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.