. Further research is recommended to explore which specific aspects of employment status contribute to better treatment outcome.

REFERENCE LIST

1. Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005, 365: 82-93.
2. Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *Am J Psychiatry* 1999, 156: 5-10.
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: Generalisability of results from randomised drug trials. A trial on antimanic treatment. *Br J Psychiatry* 1997, 170:264-7.
6. Stewart JW, McGrath PJ, Quitkin FM: Can mildly depressed outpatients with atypical depression benefit from antidepressants? *Am J Psychiatry* 1992, 149: 615-619.
7. Moller HJ: How close is evidence to truth in evidence-based treatment of mental disorders? *Eur Arch Psychiatry Clin Neurosci* 2011.
8. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
9. 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.
10. Zetin M, Hoepner CT: Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol* 2007, 27: 295-301.
11. van der Lem R, van der Wee NJ, van Veen T., Zitman FG: The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychol Med* 2011, 41: 1353-1363.
12. Button K.S., Wiles N.J., Lewis G, Peters T.J., Kessler D.: Factors associated with differential response to online cognitive behaviavoural therapy. *Soc Psychiatry Psychiatr Epidemiol* 2012, 47: 827-833.
13. Aberg-Wistedt A, Agren H, Ekselius L, Bengtsson F, Akerblad AC: Sertraline versus paroxetine in major depression: clinical outcome after six months of continuous therapy. *J Clin Psychopharmacol* 2000, 20: 645-652.
14. Kornstein SG, Schatzberg AF, Thase ME, Yonkers KA, McCullough JP, Keitner GI *et al.*: Gender differences in treatment response to sertraline versus imipramine in chronic depression. *Am J Psychiatry* 2000, 157: 1445-1452.
15. Joyce PR, Mulder RT, Luty SE, Sullivan PF, McKenzie JM, Abbott RM *et al.*: Patterns and predictors of remission, response and recovery in major depression treated with fluoxetine or nortriptyline. *Aust N Z J Psychiatry* 2002, 36: 384-391.
16. Papakostas GI, Petersen T, Mischoulon D, Hughes ME, Spector AR, Alpert JE *et al.*: Functioning and interpersonal relationships as predictors of response in treatment-resistant depression. *Compr Psychiatry* 2003, 44: 44-50.
17. Baca E, Garcia-Garcia M, Porras-Chavarino A: Gender differences in treatment response to sertraline versus imipramine in patients with nonmelancholic depressive disorders. *Prog Neuropsychopharmacol Biol Psychiatry* 2004, 28: 57-65.
18. Lowe B, Schenkel I, Bair MJ, Gobel C: Efficacy, predictors of therapy response, and safety of sertraline in routine clinical practice: prospective, open-label, non-interventional postmarketing surveillance study in 1878 patients. *J Affect Disord* 2005, 87: 271-279.
19. 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.
20. Jarrett RB, Eaves GG, Grannemann BD, Rush AJ: Clinical, cognitive, and demographic predictors of response to cognitive therapy for depression: a preliminary report. *Psychiatry Res* 1991, 37: 245-260.

21. Sotsky SM, Glass DR, Shea MT, Pilkonis PA, Collins JF, Elkin I *et al*.: Patient predictors of response to psychotherapy and pharmacotherapy: findings in the NIMH Treatment of Depression Collaborative Research Program. *Am J Psychiatry* 1991, 148: 997-1008.
22. Thase ME, Reynolds CF, III, Frank E, Simons AD, McGeary J, Fasiczka AL *et al*.: Do depressed men and women respond similarly to cognitive behavior therapy? *Am J Psychiatry* 1994, 151: 500-505.
23. Hatcher S: The STAR*D trial: the 300 lb gorilla is in the room, but does it block all the light? *Evid Based Ment Health* 2008, 11: 97-99.
24. Percevic R, Lambert MJ, Kordy H: Computer-supported monitoring of patient treatment response. *J Clin Psychol* 2004, 60: 285-299.
25. Lem Rvd, Stamsnieder P, Wee Nvd, Veen Tv, Zitman FG (Eds):Socio-demographic features in randomized controlled trials for major depression: generalizability and individualization. In *Int J Person Cent Medicine* 2011, 1: 268-278.
26. 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.
27. 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.
28. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
29. Altman DG: Practical Statistics for Medical Research. 1991:229-276.
30. Donders AR, van der Heijden GJ, Stijnen T, Moons KG: Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006, 59: 1087-1091.
31. Royston P: Multiple imputation of missing values: update. *Stata Journal* 2005, 5: 188-201.
32. Hollon SD, Derubeis RJ, Evans MD, Wiemer MJ, Garvey MJ, Grove WM *et al*.: Cognitive therapy and pharmacotherapy for depression. Singly and in combination. *Arch Gen Psychiatry* 1992, 49: 774-781.
33. Falconnier L: Socioeconomic status in the treatment of depression. *Am J Orthopsychiatry* 2009, 79: 148-158.
34. Goekoop JG, Hoeksema T, Knoppert-Van der Klein EA, Klinkhamer RA, Van Gaalen HA, Van Londen L *et al*.: Multidimensional ordering of psychopathology. A factor-analytic study using the Comprehensive Psychopathological Rating Scale. *Acta Psychiatr Scand* 1992, 86: 306-312.
35. Van HL, Schoevers RA, Dekker J: Predicting the outcome of antidepressants and psychotherapy for depression: a qualitative, systematic review. *Harv Rev Psychiatry* 2008, 16: 225-234.
36. Butterworth P, Leach L.S., Pirkis J., Kelaher M: Poor mental health influences risk and duration of unemployment: a prospective study. *Soc Psychiatry Psychiatr Epidemiol* 2012, 47: 1013-1021.
37. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
38. van der Lem R, de Wever WW, van der Wee NJ, van VT, Cuijpers P, Zitman FG: The generalizability of psychotherapy efficacy trials in major depressive disorder: an analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry* 2012, 12: 192.

SUPPLEMENTARY MATERIAL

Table 2. Comparison of all patients who had baseline assessment (follow-up: treatment + follow-up assessment; lost to follow-up: drop out or treatment + no follow up assessment).

	Baseline severity of depression (MADRS)	**Age (years)**	**Gender (% female)**	**Race (% white)**	**Marital status (% marrried)**	**Employment I (% paid work)**	**EmploymentII (% paid work or sickness benefit)**
Selected Patientgroup	25.8	39.2	66.3%	82.0%	52.6%	34.1%	74.4%
Lost to follow-up	25.6	37.3	66.6%	80.5%	46.9%	38.8%	70.8%
Statistical Comparison	CI of difference: -090–0.63; p=0.73	CI of difference: 0.71–3.09; **p=0.002**	X^2=0.02 Df1 p=0.89	X^2=0.65 Df2 p=0.72	X^2=3.47 Df1 **p=0.05**	X^2=2.82 Df1 p=0.09	X^2=2.01 Df 1 p=0.16

For 82 of the selected patients no information was available on race, marital status, and employment status. For 291 patients in the lost-to follow up group no information was available on race, marital status, and employment status.

Table 3. Comparison of patients who had treatment+ follow-up assessment and patient who had treatment without follow-up assessment.

	Baseline severity of depression (MADRS)	**Age (years)**	**Gender (% female)**	**Race (% white)**	**Marital status (% marrried)**	**Employment I (% paid work)**	**EmploymentII (% paid work or sickness benefit)**
Selected Patientgroup	25.8	39.2	66.3%	82.0%	52.6%	34.1%	74.4%
Lost to follow-up	25.4	37.7	68.1%	80.6%	47.2%	38.5%	71.8%
Statistical Comparison	CI of difference: -1.19–0.43 p=0.35	CI of difference: 0.30–2.83 **p=0.015**	X^2=0.49 Df1 p=0.48	X^2=0.29 Df2 p=0.86	X^2=3.09 Df1 p=0.08	X^2=2.15 Df1 p=0.14	X^2=0.96 Df1 p=0.33

For 82 of the selected patients no information was available on race, marital status, and employment status. For 153 patients in the lost-to follow up group no information was available on race, marital status, and employment status.
CI: confidence interval. Df: degrees of freedom. X^2: chi square

Chapter 7

Summary and General Discussion

SUMMARY OF OUR FINDINGS

In today's psychiatric practice in Western societies, most mental health care institutions have implemented treatment algorithms or guidelines for the treatment of major depressive disorder (MDD). These treatment algorithms/guidelines are all based on results from randomized clinical trials (RCTs), also called efficacy trials. In daily psychiatric practice, many clinicians have the impression that results found in trials (efficacy) are better than the results of the same therapies in routine care (effectiveness). In this thesis, we investigated whether the clinicians' impression that efficacy is higher than effectiveness is correct and, if the impression is substantiated, which factors explain the difference. Criticism about the generalizability of results from RCTs to daily practice has often been heard. Clinicians believe that the possible difference in effect is explained by differences between their patients and participants in MDD trials. Clinicians are supported by previous research on the generalizability of results from MDD trials. It has been shown that only a minority of "real life" patients is eligible for participation in RCTs, because of the stringent criteria for patient selection [1-3] and perhaps also because of (un)intended selection due to the methodology of recruitment of participants in trials [4]. The STAR*D trial [5] found that participants who were eligible for "classical" MDD trials had a beneficial outcome compared to participants who were not [6]. The STAR*D trial, with very broad inclusion criteria, has many similarities with daily practice. However STAR*D also has characteristics of an RCT, like the use of a baseline severity threshold for inclusion, no possibilities for patient preferences in the first treatment step, and a large investment in treatment adherence of both therapists and participants. Despite the broadness of the inclusion criteria, it is possible that the RCT characteristics of STAR*D may still have limited its generalizability to daily practice. For this project we derived our data directly from daily practice through Routine Outcome Monitoring (ROM). By doing so, we were able to investigate whether clinicians are right when they state that treatment outcome in daily practice is less hopeful than in efficacy trials. Subsequently, we were curious to see whether the evidence from the STAR*D trial of a better treatment outcome in "RCT-eligible" patients could be replicated. If clinicians would, hypothetically, exclude all of their non-eligible patients, would their treatment results improve?

In order to assess whether effectiveness is really lower than efficacy (chapter 2), we compared the within group efficacy reported in fifteen meta analyses on three types of MDD treatment; antidepressants, individual psychotherapy and a combination of both, with the effectiveness of the same treatments in daily practice, measured by ROM. A meta analysis provides an aggregated estimate of results found in RCTs. Meta analyses of RCTs are most often carried out to investigate whether the active drug/psychotherapy is superior to placebo (which is called the *between group* efficacy). However, we were not interested in this relative effect of active drug/psychotherapy, but in their overall or absolute effect (which is called the *within-group* efficacy). We compared this overall efficacy with the effectiveness

in "reality". Our overall conclusion in chapter 2 is that the impression of clinicians, that treatments in "reality" are not as effective as in scientific research, is true:

> ***Effectiveness of MDD treatment in daily practice is lower than efficacy results from RCTs on MDD treatment. This is the case for antidepressant treatment, individual psychotherapy as well as combination treatment.***

Above we mentioned that clinicians attribute the smaller treatment effects in "real life" to the fact that only a selection of patients is allowed to participate in RCTs. To investigate this, we first made an inventory of the exclusion criteria used in RCTs. Next we studied how many patients would have to be excluded if these criteria were applied in clinical practice, and then we compared treatment effectiveness in "real life" patients who meet the selection criteria versus patients who do not.

For inclusion in an MDD efficacy trial, in antidepressant efficacy trials (AETs) as well as in psychotherapy efficacy trials (PETs), participants indeed have to meet a set of eligibility criteria (in and exclusion criteria). These eligibility criteria are necessary to optimize the internal validity of the trial. In AETs, there is consistency in the use of exclusion criteria [7,8]. The most commonly used exclusion criteria in AETs are: not meeting a baseline severity threshold of 18 on HAMD17 [9]; co morbid Axis I disorders; co morbid Axis II disorders (in particular borderline personality disorder); suicidality and co morbid substance abuse. In the literature it is reported that only a minority of MDD patients from fee-for-service practices are eligible for AETs [1,2]. In this thesis, we investigated whether in the Netherlands in routine care also only a minority of patients with MDD would be eligible for AETs. We planned to do the same for PETs, yet studies on patient selection in PETs were absent. Therefore, we first had to investigate which exclusion criteria were used in a large set of PETs (chapter 4). We found that the following exclusion criteria were frequently used in PETs: not meeting a baseline severity threshold of 14 on HAMD17 [9]; co morbid substance abuse and antidepressant treatment prior to participation.

The next step was to apply the exclusion criteria of AETs (chapter 3) and PETs (chapter 4) to a "real life" (ROM) population (in which no selection takes place besides sufficient mastery of the Dutch language to complete the ROM questionnaires). We used a large dataset of MDD patients who sought treatment in Rivierduinen, a large regional mental health provider (RMHP). We found that clinicians are right when they state that their patients are very different from RCT participants.

- ***"Real life" MDD patients often do not meet the baseline severity threshold (42% for AET threshold and 22% for PET threshold).***

- ***"Real life" MDD patients do report suicidality (15%).***
- ***"Real life" MDD patients do have Axis I (63%) and Axis II co morbidity (7% borderline personality disorder).***
- ***"Real life" MDD patients do have co morbid substance abuse (9%).***
- ***"Real life" MDD patients do have often used antidepressants prior to referral to psychotherapeutic treatment (44%).***

Apart from selection based on explicit criteria also implicit selection may be important in RCTs, for instance with respect to sociodemographic and socioeconomic (SES) features. In chapter 5 we studied which SES features were reported in AETs and PETs. It became clear that educational level, socioeconomic status and income were reported insufficiently. However, for some features (age, gender, ethnicity, marital and employment status) enough data were available to enable comparison with "real life" patients (chapter 6). Our most striking finding was:

"Real life" MDD patients significantly less often have a paid job at time of treatment than RCT participants.

Having identified criteria that play a role in the selection of patients for RCTs and having demonstrated that application of these criteria in "real life" would indeed exclude a large group of patients from treatment, the next question is whether "real life" patients who are eligible for RCTs are doing better in treatment than "real life" patients who would be excluded. We found that exclusion of patients with mild depression, patients who used antidepressants prior to psychotherapy or patients without a paid job, improved treatment outcome in the remaining patient group, but only in a modest way. Besides, Axis I and Axis II co morbid disorders, substance abuse and suicidality were not associated with treatment outcome in our MDD patients. Furthermore the extent to which the difference in treatment outcome between RCTs and "real life" can be attributed to patient selection based on exclusion criteria is very small (explained variances 1–4% for the AET criteria; 4–11% for the PET criteria, dependent on definition of outcome). The same accounts for implicit selection based on the SES features age, gender, ethnicity, marital and employment status (explained variance 3–7%).

Therefore, our most striking overall finding was that:

In our "real life" patients, being eligible (meeting all criteria) for RCTs was not associated with a better treatment outcome.

GENERAL DISCUSSION

We found that if only RCT eligible patients were treated in daily practice and non eligible patients would be excluded, the treatment success in daily practice would not improve. So….yes, clinicians àre right that their MDD treatment results are less favorable than those from efficacy trials, and yes, they àre right that their patients differ very much from RCT participants due to the use of stringent exclusion criteria and (un)intended sociodemographic/socioeconomic patient selection by recruitment procedures. However…. the use of exclusion criteria and the selection of patients with a different socioeconomic status in RCTs do not explain the difference between efficacy and effectiveness. So, it might be that the items from patient selection that we analyzed are not the major threat to the generalizability of the results from MDD trials to daily practice as has been suggested in the past. In the next paragraphs, we will elaborate on the implications of our findings for clinical practice and the scientific field. We will seek further explanations for the difference between efficacy and effectiveness of MDD treatment. Although the effect of the use of exclusion criteria was modest on treatment outcome, we will however comment on the implications of our findings that patients suffering from minor depression, as well as patients who used antidepressants prior to their psychotherapy seem to benefit less from their treatment. We will also discuss the implications of our finding that patients without a paid job have a less favorable treatment outcome. Finally, we will discuss the limitations of our project and will conclude with recommendations both for future research and clinical practice.

Why do efficacy and effectiveness differ?

In this thesis, we found evidence for the assumption of clinicians that treatment results in daily practice are disappointing compared to those in MDD trials (of the same therapies). Of MDD patients in antidepressant trials, 34–47% reaches remission, whereas in daily practice only 21% of the patients are that fortunate after the first treatment step. For individual psychotherapy (cognitive behavioral therapy or interpersonal therapy), 34–58% of trial participants reach remission, while in daily practice only 27% of the patients reach remission after the first treatment step. Patients who receive combination therapy in daily practice reach remission in 21% of the cases, while in trials 45-63% do. We have shown that (un) intentional patient selection based on exclusion criteria or on socioeconomic grounds does not explain the difference between efficacy and effectiveness. Then, what does*? Does Dr. X, introduced in the Introduction section of this thesis, turn out to be a lousy therapist? Is the faith of younger colleagues in his knowledge and experience misplaced? Or do we have to look for other explanations for the difference between efficacy and effectiveness?*

Dr. X, now getting worried, provokingly states that the disappointing outcome results that we found in this thesis are typical for the RMHP Rivierduinen, for the Leiden area, or for Dutch psychiatry. Is he right? The effectiveness that we found is in line with the results of STAR*D

[5,10,11], which suggests that the modest treatment results are not typically from the RMHP Rivierduinen or Dutch psychiatric practice. The similarity of our results to those from STAR*D is a notable finding. As mentioned at the start of this chapter, STAR*D is a pragmatic trial, with on the one hand methodological characteristics of RCTs, but on the other hand resemblances with routine psychiatric practice. Much effort was put in treatment adherence and motivation, both at the side of patients and of clinicians. This may have inflated the treatment success. On the other hand, following patient's or doctor's preferences for a specific drug or psychotherapy, as is usual in daily clinical practice, was not allowed in the first treatment step in STAR*D. As the allowance of preference is associated with better treatment outcome [12-15], this might have diminished treatment success in STAR*D. In our population, no special effort was made to improve treatment adherence besides care as usual. Therefore, based on treatment adherence alone, one would probably expect less favorable treatment outcome in the ROM population than in STAR*D. On the other hand, in our daily practice population patient and doctor's preferences were of course allowed, which may have raised our treatment effect compared to STAR*D. These two factors together may have contributed to similar results in STAR*D and our ROM data. More emphasis on the improvement of treatment adherence, as is done in RCTs as well as in STAR*D, may improve the treatment results of daily practice. Furthermore, many other factors may contribute to the differences between efficacy and effectiveness. They may be features of patients, therapists, setting or RCT methodology. We will discuss them one by one in the following paragraph.

Patient features

Today, Dr. X's first patient is Ms. Y. Ms. Y is a moderately severe depressed and traumatized single, middle-aged woman who just lost her job and whose cat just died. Counseling sessions in a private practice and antidepressant treatment by her general practitioner did not improve her mood. She is somewhat sceptical about her referral to dr. X but is determined to give it a try and tell dr. X about all her problems in the first session. Right before his busy clinic starts, Dr. X quickly opens his mailbox. In his mailbox is an enthusiastic letter from a young colleague working in an academic center who asks psychiatrists to send in patients for a promising trial with a specific drug. What are the chances that Ms. Y will be willing to participate in this trial?

Likely, there are differences between RCT participants and daily practice patients, which we did not explore. For instance, participants in RCTs probably are a subgroup with a special motivation: they are willing to take the risk to be treated with a placebo. It is yet unknown which other specific characteristics this subgroup has and whether these characteristics may contribute to treatment outcome. Recruitment procedures might introduce (un) intentional selection bias by recruiting patients with a prognosis that differs from the "real life" population. Clinicians might not send patients with a poor prognosis for participation

to trials or these patients might not be motivated to participate. More importantly, in the Netherlands, and in other countries with a stepped health care system where the general practitioners (GPs) have the function of gatekeeper, many MDD patients with a good prognosis will be (successfully) treated by their GP or in private practice (so called first line treatment) and will not be referred to the RMHPs. Consequently, RMHPs only treat patient populations with a poorer prognosis. RCTs probably recruit MDD patients from both the GP population with a good prognosis and from the RMHP population (with a poorer prognosis) and in the "worst" case only from the first line population. The overall prognosis of RCT participants is therefore probably better than of RMHP patients. In order to optimize the generalizability of results from RCTs to "real life" (RMHP) psychiatric practice, it would be recommendable to conduct trials which include only patients who already went through GP or private practice treatment.

Therapist features

On a regular Friday, six a clock in the afternoon, Dr X. leaves his institution. It has been a busy week; at least twenty patients a day, staff meetings, resident supervision, two patients in severe crisis, an absent colleague who will probably be ill for a longer period, and a deadline for a report on a patient who had a complaint about his treatment. Dr X. cannot deny his feeling of tiredness and he starts to look forward to the moment that he will retire. Meanwhile, he feels a not-severe-but-nevertheless-nasty flu coming up. In his briefcase he has a brochure of a new and promising trial for MDD patients. On the cover of the brochure there is a smiling physician in a crispy white coat, who seems to be half the age of Dr. X.

Therapists who participate in trials might differ from daily practice clinicians in terms of workload, motivation, extent of updating training, and many other aspects. While for trial therapists the proper conductance of the treatment under investigation is their main goal, so to speak "real life" clinicians can be distracted by many other tasks than state of the art treatment of MDD patients. For instance, "real life" clinicians may have very limited time per patient as a result of a caseload that is too large. Furthermore they often have to stand in for absent colleagues, perform instant assessments of so called crisis patients, and sometimes also have managerial tasks. And all this in between their therapies for MDD patients…Furthermore, clinicians are probably very dedicated to their patients, yet perhaps headstrong when it comes to strictly following the protocol described in the treatment guidelines. Therapists who participate in trials might be more motivated for the treatment under investigation than "real life" clinicians. Maybe it is even so that especially highly skilled or specialized clinicians participate in trials. All these factors contribute to differences between RCT therapists and "real life" clinicians. Motivation, protocol adherence, extent of education, time per patient, and experienced workload are all factors likely to be associated

with treatment outcome. So, if *patient selection* is not the (only) answer to the difference between efficacy and effectiveness, *therapist selection* may well be one of them.

Differences between trial setting and daily practice setting

As a response to a shimmering trial brochure that calls for participants, Dr X. sends in Mr. Z. for participation. Mr. Z. is a 45 year old patient who suffers from MDD. He is a little bit sceptical about the results of antidepressant therapy, since his cousin and his neighbor did not improve on this medication. The trial therapist convinces Mr. Z. that the trial antidepressant is very new and promising. He explains the procedure of the trial to Mr. Z. and tells him that he will have a chance to either receive this new drug or a placebo. Mr. Z. is persuaded to participate. Without knowing (a double blind procedure) Mr. Z receives a placebo. The inspired trial therapist sees Mr. Z. every week during the follow-up time of the trial, which is 8-12 weeks. After 3 months, treatment results are assessed and the trial therapist says goodbye to Mr. Z. He thanks Mr. Z. for his willingness to participate and for his contribution to the development of treatment of MDD.

Every treatment has a placebo effect: the mere fact that the patient is receiving treatment has a beneficial effect. The aim of RCTs is to prove that the active drug under consideration has a significantly larger effect than a placebo (the between group efficacy, see above). In clinical practice the placebo effect also contributes to the overall treatment effect. It is likely, that the placebo effect in trials is larger than in daily practice. We will provide some arguments why this might be the case:

A proportion of participants in RCTs will spontaneously recover (like in daily practice) during participation. Spontaneous recovery will augment the proportion of patients who reach remission in a trial, while this effect cannot be attributed to the investigated treatment. In one meta analysis, spontaneous recovery was estimated to constitute one third of the placebo effect [16]. In daily practice, patients who recover spontaneously will probably not enter treatment or will drop out prematurely. They will not enter a ROM follow up assessment and therefore do not contribute to treatment outcome in ROM. Furthermore, participants in trials (and clinicians as well) have the feeling that they are treated in a special, new and promising way. This belief might contribute to improvement in RCTs and is called the Hawthorne effect [17,18]. Finally, as discussed above, in trials much effort is put in optimizing both patient's and clinician's protocol adherence. Protocol (or guideline) adherence seems to be positively associated with treatment success [19-25]. One specific aspect of protocol adherence is the frequency of follow up visits. In RCTs, frequency of appointments is closely monitored, while in daily practice appointments are sometimes cancelled by the clinician or patients for reasons of illness or otherwise. As a result of that, patients in RCTs have more regular and more frequent follow up visits. In a meta analysis on the therapeutic effect of follow up assessments in AETs, it was found that extra follow

up visits were associated with better treatment outcome and that the therapeutic effect of follow up assessments represents about 40% of the placebo response in AETs [26].

Study features

Therapist A conducted a trial with a new type of psychotherapy. She was very enthusiastic about this type of treatment, but unfortunately after a lot of effort it turns out that the results are disappointing. The effect of the new psychotherapy was comparable to that of treatment as usual. What are the chances that this therapist will lose her motivation and the results will end up in her top drawer? And if not, what are the chances that her negative results will be published in a prominent psychotherapy journal?

Negative findings are reported a lot less often than positive findings. This is the so called *publication bias* [27]. Nowadays all medication trials have to be made available in public registers that are available to everyone (e.g. Nederland's Trial Register) ahead of the start. However, treatment guidelines are based on articles published in scientific journals. Due to publication bias, efficacy may be overestimated. This may partially explain the difference between efficacy and effectiveness. Recently, several methodologies have been developed for meta-analyses in order to adjust to some extent for publication bias. It also has been suggested that the efficacy in MDD trials is exaggerated due to so called *rater bias*. The severity of MDD might be somewhat inflated by participating therapists at the beginning of the trial. At the same time, severity rating of the depression might be somewhat deflated at the end of the trial. If so, treatment success of trials (which can be the pre-post treatment difference) might be exaggerated and thus contribute to the difference between efficacy and effectiveness. However the extent of rater bias in MDD trials is still unknown. In one study rater bias was found to occur in MDD trials, yet its extent was too small to invalidate the results of the trials [28].

Minor depression, prior antidepressant use and having a paid job: implications of our findings

Although modest, we found that the exclusion of patients with mild depression, patients who used antidepressants prior to psychotherapy and patients without a paid job, improved treatment outcome in the remaining patient group. Although our most striking overall finding was a negative (absence of association) one: "In our "real life" patients, being eligible (meeting all criteria) for RCTs was *not* associated with a better treatment outcome", we didn't want to leave our positive findings undiscussed. The influence of exclusion of patients who do not meet the baseline severity threshold, who use antidepressants prior to their psychotherapy and who do not have a paid job is described in detail in the chapters 3, 4 and 6. Below we summarize our main findings.

We found that exclusion of ROM patients who suffered from minor depression (baseline severity of less than 18 on HAMD17) lead to a larger proportion of patients who reach remission (OR 2.0; 95% confidence interval 1.3–3.1). This association was found for psychotherapy, antidepressant treatment and a combination of both. As mentioned before, AETs often use a baseline severity threshold of HAMD17 ≤18 as exclusion criterion. We also found that exclusion of patients who have a baseline severity less than 14 on the HAMD17 (the threshold used in PETs) lead to more improvement in the remaining patients (β=7.23; 95% confidence interval 5.31–9.11).

We found that mild to moderate depression is very common in routine clinical practice (42% of the patients do not meet the AET severity threshold and 22% do not meet the PET severity threshold). Why this specific group rarely reach remission in their first treatment step is still unclear. Maybe these patients more often have a chronic mild depression instead of episodes of more severe MDD, and therefore have a different prognosis. It is also possible that these patients have other traits that differ from more severe MDD patients, such as lack of optimism as a personality trait. Future research is recommended on the characteristics of the large group of "real life" patients suffering from minor depression. To what extent the results of RCTs are generalizable to this group also needs to be further explored.

In addition, we found that exclusion of patients who used antidepressants prior to psychotherapy enlarges the extent of improvement of PETs (β=7.62; 95% confidence interval 1.94–13.30). These patients probably do not or only partially respond to medication, often prescribed by their GP or in private practice. As mentioned above in this chapter, these patients might have a worse treatment prognosis, than the ones who did not go through another treatment prior to their psychotherapy. Our finding accentuates that it would be recommendable to conduct trials which include only patients who already went through GP or private practice treatment, in order to optimize the generalizability of results from RCTs to "real life" (RMHP) psychiatric practice.

We compared demographic characteristics of the groups. We found a substantial difference in the proportion of patients employed at time of participation. 68% of the RCT participants had a paid job, while only 34% of the ROM patients were working at the time of treatment. ROM patients who were working had better treatment outcome than patients who were not, irrespective of the baseline severity of their depression (OR 1.76; 95% confidence interval 1.2–2.6 for the proportion of MDD patients who respond and OR 1.85; 95% confidence interval 1.2–2.8 for the proportion of patients who reach remission). In chapter 6 we showed that having financial security is probably not the aspect of having a job that contributes to treatment success. We recommend further research on which aspects of employment contribute to treatment outcome of MDD patients. The results of this future research can be used in the development of new MDD treatments or improvement of the existing ones by increasing the attention for the role of social factors in MDD treatment.

Limitations of our project

There were three major limitations to our research: Firstly, the lack of consistency in efficacy trials with respect to type of instruments, definition of outcome and use of exclusion criteria. Secondly, the missing data and large loss to follow up in the ROM data that is inherent to research in clinical practice. Lack of routinely collected treatment information and data on life history in the ROM data forced us to rely on data from extensive charts review. Thirdly, although our results seem to be representative for "real life" MDD patients, there are also limitations in the generalizability of our results. In the next paragraphs, we will first discuss the implications of the lack of consistency in RCTs, discuss the limitations of working with ROM data, and finally we will critically review the limits of the generalizability of our results.

Research on research: limitations in estimating efficacy

We investigated the efficacy of antidepressant treatment, individual psychotherapy and combination treatment. We studied the estimation of efficacy in RCTs and found an inconsistency in the use of instruments to assess depression severity. We also found an inconsistency in the definition of outcome: response is consistently defined as a 50% reduction of symptoms, but remission is defined by different cut off scores. Furthermore, we found that PETs are inconsistent in their use of exclusion criteria. These inconsistencies in the underlying data might compromise the validity of the aggregated efficacy estimates that are given in meta-analyses.

In addition, AETs and PETs have a different manner in evaluating treatment outcome, due to a different research tradition. The difference of defining outcome between AETs and PETs did not hinder our analysis, but it somewhat diminished the comprehensiveness of our results for clinicians, since we had to compute outcome in line with AETs as well as PETs. In table 1, we provide an overview of the instruments and definitions of outcome used in the meta- analyses included in our study. In the frame we describe the inconsistencies in instruments and outcome definition and their implications for our results in detail. The comparability of efficacy estimates, in meta- analyses but also in the comparison with "real life" cohorts would benefit greatly from more consensus on the instruments and the eligibility criteria for AETs and PETs. Finally, within our selection of AETs and PETs, for the exploration of eligibility criteria in PETs (chapter 4) and the reporting of sociodemographic/ socioeconomic features in AETs as well as PETs (chapter 5), many more AETs were available than PETs. For AETs we therefore limited our search to high impact journals, while we included all PETs within the same time frame. Although our selected AETs and PETs were similar with respect to countries of origin and timeframe, there is a slight possibility that our methodology of RCT selection has introduced some selection bias.

Table 1. Instruments and definitions of outcome in meta analyses.

		Type of meta-analysis	RCTs that used the following instruments were included	Definition of outcome: Response	Definition of outcome: Remission	Definition of outcome: Effectsize
1	Kasper 1997	AETs	HAMD (17 item version)	50% reduction	HAMD ≤7	-
2	Bech 2000	AETs	HAMD (17 item version)	50% reduction	-	-
3	Storosum 2001	AETs	HAMD (both 17 and 21 item version)	50% reduction	-	-
4	Steffens 1997	AETs	HAMD (version not specified)	50% reduction	-	-
5	Montgomery 2001	AETs	HAMD (17 item version)	50% reduction	HAMD ≤8	-
6	Beasley 2000	AETs	HAMD (17 item version)	50% reduction	HAMD ≤7	-
7	Einarson 1999	AETs	HAMD (version not specified) MADRS	50% reduction	-	-
8	Stahl 2002	AETs	HAMD (21 item version) MADRS	50% reduction	-	-
9	Nelson 1999	AETs	HAMD (version not specified) MADRS	50% reduction	-	-
10	Thase 2001	AETs	HAMD (both 17 and 21 item version)/ MADRS	50% reduction	HAMD17 ≤7; HAMD21≤7/≤8/10; HAMD17≤10 +CGI=1, MADRS<10	-
11	Thase 1997	PETs COMs	HAMD (17 item version)	-	HAMD<7	-
12	De Maat 2007	PETs	HAMD BDI	-	HAMD <6/<7/<8; BDI <9/<10	-
13	Minami, 2007	PETs	BDI	-	-	Δ Mean BDI pre-posttreatment / SD pre-treatment
14	Thase, 2005	AETs	HAMD (21 item version)	-	≤7 on first 17 items on HAMD21	-
15	Wexler, 1992	COMs	BDI	-	BDI≤16/23; Raskin[2] ≤9	

AET: antidepressant efficacy trial. PET: psychotherapy efficacy trial. COM: combination treatment trial; antidepressants + individual psychotherapy. RCT: randomized controlled trial. HAMD: Hamilton Depression Rating Scale. MADRS: Montgomery Asberg Depression Rating Scale. BDI: Beck Depression Inventory. CGI: Clinical Global Impression scale Raskin: Raskin Depression Scale

Cicchetti DV, Prusoff BA: Reliability of depression and associated clinical symptoms. Arch Gen Psychiatry 1983, 40: 987-990.

Δ: difference pre-post treatment. SD: standard deviation. - : Definition of outcome is not used in meta-analysis.

Inconsistencies in the use of instruments in RCTs

In AETs, the most commonly used severity scale is the HAMD [9], especially in trials from the United States. The MADRS [29] is also often used in AETs, especially in European trials, sometimes as primary outcome measurement, often as secondary instrument. In our selection of meta analyses on AETs, they all used the HAMD. Yet, two different versions of the HAMD (17 and 21 items) are used. Both versions are validated, but how a cut off score for remission on one version relates to a cut off score on the other version is not clear. In ROM the MADRS is used. The fact that we had to convert the MADRS to HAMD scores in all our analyses, might have influenced our results on the efficacy-effectiveness difference. In order to give the most reliable estimate of HAMD scores, we used three equations [30,35,36] to convert our MADRS scores. We found that two equations yielded the similar results [30,36] and we performed a validity check with another method for conversion: the Item Response Theory [32]which also yielded similar results. We therefore expect little limitations to our analyses due to the fact that HAMD is not used in ROM.

Inconsistencies in the use of cut offs for remission in RCTs

All meta analyses on AETs used the same definition of response and therefore we did not encounter difficulties in the efficacy–effectiveness comparison. However, most patients in ROM suffered from mild to moderate depression, which lead to very similar proportions of response and remission. Therefore, we did not report separately on the efficacy-effectiveness difference for response (chapter 2). The definition of remission varies between AETs and some meta analyses include trials with different definitions of remission. In our selection of meta analyses four different cut off scores to define remission were used. For the computation of the effectiveness of MDD treatment in "real life" we used a stringent (and scientifically investigated) cut off of MADRS ≤10 [30], which equals a score of 6.4 on the HAMD17 [31,32]. By using this stringent cut off score, it might be that we were too harsh in estimating the efficacy-effectiveness difference. In reality the efficacy-effectiveness difference might be a little smaller, especially for the efficacy-effectiveness difference in meta analyses that included only trials that used a less stringent cut off [33,34]. However, the most often used cut off score for remission (HAMD17 score of 7) in the meta analyses is, to our opinion, close enough to our definition of remission in "real life" (MADRS ≤10 ≈ HAMD17≤6.4)) to give a reliable estimate of the difference between efficacy and effectiveness.

Inconsistencies in the use of exclusion criteria in PETs

We found that PETs are not consistent in the exclusion criteria they use. Only 4 of the 38 criteria were used in 75% of the papers (chapter 4). This, of course, hampers the comparability of PETs and thus the reliability of meta analyses of PETs (and the comparability of PETs with AETs). It also has consequences for the interpretation of the results described in chapter 4. Firstly, calculating the overall efficacy of PETs as is done in three of our selected meta analyses [37-39] while the comparability of PETs is low, raises questions about the reliability of the results from these meta analyses. Therefore, the reliability of our results on the efficacy-effectiveness difference might likewise be jeopardized. Secondly, it was impossible to take all exclusion criteria into account when we investigated which "real life" patients would have been eligible for PETs. We restricted ourselves to the four most consistently used criteria, making the comparison of treatment effects in eligible and non-eligible patients just an approximation.

Research on ROM data: limitations in estimating effectiveness

The ROM data were gathered in clinical practice, as part of the routine diagnostic and treatment processes. Although such data have the advantage of offering insight into the vicissitudes of "real life" patients, they also have limitations just because of these vicissitudes. First of all, data integrity is not guaranteed. By using computers with touch screens and software that makes it impossible to skip a question in a questionnaire and by having test nurses supervising the filling out of the questionnaires, we tried to make the data as complete as possible. However, it was clinical practice, not a research project in which double checking of data and data gathering are the standard procedure. Thus incompleteness was inevitable. Also the large number of questionnaires may have impeded completeness. We addressed the problem of missing data as good as possible by using elaborate statistical methods (MICE, multivariate imputation by chained equations, [40]). Second, in the period in which the data for our project were gathered, the follow up assessments in ROM were not organized properly. The consequence is an almost 50% loss to follow-up. In the relevant chapters of this thesis we discussed how we tried to handle this loss. On the other hand, a large loss to follow up may be inherent to studies with a naturalistic design: STAR*D had reached a loss-to-follow-up of 48% in step II of the study. Third, in ROM data on the history of the patient's life and his illnesses are rudimentary. Unfortunately, as those data are also not available in a useful digital format, we had to depend on an extensive chart review. All these factors will have reduced the reliability of the data. However, they are more extensive and relate to a larger number of patients than in any other project. Therefore, we felt that our data are a significant contribution to this new field of research.

Generalizability of our results: limitations

In this thesis, we explored the outcome of antidepressant and psychotherapeutic treatment of MDD from baseline assessment to the first follow up assessment. We have not addressed patient selection and its influence on outcome of RCTs on combination treatment, as combination treatment is a second step in the treatment algorithm of MDD.

As mentioned earlier, all ROM patients at the RMHP are referred by either their GP (most often) or by a psychiatrist working in private practice. Many patients already underwent treatment for their MDD prior to referral. Our results are therefore only generalizable to outpatient clinics that treat similar patients. The generalizability of our results to private practices, GP practices and mental health providers who treat only or merely patients that are treatment naïve (and who probably have a more favorable prognosis) or patients who are non responders to several therapies (so called third line institutions) is most likely limited.

The meta-analyses that we used in this project were carried out in the United States of America and Europe. These studies included a predominantly white patient population. Our ROM population also is a predominantly white patient population treated in Western psychiatric practice. We do not know whether our results are also valid in other cultures. Neither do we know to what extent they apply to non-Western immigrants in the Netherlands who were unable to fill out the questionnaires.

Future directions in effectiveness research and opportunities for clinical innovation

In this last section of the discussion we will present some recommendations for future research in line with our project and also for clinical development.

We will start with recommendations for future research.

- As described earlier, our loss to follow-up was considerable. From personal communications with other centers using ROM it is clear that this is a nearly universal problem. The large loss to follow up in the STAR*D trial also emphasizes the problem of loss to follow up in research done in clinical practice. Of course, the loss to follow up could be decreased by a better organization. Probably the covenant with the insurance companies to increase the proportion of patients with follow up data may help. However, the high loss to follow up should also become a focus of research. Almost one third of these lost to follow up patients remain in treatment, so do these patients refuse to participate in ROM or do clinicians forget to sign up their patients? Future research will have to focus on reasons why patients do not participate in follow up assessments. And the other two thirds of the patients? Did they recover and then disappear? Or were they unsatisfied with their treatment and no longer showed up? It is remarkable that the urge to investigate these topics is not felt widely. Perhaps patients who are lost to follow up have specific features such as a common social background, more co morbid disorders or specific personality traits. From our lost to follow up analysis, we learned

that although the patients who were lost to follow up were very similar to the ones who were not, especially single male MDD patients suffering from co morbid post traumatic stress disorder were at risk of dropping out and being lost to follow up. Future research might reveal specific subgroups that are at risk for drop out or loss to follow up and need a specific approach to stay in treatment and have a proper evaluation of it. Also more research is needed on the side of the therapists: are there specific professional groups that do not support ROM? And what do they need to feel the need for routinely systematic evaluation of their treatment?

- Research on treatment effectiveness and benchmarking requires large databases. Therefore, it is important that the ROM of mental health care centers use, as far as possible, the same questionnaires and procedures. In the current financial crisis, many policy makers need/tempt to make stringent cutbacks in the budgets of mental health care. One way to reduce the costs of ROM is to reduce the number of instruments in ROM as much as possible. This, however, may seriously jeopardize the usefulness of ROM data as a reliable instrument for the evaluation of treatment progress in clinical practice. Furthermore, it certainly jeopardizes the usefulness of ROM data for scientific research. A discussion about what the necessary ingredients of ROM are, is necessary. The data of the Leiden Routine Outcome Monitoring Study may be helpful to provide this discussion with data, i.e. by the exploration of the validity of key items in the available instruments and the possibilities of answer-steered exposure to new items of questionnaires (patients do not fill out complete questionnaires, but will get new items based on their response to the former ones).
- Further research on the influence of factors in which AETs and PETs on one hand and "real life" patient cohorts on the other hand differ, should be continued. More specifically, data not included in this study, for instance on earlier treatments and patient history, should be included. Also, then, replication studies on our findings can be carried out, preferably in real life cohorts with more complete data and less loss to follow-up.
- We investigated MDD. It would be useful to extend this type of research to other disorders, for instance anxiety disorders. Such research would elaborate which problems are unique for MDD and which are general.

Is our finding of modest effects of the first step in evidence based MDD treatments a reason to discard the guidelines, throw away evidence based medicine and go back to experience based medicine? Back to the "good old days" where individual doctors knew best for individual patients and where clinicians acted on personal experience? No. Research indicates that there is a positive association between the introduction of evidence based therapies in daily practice and the improvement of MDD treatment, yet its relation is still not unmistakably clear. It is time to answer the question that was asked by A.J Rush in 1993: "Clinical Practice Guidelines: good news, bad news, or no news?"[41]. Many researchers have

tried to answer this significant question, and it is a difficult one to answer. Many factors are involved in treatment in daily practice and therefore daily practice is very complicated to address scientifically [19]. ROM is a very promising and valuable methodology to get insight in the many aspects of routine psychiatric practice. So, indeed, effectiveness in daily practice is not as positive as we hoped for. But go back to experience based treatment? No one knows what treatment results were before the introduction of evidence based medicine. And no one will ever find out, because experience based treatments cannot be explored in terms of effectiveness, since they differ between each patients and nothing is recorded automatically. So one of the big yet bitter advantages of the introduction of guidelines is that we now *know* that effectiveness is currently not as good as the promising results from RCTs. ROM may provide data to improve in an evidence based way the treatment results in clinical practice, e.g. by the future possibilities to identify patients who are at risk of non response or to define subgroups of patients that respond better to a certain type of treatment.

Are the modest MDD treatment results in daily practice a reason for panic or despair, then? A reason to become depressed? A reason to cut down the budget on mental health? We don't think so. The age of evidence based medicine went hand in hand with the age of optimistic belief in antidepressant treatment, efficacy trials on ssri's and an enormous increase in the prescription of antidepressants. Among others, pharmaceutical industries conducted efficacy trials on antidepressants and showed that depression is a treatable disorder. Those days of optimism are over. Antidepressants seem not be as effective as was believed [42], not even in the short term, 6-8 weeks follow-up trials in which no effort is spared to optimize adherence, and in which only patients with moderate to severe MDD are treated. From our results and from those of the STAR*D trial [10,11], it is clear that in daily practice even short term treatment of MDD is hard, and the results modest. In addition, it has become evident that depression is a chronic illness [43-45], which remits and recurs, and rarely disappears.

Depression causes a lot of suffering, some patients who suffered both from very severe "somatic" illnesses or terrible personal losses and major depressive disorder, stated that their depression was the worst. The loss of hope, a continuous feeling of worthlessness and/or despair, the inability to participate in daily life in the broadest sense of the word together with all the physical complaints that may occur when one suffers from major depressive disorder, surely makes MDD a disease that justifies all efforts from patients, caregivers, clinicians, researchers, and mental health policy makers. MDD is an expensive disease with respect to direct costs on the health care budget (not only the mental health care budget) and indirect costs with respect to absenteeism. Depression is the leading cause of disability and the fourth leading contributor to the global burden of disease according to the World Health Organization (https://www.who.int/en). MDD is a very serious medical issue, like other chronic diseases such as diabetes or chronic obstructive pulmonary disease (COPD), not a temporarily suffering from worries that will go away after a good talk with

your neighbour. It is of importance that the people in charge of the mental health care budgets, the policymakers, the government and the minister of Public Health take notice of the complexity, severity and chronicity of MDD.

Having learned that we, as clinicians and researchers, have to be modest about the prognosis of patients suffering from MDD who seek treatment for it, we cannot just sit and wait... We can improve treatment adherence of patients and clinicians and we can develop staging and profiling of MDD. This discussion will be closed by elaborating on these three topics.

Firstly, we can ameliorate our methods to improve treatment adherence of patients. Many new developments may improve adherence: e-health, apps with medication instruction, sms alerts for medication, technical devices that help patients to monitor changes in their mood by providing feedback several times a day, and collaborative care (an integrated approach of the biological, social and psychological aspects of MDD). For better treatment adherence, we have to invest in the education of patients suffering from MDD. It has been proven that informing patients about the nature of their disease and its treatment, the duration, the expected results and time span, the expected investment of the patients and possible side effects of the treatment, which accounts for both pharmacotherapy and psychotherapy will improve adherence [46,47]. It is hard to tell a patient who just got out of a period of feeling worthless and guilty, who had nights without sleep, days without energy or appetite and who nearly came to the edge of committing suicide, that it is likely that this illness will return, sooner or later. Nevertheless, education is a very important part of MDD treatment. Future research on the effect of improvement of treatment adherence on outcome in daily practice is highly recommended. Secondly, improvement of the protocol adherence of clinicians in daily practice might also lead to an increase of effectiveness. Due to a variety of reasons, clinicians in daily practice sometimes find it difficult to strictly follow the protocol (especially when it comes to the frequency of follow up contacts or taking blood levels of antidepressants). Further research on the association between protocol adherence and outcome is recommended [19]. In this project, we presented ROM as a valuable methodology to do scientific research in daily practice. Other potential benefits of ROM remained underexposed in this thesis so far, but need to be mentioned. In Rivierduinen, ROM was primarily designed to ameliorate the evaluation of the treatment of individual patients and patient groups. A systematic evaluation of treatment progress after each treatment step helps clinicians and patients to see whether they are on the right track, and what further steps need to be taken. If ROM is fully incorporated in daily routine, it can be a helpful tool for clinicians to remain adherent to their treatment protocol and to switch in time to a next treatment step in the protocols for MDD treatment.

Finally, depression treatment itself can be improved by so called staging and profiling (a specific therapy for a specific stage or subtype of the disease). At this moment, almost all patients suffering from MDD are treated the same way, either with antidepressants or with

psychotherapy (or a combination of both). The choice for either one of treatment modalities is based only on severity of the depression and the preferences of the patients. Yet, there are many clues that not all MDDs are the same. Within the disorder, different symptoms or symptom dimensions may have different etiology. Different symptom dimensions [48] have been demonstrated to be associated with different genetic pathways [Van Veen, in press], with differences in the dysregulation of the HPA-axis [49,50], and different types of childhood trauma [51] [Van Veen, submitted] and life events [Wardenaar, submitted]. Currently more and more results become available indicating that different subtypes of depression need different treatment. For instance in the STAR*D trial was found that specific genotypes together with co morbid anxiety disorders (in our ROM sample 43% suffered from co morbid anxiety and/or somatoform disorders) are associated with non-response to antidepressants [52]. Similar results were found in the Genome Based Therapeutic Drugs for Depression (GENDEP) study [53].Therefore, patients suffering from MDD should not all be treated in the same way, but with treatment tailor made for their type of depression. Future research should focus on those tailor made treatments.

REFERENCE LIST

1. Zimmerman M, Mattia JI, Posternak MA: Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry* 2002, 159: 469-473.
2. Zetin M, Hoepner CT: Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol* 2007, 27: 295-301.
3. Partonen T, Sihvo S, Lonnqvist JK: Patients excluded from an antidepressant efficacy trial. *J Clin Psychiatry* 1996, 57: 572-575.
4. Rothwell PM: External validity of randomized controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005, 365: 82-93.
5. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA *et al*.: Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials* 2004, 25: 119-142.
6. 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.
7. Posternak MA, Zimmerman M, Keitner GI, Miller IW: A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry* 2002, 159: 191-200.
8. Zimmerman M, Chelminski I, Posternak MA: Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *J Nerv Ment Dis* 2004, 192: 87-94.
9. Hamilton M: Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967, 6: 278-296.
10. 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.
11. 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.
12. McPherson K: Do patients' preferences matter? *BMJ* 2008, 337:a2034. doi: 10.1136/bmj.a2034.: a2034.
13. Churchill R, Khaira M, Gretton V, Chilvers C, Dewey M, Duggan C *et al*.: Treating depression in general practice: factors affecting patients' treatment preferences. *Br J Gen Pract* 2000, 50: 905-906.
14. Dwight-Johnson M, Sherbourne CD, Liao D, Wells KB: Treatment preferences among depressed primary care patients. *J Gen Intern Med* 2000, 15: 527-534.
15. Persaud N, Mamdani MM: External validity: the neglected dimension in evidence ranking. *J Eval Clin Pract* 2006, 12: 450-453.
16. Kirsch I, Sapirstein G: Listening to Prozac, but hearing Placebo: a meta-analysis of antidepressant medications. In *How Expectancies shape Experience*. American Psychological Association; 1999:302-320.
17. 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.
18. Kirsch I (Eds):Response Expectancy of Experience and Behavior. In *American Psychologist* 1985, 40: 1189-1202.
19. Fenema EMv, Wee Nvd, Bauer M, Witte C.J., Zitman FG: Assessing adherence to guidelines for common mental disorders in routine clinical practice. *International Journal for Quality in Health Care* 2011, 1-8.
20. Katon W, Von KM, Lin E, Walker E, Simon GE, Bush T *et al*.: Collaborative management to achieve treatment guidelines. Impact on depression in primary care. *JAMA* 1995, 273: 1026-1031.

21. Badamgarav E, Weingarten SR, Henning JM, Knight K, Hasselblad V, Gano A, Jr. *et al*.: Effectiveness of disease management programs in depression: a systematic review. *Am J Psychiatry* 2003, 160: 2080-2090.
22. Andrews G, Issakidis C, Sanderson K, Corry J, Lapsley H: Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment of ten mental disorders. *Br J Psychiatry* 2004, 184: 526-533.
23. Pirraglia PA, Rosen AB, Hermann RC, Olchanski NV, Neumann P: Cost-utility analysis studies of depression management: a systematic review. *Am J Psychiatry* 2004, 161: 2155-2162.
24. Neumeyer-Gromen A, Lampert T, Stark K, Kallischnigg G: Disease management programs for depression: a systematic review and meta-analysis of randomized controlled trials. *Med Care* 2004, 42: 1211-1221.
25. Trivedi MH, Claassen CA, Grannemann BD, Kashner TM, Carmody TJ, Daly E *et al*.: Assessing physicians' use of treatment algorithms: Project IMPACTS study design and rationale. *Contemp Clin Trials* 2007, 28: 192-212.
26. Posternak MA, Zimmerman M: Therapeutic effect of follow-up assessments on antidepressant and placebo response rates in antidepressant efficacy trials: meta-analysis. *Br J Psychiatry* 2007, 190: 287-292.
27. 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, 196: 173-178.
28. Petkova E, Quitkin FM, McGrath PJ, Stewart JW, Klein DF: A method to quantify rater bias in antidepressant trials. *Neuropsychopharmacology* 2000, 22: 559-565.
29. Asberg M, Montgomery SA, Perris C, Schalling D, Sedvall G: A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl* 1978, 5-27.
30. Zimmerman M, Posternak MA, Chelminski I: Derivation of a definition of remission on the Montgomery-Asberg depression rating scale corresponding to the definition of remission on the Hamilton rating scale for depression. *J Psychiatr Res* 2004, 38: 577-582.
31. Zimmerman M, Posternak MA, Chelminski I: Implications of using different cut-offs on symptom severity scales to define remission from depression. *Int Clin Psychopharmacol* 2004, 19: 215-220.
32. 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.
33. 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.
34. Thase ME, Entsuah AR, Rudolph RL: Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *Br J Psychiatry* 2001, 178:234-41.
35. Mittmann N, Mitter S, Borden EK, Herrmann N, Naranjo CA, Shear NH: Montgomery-Asberg severity gradations. *Am J Psychiatry* 1997, 154: 1320-1321.
36. Hawley CJ: Depression rating scales can be related to each other by simple equations. 1998.
37. 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.
38. 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.
39. 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.
40. Royston P: Multiple imputation of missing values: update. *Stata Journal* 2005, 5: 188-201.
41. Rush AJ: Clinical practice guidelines. Good news, bad news, or no news? *Arch Gen Psychiatry* 1993, 50: 483-490.

42. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R *et al*.: Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009, 373: 746-758.
43. Andrews G: Should depression be managed as a chronic disease? *BMJ* 2001, 322: 419-421.
44. Gask L: Is depression a chronic illness? For the motion. *Chronic Illn* 2005, 1: 101-106.
45. Judd LL, Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W *et al*.: A prospective 12-year study of subsyndromal and syndromal depressive symptoms in unipolar major depressive disorders. *Arch Gen Psychiatry* 1998, 55: 694-700.
46. Trivedi MH, Lin EH, Katon WJ: Consensus recommendations for improving adherence, self-management, and outcomes in patients with depression. *CNS Spectr* 2007, 12: 1-27.
47. Vergouwen AC, Bakker A, Katon WJ, Verheij TJ, Koerselman F: Improving adherence to antidepressants: a systematic review of interventions. *J Clin Psychiatry* 2003, 64: 1415-1420.
48. den Hollander-Gijsman ME, de BE, van der Wee NJ, van Rood YR, Zitman FG: Distinguishing between depression and anxiety: a proposal for an extension of the tripartite model. *Eur Psychiatry* 2010, 25: 197-205.
49. Wardenaar KJ, Vreeburg SA, van VT, Giltay EJ, Veen G, Penninx BW *et al*.: Dimensions of depression and anxiety and the hypothalamo-pituitary-adrenal axis. *Biol Psychiatry* 2011, 69: 366-373.
50. Veen G, van Vliet IM, DeRijk RH, Giltay EJ, van PJ, Zitman FG: Basal cortisol levels in relation to dimensions and DSM-IV categories of depression and anxiety. *Psychiatry Res* 2011, 185: 121-128.
51. van Reedt Dortland AK, Giltay EJ, van VT, Zitman FG, Penninx BW: Personality traits and childhood trauma as correlates of metabolic risk factors: the Netherlands Study of Depression and Anxiety (NESDA). *Prog Neuropsychopharmacol Biol Psychiatry* 2012, 36: 85-91.
52. Ising M, Lucae S, Binder EB, Bettecken T, Uhr M, Ripke S *et al*.: A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression. *Arch Gen Psychiatry* 2009, 66: 966-975.
53. Uher R, Perroud N, Ng MY, Hauser J, Henigsberg N, Maier W *et al*.: Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *Am J Psychiatry* 2010, 167: 555-564.

Nederlandstalige Samenvatting

Titel: ZIJN DEPRESSIE-TRIALS GENERALISEERBAAR NAAR DE KLINISCHE PRAKTIJK?

Subtitel: *Wat clinici altijd al hadden willen weten over RCTs, maar niet durfden te vragen….*

Dr. X is een 60-jarige, alom gerespecteerde, psychiater die in een grote psychiatrische polikliniek werkt. Hij ziet iedere dag vele patiënten met uiteenlopende psychiatrische stoornissen. Jongere collega's verwijzen vaak complexe patiënten naar hem vanwege zijn lange ervaring. Dr. X heeft gedurende zijn carrière al veel ontwikkelingen in de psychiatrie meegemaakt: nieuwe psychofarmaca, de antipsychiatrie, de zelfbewustwording van patiënten, de afnemende populariteit van de psychoanalytische therapie, het toenemend belang van behandelprotocollen en de vooruitgang in inzicht in biologische aspecten van psychiatrische stoornissen. In de polikliniek waar dr. X werkt, zijn, zoals in vele poliklinieken, de nationale richtlijnen voor behandeling van psychiatrische stoornissen omarmd en geïmplementeerd. Zoals veel van zijn collega's was dr. X geïnteresseerd, maar ook wat sceptisch, en hij maakte zich zorgen dat deze richtlijnen alle creativiteit uit het vak zouden doen verdwijnen. Toch zette dr. X zich in voor navolging van de behandelrichtlijnen. Hij hield zijn vakliteratuur bij over psychofarmaca en psychotherapie, vooral voor depressie, aangezien de meeste van zijn patiënten daaraan lijden. Hij las de veelbelovende resultaten uit randomized controlled trials (RCTs) voor verschillende antidepressiva en nieuwe methodes van psychotherapie. Ondertussen waren de resultaten van medicatie of psychotherapie in zijn praktijk vaak teleurstellend en bleven zijn patiënten worstelen met hun depressie. Dr. X kreeg de indruk dat het effect van depressie behandeling een stuk groter is in RCTs dan in "de echte wereld". Hij ging zich het volgende afvragen: zijn mijn patiënten wel hetzelfde als die deelnemers aan depressie-trials? Hoe moet ik de resultaten uit RCTs interpreteren? Vertellen RCTs ons eigenlijk wel iets over de "echte wereld"? En klopt het eigenlijk wel dat wij onze behandelrichtlijnen baseren op resultaten uit RCTs die misschien zo ver van de dagelijkse praktijk staan?

In dit proefschrift hebben wij geprobeerd de vragen van dr. X te beantwoorden.

Dr. X behandelt veel patiënten met een depressie. Depressie is een stemmingsstoornis, die zich kenmerkt door aanhoudende somberheid of neerslachtigheid en/of het verlies van belangstelling of genoegen. Daarnaast is er bij depressie sprake van een aantal van de volgende symptomen: verandering in eetlust; verstoord slaappatroon; rusteloosheid of traagheid; vermoeidheid of energieverlies; schuldgevoel, bezorgdheid of angst; concentratieproblemen en gedachten aan zelfmoord. Een depressie is een veelvoorkomende en ernstige psychiatrische aandoening, die bij een groot aantal patiënten meerdere malen terugkeert in het leven. Wereldwijd staat depressie op nummer één als het gaat om invaliderende ziektes, zo stelt de World Health Organisation. Depressie komt zowel

bij mannen als vrouwen voor, in iedere leeftijdscategorie en bij iedere etnische of sociaal -economische achtergrond. Depressie komt meer voor bij vrouwen dan bij mannen (ruim één op de tien mannen maakt eens in zijn leven een depressie door, terwijl voor vrouwen dit één op de vijf is). Depressieve patiënten lijden erg onder hun ziekte. Daarnaast vormt depressie een maatschappelijk probleem: depressieve mensen maken meer gebruik van gezondheidszorg en uitkeringen en depressieve mensen zijn vaak niet in staat om te werken, waardoor er veel arbeidsproductiviteit verloren gaat. In Nederland zijn de geschatte kosten, in totaal, van depressie ongeveer 1.1% van de totale kosten van de gezondheidszorg.

Depressie is een behandelbare aandoening. Psychiaters en psychologen in Nederland volgen, net als Dr. X, daarvoor de multidisciplinaire richtlijnen voor behandeling van depressie. Bij een depressie zijn behandeling met medicatie (antidepressiva) en psychotherapie ongeveer even effectief gebleken. Als de patiënt de voorkeur geeft aan een psychotherapeutische behandeling, kan hij kiezen voor cognitieve gedragstherapie, waarbij de patiënt door middel van oefeningen zijn depressieve gedachten en gedragingen probeert te veranderen. Hij kan ook kiezen voor interpersoonlijke therapie, waarin een patiënt de levensfase of gebeurtenis die de aanleiding was voor het ontwikkelen van een depressie doorwerkt met de psychotherapeut. Beide vormen van psychotherapie zijn ongeveer even effectief, al is er veel meer onderzoek gedaan naar cognitieve gedragstherapie. Pas bij onvoldoende effect van psychotherapie of antidepressiva wordt aangeraden om te starten met combinatietherapie (een combinatie van medicatie en psychotherapie). Dit principe van enkelvoudig beginnen en pas later behandelingen combineren noemt men *stepped care*. In Nederland worden de behandelrichtlijnen samengesteld door speciale werkgroepen met experts uit het werkveld en uitgegeven door het Trimbos Instituut.

In de richtlijnen voor behandeling van depressie wordt aangegeven welke therapieën er *bewezen* effectief zijn, ook in het Nederlands meestal aangeduid met de Engelse term *evidence based*. Het bewijs voor effectiviteit wordt verkregen uit wetenschappelijk onderzoek. De onderzoeksmethode die beschouwd wordt als de methode die het meeste "harde" bewijs oplevert is de randomized controlled trial (RCT). In een RCT wordt in een van te voren vastgestelde groep patiënten gekeken of een behandeling effectiever is dan placebo behandeling (een medicament dat er net zo uitziet als het onderzochte medicament, maar dan zonder werkzame stoffen). In het geval van psychotherapie is een placebo niet haalbaar en wordt gekeken of een bepaalde psychotherapie effectiever is dan de behandeling die gebruikelijk is voor de ziekte, in het Engels aangeduid als Treatment As Usual (TAU) of op een wachtlijst staan (de zogenoemde "wachtlijst groep"). De onderzoeksopzet RCT is heel strikt omdat hij bedoeld is om aan te tonen dat er een heldere relatie is tussen een behandelinterventie en de uitkomst van deze behandeling op een specifieke stoornis zonder dat er sprake is van placebo effect. Men spreekt van placebo effect als patiënten "spontaan" opknappen als zij het idee hebben een behandeling te krijgen, zonder dat er sprake is van de werkzame component van die behandeling. Om in een RCT zo ondubbelzinnig mogelijk

te kunnen aantonen dat het effect echt het gevolg is van de onderzochte behandeling is er niet alleen een controlegroep nodig; de deelnemende patiënten mogen bijvoorbeeld ook geen andere ziekten hebben dan die waarop de behandeling gericht is.

In de dagelijkse praktijk, zoals die van dr. X is dat anders. Vaak hebben patiënten naast hun depressie nog andere psychiatrische stoornissen. De combinatie van een angststoornis en een depressie komt erg vaak voor. Ook drinken depressieve patiënten vaker dan gemiddeld alcohol en soms gebruiken zij drugs. Al deze problemen zijn een reden om niet mee te mogen doen aan een depressie trial. De vraag is dan natuurlijk: als de meeste depressieve patiënten uit de dagelijkse praktijk anders zijn dan deelnemers aan zo'n depressie trial, zijn de resultaten van zulke trials dan wel van toepassing op (generaliseerbaar naar) de dagelijkse praktijk? Die vraag hebben wij geprobeerd te beantwoorden in dit proefschrift.

Tot tien jaar geleden was het onduidelijk hoe effectief behandelingen voor depressie in de dagelijkse praktijk waren. Het effect van behandeling werd weinig gemeten door clinici en als ze het al deden werd het niet op een systematische manier gedaan. Daardoor waren er geen gegevens over de behandelresultaten bij depressieve patiënten in de dagelijkse praktijk. Onderzoekers hebben toen een nieuwe onderzoeksopzet ontworpen die de dagelijkse praktijk veel meer benadert dan de RCTs dat deden. Patiënten hoefden niet aan een veelheid aan strikte criteria te voldoen voor deelname en de behandeling vond plaats in de dagelijkse praktijk. Deze onderzoeksopzet wordt ook wel de *pragmatische trial* genoemd. Een bekend voorbeeld hiervan is de STAR*D trial, een heel grote pragmatische trial uit 2004 naar behandeling van depressie in de Verenigde Staten. Er deden 4000 patiënten aan mee. Een nadeel van pragmatische trials is, dat hoewel zij de dagelijkse praktijk zoveel mogelijk benaderen, ze toch altijd eigenschappen van een RCT houden. Een andere manier om behandeleffect te kunnen meten in de dagelijkse praktijk is Routine Outcome Monitoring (ROM). Bij Routine Outcome Monitoring wordt de psychiatrische stoornis van patiënten in dagelijkse praktijk bij binnenkomst en daarna steeds na een vaste periode gemeten met gevalideerde meetinstrumenten (vragenlijsten). In 2002 heeft Rivierduinen, in samenwerking met de afdeling Psychiatrie van het LUMC ROM ingevoerd in de dagelijkse praktijk van poliklinische behandeling van patiënten met een stemmings-, angst-, of somatoforme stoornissen. Na enkele jaren is ROM in Rivierduinen uitgebreid naar andere psychiatrische stoornissen. Rivierduinen, een grote GGZ instelling met meerdere vestigingen in Zuid-Holland, heeft een verzorgingsgebied van ongeveer een miljoen Nederlanders. ROM heeft als enige criterium voor deelname dat patiënten het Nederlands voldoende moeten beheersen en dat zij niet te ernstig ziek zijn voor het invullen van vragenlijsten. Bij binnenkomst krijgen patiënten allemaal, naast het gebruikelijke intakegesprek met een behandelaar, een serie vragenlijsten waarmee systematisch wordt nagegaan welke klachten zij hebben en hoe ernstig deze zijn. Daarnaast wordt bij alle patiënten gekeken hoe ze hun kwaliteit van leven ervaren, en hoe ze sociaal en maatschappelijk functioneren. Bij de vervolgmetingen wordt steeds gekeken hoe de ernst van de klachten op dat moment is en in hoeverre de patiënten tevreden zijn over hun functioneren. De patiënten vullen

een deel van de vragenlijsten zelf in op een computer d.m.v. een touch screen. De overige vragenlijsten worden ingevuld door een testverpleegkundige die de patiënt de vragen stelt. Het eerste doel van ROM is om behandelaar en patiënt bij de intake te informeren over de aard en ernst van de klachten en later over de voortgang van behandeling. Daarnaast is door de uitgebreide opzet van ROM van LUMC/Rivierduinen het goed mogelijk om allerlei vragen uit de dagelijkse praktijk wetenschappelijk te onderzoeken. Voor dit proefschrift maakten wij gebruik van gegevens die met ROM verzameld zijn.

BELANGRIJKSTE BEVINDINGEN

In het **eerste hoofdstuk,** de inleiding van dit proefschrift, hebben wij een overzicht gegeven van de opbouw van de Nederlandse, Engelse en Amerikaanse richtlijnen voor de behandeling van depressie. We hebben laten zien hoe het bewijs voor effectiviteit van verschillende behandelingen voor depressie gewogen wordt en hoe zwaar de resultaten van RCTs wegen voor de verschillende richtlijnen. Vervolgens hebben we uiteen gezet welke behandelingen worden aanbevolen in de richtlijnen als eerste stap in de behandeling van depressie. Daarna zijn we ingegaan op de methodologie van RCTs en de beperkingen die deze strenge methodologie met zich meebrengt voor de generaliseerbaarheid van resultaten uit RCTs naar de dagelijkse psychiatrische praktijk. We hebben het verschil uitgelegd tussen *efficacy* (de effectiviteit van een behandeling gemeten in RCTs) en *effectiveness* (de werkzaamheid van deze behandeling in de dagelijkse praktijk). In dit hoofdstuk hebben we aangegeven wat de meest gebruikte meetinstrumenten zijn om de effectiviteit van behandeling van depressie te meten: de Hamilton Rating Scale for Depression (HAMD), de Montgomery Asberg Depression Rating Scale (MADRS) en de Beck Depression Inventory (BDI-II). De meest gebruikte manieren om effectiviteit weer te geven zijn: responspercentage, remissiepercentage en effectsize. Men spreekt van respons als de patiënt 50% minder symptomen op een vragenlijst scoort dan bij het begin van de behandeling. Men spreekt van remissie als patiënten beneden een bepaalde cut-off scoren (MADRS≤10). Effectsize is het verschil in de score van symptomen na en vóór behandeling, gecorrigeerd voor de standaarddeviatie (spreiding in scores) vóór behandeling. We hebben uitgelegd hoe ROM gebruikt kan worden om de werkzaamheid (effectiveness) van behandeling van depressie in de dagelijkse praktijk te meten. Tot slot hebben we in dit eerste hoofdstuk uiteengezet hoe de selectie van depressieve patiënten in RCTs, de generaliseerbaarheid van resultaten naar de dagelijkse praktijk negatief zou kunnen beïnvloeden.

In het **tweede hoofdstuk**, de eerste studie in dit proefschrift, hebben we de effectiviteit van behandeling van depressie gemeten in 15 meta-analyses en die in onze dagelijkse praktijk gemeten met ROM vergeleken. Iedere meta-analyse geeft een geaggregeerde maat (het gemiddeld effect van een grote verzameling RCTs, gecorrigeerd voor het aantal deelnemers

aan de RCTs) voor het behandeleffect. Ook hebben we een vergelijking gemaakt tussen het effect van behandeling in de dagelijkse (ROM) praktijk en die van de STAR*D trial. Hiervoor hebben we 598 depressieve patiënten geïncludeerd die tussen 2002 en 2006 behandeling zochten bij Rivierduinen. Deze 598 patiënten hadden een ROM meting ondergaan bij binnenkomst en hadden tenminste 1 vervolgmeting. We hebben gevonden dat de remissiepercentages voor alle behandelingen: antidepressiva, psychotherapie en combinatietherapie lager waren in de dagelijkse praktijk dan in RCTs (32% vs.40–74%). Dit verschil was het meest duidelijk voor psychotherapie en combinatietherapie. Er bleek geen verschil te zijn tussen de behandelresultaten in onze dagelijkse praktijk en die van STAR*D.

In het **derde hoofdstuk** hebben wij onderzocht hoeveel depressieve patiënten uit dagelijkse praktijk in aanmerking zouden komen voor een antidepressivatrial. Hiervoor hebben we 1653 depressieve patiënten geïncludeerd die tussen 2002 en 2006 behandeling zochten bij Rivierduinen en die een ROM meting hadden ondergaan bij binnenkomst. We hebben berekend welk percentage van deze patiënten in aanmerking zou komen volgens de meest gebruikte exclusiecriteria in RCTs voor antidepressiva. Dat was slechts bij 17–25% van onze patiënten het geval. De belangrijkste redenen voor exclusie waren: niet voldoen aan de minimum ernst van de depressie (HAMD ≤17) en de aanwezigheid van comorbide psychiatrische stoornissen. Andere veelvoorkomende redenen voor exclusie waren: suïcidaliteit en misbruik of afhankelijkheid van alcohol en/of drugs. Vervolgens hebben wij bij 626 patiënten van de 1653 patiënten onderzocht wat de invloed van de veelgebruikte exclusiecriteria was op het behandeleffect. Deze 626 patiënten werden geselecteerd omdat zij tenminste 1 vervolgmeting hadden. Onze belangrijkste bevinding in deze studie was dat "in aanmerking komen voor deelname (voldoen aan alle criteria)" niet van invloed is op het behandelresultaat in de dagelijkse praktijk. Onze interpretatie is dat het gebruik van exclusiecriteria waarschijnlijk niet een zodanige bedreiging is voor de generaliseerbaarheid van resultaten uit antidepressiva-trials als in eerder onderzoek werd gesuggereerd. Waarschijnlijk zijn er andere factoren die het verschil in behandeleffect in antidepressiva trials en de dagelijkse praktijk kan verklaren. Mogelijke verklaringen zijn de veel grotere inspanning die in trials wordt geleverd om therapietrouw te bevorderen, waarschijnlijk is de frequentie van behandelcontacten in trials hoger dan in de praktijk en speelt in trials het zogenaamde Hawthorne effect mee (een gunstiger uitkomst doordat patiënten hoopvoller zijn omdat zij meedoen aan een bijzondere behandeling in een bijzondere setting).

In het **vierde hoofdstuk** hebben wij dezelfde vraagstelling onderzocht, maar dan voor psychotherapie-trials. Van psychotherapie-trials was niet bekend welke exclusiecriteria het meest gebruikt worden. Dit hebben wij onderzocht in 20 psychotherapie-trials voor depressie. We hebben gevonden dat psychotherapie-trials minder consistent waren in het gebruik van exclusiecriteria dan antidepressiva-trials. De volgende criteria worden veel gebruikt en zijn mogelijk van invloed op de generaliseerbaarheid van de resultaten van psychotherapie-trials voor depressie: 'niet voldoen aan de minimum ernst (HAMD≤14)', 'misbruik of afhankelijkheid

van drugs en/of alcohol' en 'gebruik van medicatie of electroconvulsie therapie voorafgaand aan de psychotherapie'. Aangezien de exclusiecriteria in psychotherapie-trials niet consistent worden gebruikt, was het niet mogelijk om een percentage ROM patiënten te berekenen dat in aanmerking zouden komen voor deelname. Wel hebben we gevonden dat de invloed van het gebruik van de afzonderlijke exclusiecriteria op behandelresultaat in de dagelijkse praktijk laag was: 'misbruik van alcohol en/of drugs' had geen invloed en de invloed van de andere twee exclusiecriteria was gering. Ook bij de psychotherapietrials is onze interpretatie van de gevonden resultaten de generaliseerbaarheid van de resultaten waarschijnlijk minder ernstig wordt bedreigd door het gebruik van exclusiecriteria dan eerder werd gedacht. Waarschijnlijk geldt ook voor psychotherapie dat andere factoren, zoals genoemd in hoofdstuk 3, het verschil in behandeleffect tussen de dagelijkse praktijk en trials kunnen verklaren.

In het **vijfde hoofdstuk** hebben wij de socio-demografische en socio-economische kenmerken (SES kenmerken) van deelnemers aan depressie-trials in kaart gebracht. Hiervoor hebben wij 45 antidepressiva-trials en 19 psychotherapie-trials geïncludeerd. We hebben gevonden dat de rapportage van SES kenmerken niet eenduidig is en dat er vaak beperkte informatie gegeven wordt. Vooral vermelding van opleidingsniveau, sociaaleconomische status en inkomen wordt vaak achterwege gelaten, terwijl die wel van invloed kunnen zijn op de behandeluitkomst. Uit dit onderzoek is gebleken dat deelnemers aan depressie-trials gemiddeld 41 jaar zijn, voornamelijk vrouw (62%) en voornamelijk blank (89%) zijn. Onze conclusie is dat standaardisatie van de rapportage van SES kenmerken in RCTs de vergelijking tussen trials en met de dagelijkse praktijk ten goede zou komen.

In het **zesde hoofdstuk** hebben wij de SES kenmerken van deelnemers aan depressie-trials (zoals gevonden in hoofdstuk 5) met die van patiënten uit de dagelijkse praktijk vergeleken. Wij hebben hiervoor weer de 626 depressieve patiënten die hierboven genoemd werden geïncludeerd. Deze patiënten ondergingen een ROM meting bij binnenkomst en hadden tenminste 1 vervolgmeting. Wij hebben gevonden dat trialdeelnemers ouder, vaker van het mannelijk geslacht, vaker blank en vaker ongetrouwd waren dan patiënten uit de dagelijkse praktijk. Deze verschillen bleken echter klein te zijn. Opvallend was dat veel meer trialdeelnemers betaald werk hadden gedurende hun behandeling dan patiënten uit de dagelijkse praktijk (verschil 34%). Het doen van betaald werk voorspelde een betere behandeluitkomst bij depressie in de dagelijkse praktijk.

ALGEMENE DISCUSSIE

Onze studies geven inzicht in het verschil in behandelresultaat voor depressie tussen trialsettings en de dagelijkse praktijk. We kunnen concluderen dat clinici, zoals dr. X, gelijk hebben, als zij stellen dat de resultaten in de dagelijkse praktijk vaak minder hoopvol zijn dan RCTs suggereren. Ook kunnen we concluderen dat clinici gelijk hebben als zij stellen dat hun patiënten uit de "echte wereld" verschillen van deelnemers aan trials; patiënten uit de dagelijkse praktijk zijn vaak minder ernstig ziek dan die in trials, maar ze zijn wel vaak suïcidaal en hebben in meerderheid andere psychiatrische stoornissen of verslavingen, allemaal verschijnselen die bij trials juist redenen zijn om mensen uit te sluiten voor deelname. Ook hebben patiënten uit de dagelijkse praktijk vaak wél antidepressiva gebruikt voordat ze aan psychotherapie beginnen. Daarnaast hebben depressieve patiënten in de dagelijkse praktijk veel minder vaak betaald werk tijdens hun behandeling dan deelnemers aan depressietrials. Onze belangrijkste bevinding is echter, in tegenstelling tot het vermoeden van veel clinici zoals dr. X: "in aanmerking komen voor een depressie trial (voldoen aan alle criteria) is niet van invloed op het behandeleffect in de dagelijkse praktijk". Met andere woorden: als in de dagelijkse praktijk alleen nog de "schone" depressieve patiënten die in aanmerking komen voor deelname aan RCTs zouden worden behandeld, dan zal het behandelresultaat niet sterk verbeteren. De verschillen tussen trial deelnemers en patiënten uit de dagelijkse praktijk verklaren het verschil in behandelresultaat tussen RCTs voor depressie en de dagelijkse praktijk dus niet.

In de algemene discussie zijn we uitgebreider ingegaan op een aantal andere mogelijke verklaringen voor het verschil tussen efficacy en effectiveness bij depressiebehandeling. Waarschijnlijk zijn er andere verschillen tussen trialdeelnemers en patiënten uit de dagelijkse praktijk die van invloed zijn op het behandeleffect: verschillen in motivatie en therapietrouw en in de prognostische kenmerken van de depressie. Daarnaast zijn er verschillen tussen behandelaars die meedoen aan een trial en behandelaars in de dagelijkse praktijk: verschillen in werkdruk, motivatie en protocolgetrouwheid. De polikliniek waar een trial wordt uitgevoerd verschilt van de dagelijkse praktijk: trials worden vaak uitgevoerd in gespecialiseerde centra en er wordt voor een trial veel geïnvesteerd in therapietrouw en het voorkomen van uitval van behandelafspraken. Waarschijnlijk is het placebo effect in RCTs groter dan in de dagelijkse praktijk, wat bijdraagt aan het verschil in behandelresultaat tussen trials en de dagelijkse praktijk. Tot slot is er meestal sprake van publicatiebias bij RCTs: trials die een positief resultaat laten zien worden vaker gepubliceerd dan trials die een negatief effect of geen effect hebben gevonden. Het feit dat deze laatste trials niet worden gepubliceerd, vergroot de verschillen tussen in meta-analyses gerapporteerde efficacy en effectiveness.

Er zijn drie belangrijke beperkingen in dit onderzoeksproject: beperkingen aan de kant van de RCT resultaten, beperkingen in de ROM data en beperkingen in de generaliseerbaarheid van onze bevindingen. Ten eerste was er in RCTs voor depressie gebrekkige consistentie in het gebruik van meetinstrumenten, definitie van behandeleffect en het gebruik van exclusie criteria. Door inconsistentie in gebruik van meetinstrumenten en definities van behandeluitkomst in RCTs is het mogelijk dat de efficacy uitkomsten in meta-analyses minder valide zijn dan ze lijken. In ons onderzoek hebben we een "strenge" cut-off gebruikt voor onze definitie van remissie. Het is daardoor mogelijk dat in werkelijkheid het verschil tussen efficacy en effectiveness bij depressiebehandeling minder groot is dan wij gevonden hebben. Ten tweede worden behandelgegevens niet meegenomen in ROM. Informatie over de behandeling hebben wij verzameld door middel van uitgebreid status onderzoek. Naast de grote tijdsinvestering die dit status onderzoek vergde, was het soms lastig om de benodigde informatie uit de statussen te halen. Het is erg aanbevelenswaardig als in de toekomst behandelgegevens onderdeel zijn van ROM. Daarnaast was er bij ROM sprake van een hoge loss-to-follow- up: slechts 50% van de patiënten die een eerste ROM meting hadden kregen een vervolgmeting. De hoge loss-to-follow-up in onze ROM data, lijkt inherent aan onderzoek in de dagelijkse praktijk: bij STAR*D worden vergelijkbare percentages gevonden. We hebben geprobeerd om selectiebias door loss-to-follow up (patiënten die uitvallen zouden kunnen verschillen van patiënten die niet uitvallen, waardoor het behandeleffect niet meer generaliseerbaar is naar de hele patiëntengroep) zoveel mogelijk te ondervangen. Wij hebben de verschillen tussen de patiënten met alleen een ROM meting bij binnenkomst en die met tenminste één vervolgmeting onderzocht: de verschillen waren minimaal. Ook hebben wij met een statistische methode gecorrigeerd voor ontbrekende data. Tot slot zijn er, ondanks de hoge mate van representativiteit van onze ROM populatie, toch enige beperkingen in de generaliseerbaarheid van onze resultaten. Onze bevindingen zijn waarschijnlijk alleen van toepassing op poliklinisch behandelde depressieve patiënten uit de tweede lijn (verwezen door hun huisarts voor behandeling bij een GGZ instelling) en niet op patiënten met een depressie in de huisartsenpraktijk, in vrijgevestigde praktijken (de eerste lijn) of in academische centra (de derde lijn). Aangezien patiënten alleen gemeten kunnen worden in ROM als zij het Nederlands voldoende machtig zijn, hebben we niet kunnen onderzoeken in hoeverre onze resultaten generaliseerbaar zijn naar alle niet-westerse immigranten.

Mogelijkheden voor toekomstig onderzoek

Aangezien zoveel depressieve patiënten geen ROM vervolgmeting blijken te krijgen, zou dit een belangrijk onderwerp kunnen (en moeten?) zijn voor toekomstig onderzoek. Wellicht bestaan er bepaalde groepen depressieve patiënten die een hoog risico lopen om hun behandeling vroegtijdig af te breken? Indien er door middel van onderzoek achterhaald kan worden om welke patiënten dit gaat, zouden deze patiënten in een zeer vroeg stadium

kunnen worden opgespoord. Zij hebben misschien een specifieke benadering nodig om in behandeling te blijven en een goede evaluatie te krijgen van die behandeling. Dit is belangrijk, aangezien het afronden van een (succesvolle) behandeling voorspellend is voor de prognose van deze patiënten. Ook is het van belang om te weten welke behandelaars patiënten niet doorsturen voor een ROM meting. Het is mogelijk dat er specifieke beroepsgroepen (artsen, psychologen, of sociaal-psychiatrisch verpleegkundigen) zijn die het belang van ROM niet direct onderschrijven. Ook is het mogelijk dat clinici die hun patiënten niet doorsturen bepaalde andere gemeenschappelijke kenmerken hebben zoals leeftijd, mate van ervaring, werkdruk etc. Onderzoek onder clinici zou aan het licht kunnen brengen wat zij nodig hebben om de noodzaak te voelen hun behandelingen systematisch te (laten) evalueren. ROM zou kunnen worden verbeterd met de kennis die opgedaan wordt uit dergelijk onderzoek. Daarnaast is verder onderzoek naar de opzet van ROM zeer aan te bevelen. In tijden van financiële malaise en bezuinigingen wordt de roep om minder vragenlijsten per ROM-meting groter. Dit zou echter de bruikbaarheid van ROM als behandelevaluatie en als wetenschappelijke instrument in gevaar kunnen brengen. Toekomstig onderzoek naar de noodzakelijke ingrediënten van ROM is daarom hard nodig.

In dit proefschrift hebben wij een aantal verschillen tussen deelnemers aan depressietrials en patiënten uit de dagelijkse praktijk uit Rivierduinen onderzocht. Om na te gaan of onze resultaten inderdaad generaliseerbaar zijn naar depressieve patiënten die ambulante behandeling zoeken bij GGZ instellingen zou het waardevol zijn om na te gaan of in andere GGZ instellingen dezelfde verschillen worden gevonden (replicatie van ons onderzoek). Uiteraard zijn er meer verschillen te vinden tussen trial deelnemers en patiënten uit de dagelijkse praktijk dan wij hebben onderzocht. Verder onderzoek naar deze verschillen zou waardevol zijn, zeker naar kenmerken die wij in dit onderzoek niet konden bekijken wegens onvoldoende beschikbare informatie, zoals de voorgeschiedenis van de patiënt. Tot slot zou het heel interessant zijn om het huidige onderzoek uit te breiden naar andere psychiatrische stoornissen, zoals angststoornissen.

In dit proefschrift hebben wij gevonden dat het behandeleffect uit de dagelijkse praktijk achter blijft bij RCTs naar depressie behandeling. We hebben ook gevonden dat er weliswaar veel verschillen zijn tussen deelnemers aan depressietrials en depressieve patiënten uit de dagelijkse praktijk, maar dat die het verschil in behandeleffect niet kunnen verklaren. Wij hopen dat de resultaten die wij hebben gepresenteerd in dit proefschrift geen aanleiding zijn tot somberheid, maar juist een stimulans zal zijn voor verder onderzoek naar behandeluitkomsten voor depressie in de praktijk en alle aspecten die daarmee samenhangen. Met de resultaten van dit toekomstig onderzoek zal men waarschijnlijk in staat zijn om therapietrouw bij patiënten en behandelaars te verbeteren en in de toekomst wellicht zelfs "behandeling op maat" voor de individuele patiënt die aan een depressie lijdt te leveren.

List of publications

1. **van der Lem R** Voordelen van een multidisciplinaire aanpak bij chronische buikpijn. *Patiënt Care* 2000, 27 (4): 33-38
2. de Beurs E, den Hollander-Gijsman ME, van Rood YR, van der Wee NJ, Giltay EJ, van Noorden MS, **van der Lem R**, van Fenema E, Zitman FG: 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.
3. **van der Lem R**, van der Wee NJ, van Veen T., Zitman FG: The generalizability of anti-depressant efficacy trials to routine psychiatric out-patient practice. *Psychological Medicine* 2011, 41: 1353-1363.
4. **van der Lem R**, Stamsnieder P, van der Wee NJA, van Veen T, Zitman FG : Socio-demographic features in randomized controlled trials for major depression: generalizability and individualization. *International Journal of Person Centered Medicine* 2011, **1:** 268-278.
5. Grootenboer EMV, Giltay EJ, **van der Lem R**, van Veen T, van der Wee NJA, Zitman FG. Reliability and validity of the Global Assessment of Functioning Scale in clinical outpatients with depressive disorders. *Journal of Evaluation in Clinical Practice* 2012, 18: 502-507.
6. **van der Lem R**, van der Wee NJ, van VT, Zitman FG: Efficacy versus effectiveness: a direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice**.** *Psychotherapy and Psychosomatics* 2012, 81**:** 226-234.
7. **van der Lem R**, de Wever WW, van der Wee NJ, van VT, Cuijpers P, Zitman FG: The generalizability of psychotherapy efficacy trials in major depressive disorder: an analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry* 2012, 12: 192 (published online, open access)
8. **van der Lem R:** Behandelresultaten bij depressie: gerandomiseerd gecontroleerd onderzoek versus de dagelijkse praktijk**.** *Tijdschrift voor Psychiatrie* 2012, 10: 905-906.
9. **van der Lem R,** Stamsnieder PM, van der Wee NJA, van Veen T, Zitman FG: Influence of sociodemographic and socioeconomic features on treatment outcome in RCTs versus daily psychiatric practice. *Social Psychiatry and Psychiatric Epidemiology* 2012, 12 DOI:10.1007/s00127-012-0624-4 (published online)

Curriculum Vitae

Rosalind van der Lem is geboren op 11 juli 1974 te 's Gravenhage. In 1992 behaalde zij cum laude haar atheneum diploma aan het Nienoord College te Leek. Hierna studeerde zij geneeskunde aan de Rijksuniversiteit Groningen. Gedurende haar studie volgde zij twee wetenschappelijke stages: een stage bij de afdeling Kinderneurologie van het Universitair Medische Centrum Groningen en een stage bij The Hospital of the University of Pennsylvania in Philadelphia,Verenigde Staten, bij de afdeling Cognitive Neurology/Neurolinguistics. In 1999 behaalde zij cum laude haar artsenbul. Daarna werkte zij een jaar als arts-assistent (AGNIO) psychiatrie voor GGZ Delfland te Delft en een jaar als arts-assistent neurologie voor het Sint Franciscus Gasthuis te Rotterdam. In 2001 startte zij haar opleiding tot psychiater bij het Haags Leids Opleidings Consortium Psychiatrie (HLOCP), waarbij zij werkzaam was voor Parnassia in Den Haag en GGZ Delfland in Delft. Gedurende haar opleiding volgde zij 2 keuzestages bij het Centrum Intensieve Behandeling van Parnassia te Den Haag en in het Erasmus Universitair Medisch Centrum te Rotterdam. Sinds april 2006 was zij werkzaam als psychiater bij Rivierduinen GGZ Leiden en startte haar promotie onderzoek, wat resulteerde in dit proefschrift, in combinatie met poliklinische patiëntenzorg. Zij ontving hiervoor een subsidie van ZonMW die tot doel heeft om clinici op te leiden tot bruggenbouwer tussen wetenschap en klinische praktijk in de GGZ. Zij was bij Rivierduinen actief in begeleiden van medische studenten en psychiaters in opleiding bij hun wetenschappelijke stage/ eindreferaat, coördinator van het coschap psychiatrie bij GGZ Leiden en nauw betrokken bij de ontwikkeling van de bipolaire polikliniek. Sinds 2011 is zij werkzaam als psychiater bij Het Dok, een forensische poli- en dagkliniek in Rotterdam, onderdeel van FPC de Kijvelanden. Zij is daar verantwoordelijk voor de ontwikkeling van een gespecialiseerde Forensische ADHD poli, waarin patiëntenzorg, behandelinnovatie, wetenschap, en opleiding vertegenwoordigd zullen zijn. Rosalind van der Lem is getrouwd en heeft een zoon en een dochter.

Rosalind van der Lem was born on the 11th of July 1974 in the Hague. In 1992 she graduated cum laude from secondary school and started studying Medicine at the University of Groningen. In 1999 she obtained her medical degree cum laude. Subsequently, she worked as a physician in psychiatry and neurology for two years. In 2001 she started her psychiatry residency at GGZ Delfland in Delft and Parnassia, the Hague. She completed her psychiatry training in April 2006 and started working as a psychiatrist for Rivierduinen, GGZ Leiden. She combined working as a psychiatrist in an outpatient clinic of GGZ Leiden and working on the current PhD project. She was involved in supervising medical students and psychiatry residents in their scientific training, coordinated the internship psychiatry at GGZ Leiden and participated in the development of a specialized outpatient clinic for bipolar disorders. Since 2011 she is working as a psychiatrist at Het Dok, a forensic outpatient clinic of the FPC the Kijvelanden. In this clinic, she develops a specialized forensic outpatient clinic for patients suffering from ADHD. In this clinic, the treatment of patients, the development of new treatments, training of clinicians and research will be combined. Rosalind van der Lem is married and has a son and a daughter.

Acknowledgements (dankwoord)

Ik heb genoten van het schrijven van dit proefschrift gedurende een periode van maar liefst zeven jaar. Het is een leerzame, inspirerende tijd geweest, met name dankzij de hulp van velen.. Ik wil deze paragraaf benutten om jullie heel hartelijk te bedanken.

Frans, mijn promotor: jouw begeleiding, kritische blik en het vermogen om steeds boven het onderzoek uit te stijgen hebben dit proefschrift gemaakt tot een onderzoeksproject met "evidence based" antwoorden op een al lang bestaande vraag uit de klinische praktijk.
Nic, mijn copromotor, van jou leerde ik niet alleen wetenschappelijke artikelen te schrijven, maar ook om met plezier te laveren in het soms ingewikkelde wetenschappelijke werkveld. Tineke, mijn copromotor, jij gaf me in dit project de vrijheid die ik nodig had en je weet nuchterheid te combineren met geestdrift, een eigenschap die misschien wel zeldzaam is in een academische wereld.

Ik wil graag de testverpleegkundigen van Rivierduinen, de medische studenten en aios psychiatrie, in het bijzonder Wouter de Wever en Purdey Stamsnieder, bedanken voor het verzamelen van de ROM data en de hulp bij het omvangrijke status- en literatuuronderzoek dat nodig was voor dit proefschrift. Wouter en Purdey, het was fijn samenwerken met jullie! Ik ben prof. dr. Pim Cuijpers zeer erkentelijk voor zijn enthousiasme en bijdrage aan hoofdstuk vier van dit proefschrift.

Evert Onstein en Emke Osinga, mijn leidinggevenden bij Rivierduinen, ondersteunden mij in het combineren van wetenschappelijk onderzoek met dagelijkse patiëntenzorg. Veel dank ben ik verschuldigd aan Renske de Reus, die van 2006 tot 2011 mijn supervisor was in het combineren van werk als promovendus en als clinicus. Renske, dankzij jou lukte het om de brug tussen GGZ en wetenschap te slaan. Verder alle lof voor mijn collega's van het SAS-team van Rijnveste voor hun inzet voor ROM en onderzoek in de dagelijkse praktijk.

Ik dank mijn huidige werkgever, in het bijzonder Machiel Polak en Hans Vermeulen, voor de mogelijkheid om mijn proefschrift af te ronden terwijl er meer dan genoeg te doen was en is voor het Dok. En natuurlijk alle betrokken collega's van het Dok voor de waarneming, steun, het enthousiasme en de relativering.

Tot slot wil ik een nog aantal vrienden noemen. Allereerst Michiel van Vreeswijk, Gerthe Veen en Laura van Goor, dank voor jullie eigenzinnige gesprekken over psychiatrie, wetenschap en vele andere onderwerpen. Onze gesprekken hielpen me een vrolijke scherpte te ontwikkelen. En dan natuurlijk Annet Spijker en Esther van Fenema. Ik ben er trots op dat jullie mijn paranimfen zijn. Annet, jij bent degene die ik bel als ik niet meer weet hoe ik wetenschap, patiëntenzorg en moederschap moet combineren. Esther, het is zo jammer dat er geen markt is voor een forensische muziekpoli voor patiënten met ADHD. En tot slot Brian Comanne, wat had ik graag gewild dat je erbij kon zijn als ik mijn proefschrift verdedig. Ik ben nog steeds erg dankbaar voor je vriendschap.

Lieve familie, wat fijn dat jullie voor mij klaar hebben gestaan en altijd geïnteresseerd zijn gebleven in mijn proefschrift. Lieve Warre en Rifka, er was ooit een tijd dat ik dacht dat promoveren één van de mooiste dingen zou zijn die je in je leven meemaakt. Wat een feest dat jullie er zijn om te bewijzen dat er nog veel mooier dingen zijn! Lieve Feico, mijn lief, zonder jou stond al het voorgaande er niet. Mijn dank is groot, mijn liefde groter.