ASSESSING THE EFFICACY OF PSYCHOLOGICAL TREATMENTS FOR MAJOR DEPRESSION:

An Investigation of Methodological Issues.

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Martin Connor.

May 2013

Abstract

Major depressive disorder (MDD) is a substantial healthcare concern. Conventionally conducted meta-analyses support the efficacy of both psychological and pharmacological interventions for MDD, but methodological limitations of meta-analyses may obfuscate rather than clarify the clinical efficacy of available interventions.

The thesis begins with a systematic review of meta-analyses of high quality psychological treatment studies for MDD. The results of the systematic review indicated that 48% of patients achieved remission after a course of psychological treatment. However, approximately 70% of remitted patients relapsed within 3 years after the discontinuation of psychological therapy. Consistent methodological limitations were identified in the primary outcome studies contributing to the meta-analyses. The primary studies typically published insufficient evidence on treatment fidelity. There was considerable variability in the overall treatment duration, the mean severity of samples and the definition of clinical significance. These factors pose a risk to the validity of meta-analytic results of psychological interventions for MDD.

The next component of the thesis investigated the impact of idiosyncratic clinical significance definitions on the published conclusions of studies that used the Beck Depression Inventory (BDI) or Hamilton Rating Scale for Depression (HRSD) to assess outcome. The availability of individual patient data (IPD) for 7 published studies enabled the empirically-based Jacobson Method of clinical significance to be used as a standard definition of recovery across IPD studies. Comparisons of published and Jacobson method clinical significance rates for each IPD study showed that idiosyncratic outcome definitions typically overestimated treatment efficacy. Moreover, treatment efficacy was confounded with the definition of clinical significance employed. This indicates that to reduce the risk of bias in meta-analysis, a standard and empirically-based definition of clinical significance should be used across primary MDD treatment studies. Subsequently, the moderating role of pre-treatment severity on clinical significance rates was investigated via individual patient data meta-analysis. It was found that being male and having higher pre-treatment severity both predicted a significantly reduced likelihood of achieving clinical recovery.

It is evident that between-study methodological differences means that even high quality conventional meta-analyses of psychological treatments for MDD remain at risk of bias. The novel finding that gender significantly moderated treatment outcome indicates that IPD meta-analyses are both more powerful and flexible than conventional meta-analyses based

on summary data. Ideally, future meta-analyses of primary MDD treatment studies should be based on individual patient data.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This thesis is dedicated to my son and finest achievement Joseph Connor.

I would like to thank the following people in helping me complete my work:-

Supervisors at the University of Liverpool

Prof. Rumona Dickson, Prof. Peter Kinderman but especially Dr. Peter Fisher.
Whilst not directly involved in my supervision, I am also very grateful to Prof. Peter Salmon & Dr. Paula Byrne for their additional feedback and encouragement.

Contributors to the systematic review in study 1

Prof. R. Dickson:	Independent screening of abstracts
Dr. Y. Dundar:	Search filters in appendix A
	Independent quality assessment of included reviews
Dr P. Fisher:	Resolved disagreements between independent reviewers concerning review eligibility
Mr. C. Huntley:	Independent assessment of review eligibility & data extraction

Providers of individual patient data used in studies 2 & 3

Dr. M. J. Constantino, University of Massachusetts at Amherst

Dr. D. David, Babes-Bolyai University, Romania

Dr. J. J. M. Dekker, Vrije Universiteit & Mentrum Depression Research Group, Amsterdam

Dr. K. S. Dobson, University of Calgary (Jacobson et al., 1996)

Dr. R. J. DeRubeis & Dr. J. Fournier, University of Pennsylvania

Dr. R. B. Jarrett, University of Texas Southwestern Medical Centre at Dallas

Dr. J. K. Salminen, University of Turku, Finland.

Declaration

I, Martin Connor, declare that I am the author of this thesis; that unless otherwise stated, all references cited have been consulted by me; that unless otherwise stated, the work of which this thesis is a record has been done by myself and has not been previously accepted for a higher degree.

Martin Connor

May 2013

Chapter One

Overview of Thesis

As depression places many burdens on individuals and the wider economy it is imperative that highly efficacious treatments are available. Consequently, identifying those treatments which best promote remission of depressive symptoms is of great value to stakeholders, clinicians and individuals. However, this is no simple task, as remission rates for the same treatment type often differ markedly between treatment studies. The variability in treatment outcomes makes it difficult to assess the relative and absolute efficacy of psychological treatments for major depression. Synthesising outcome data across multiple outcome studies has become the preserve of meta analytic reviews. However, although meta-analyses can accommodate variability between the results of primary studies and are thus widely used in healthcare research, their application within reviews concerned with the efficacy of psychological treatments for depression has not led to consistent conclusions. The lack of consistency between the results of meta-analytic reviews has arisen for a variety of methodological reasons. For example, reviews have differed concerning which treatment studies should be included, as well as the specific statistical methods employed to perform meta- analysis.

This thesis uses the findings of meta-analyses of psychological treatment studies for major depression as a starting point to examine several methodological issues which contribute to between-study variability in treatment outcome. In addition, this thesis examines whether these factors reduce the validity of meta-analyses of psychological treatment studies for major depression. The research presented here is based on examinations of the methods and results of published meta-analyses as well as original analyses of patient outcome data obtained from the authors of published treatment studies.

The first part of this thesis presents a description of major depression in terms of its diagnosis, epidemiology and burden. Following this a historical review and methodological critique of meta-analysis in the context of psychological treatment studies for depression is presented. Here, methodological factors that potentially bias meta-analyses of depression treatment studies are described. Subsequently, study 1 examines the findings and methodological limitations of seven meta-analytic reviews of psychological treatment studies for major depression that obviated previously identified sources of bias. The most

reliable available evidence from meta-analytic reviews suggested that less than 50% of patients starting individual psychotherapy achieved remission. Whilst there was no evidence that medication was superior to psychological treatment in general, the results indicate that the efficacy of both treatment types may be considerably improved. However, study 1 revealed four methodological factors which reduced confidence in the validity of the conclusions reached in all meta-analytic reviews. First, it was possible that psychological treatments were poorly implemented in some of the studies included in reviews. Second, there was considerable variability between the included studies in reviews concerning (i) the duration and intensity of psychotherapy, (ii) the mean pre-treatment severity of samples and (iii) the methods used to define remission. Moreover, the employment of idiosyncratic remission definitions across included studies meant that it was unclear to what degree overall review findings represented the proportion of patients who genuinely achieved remission.

Quantifying the proportion of patients who achieve a clinically significant outcome across treatment studies requires the use of empirically-based and standardised definition. The Jacobson Method of clinical significance (Jacobson et al., 1984; Jacobson and Revenstorf, 1988; Jacobson and Truax, 1991) is ideally suited for this purpose. However, application of the Jacobson Method requires that individual patient data (IPD) be made available by the authors of primary studies. Following a review and critique of the Jacobson Method, it was employed in the two studies which comprise the second part of the thesis.

In study 2, the proportion of patients who recovered according to the Jacobson Method following psychotherapy for major depression was quantified. Individual patient data (IPD) was obtained from seven primary studies where treatment outcome was assessed using either the Beck Depression Inventory (BDI, Beck et al., 1961) or Hamilton Rating Scale for Depression (HRSD, Hamilton, 1960). The results showed that less than 50% of patients across IPD studies achieved recovery according to the Jacobson Method and that published rates for the treatments in individual IPD studies could differ considerably from their corresponding Jacobson Method recovery rate. The use of idiosyncratic published outcome definitions also meant that the rank-ordering of treatment efficacy in IPD studies could differ according to published or Jacobson Method definitions of clinical significance. Finally, when recovery according to the Jacobson Method was used to compare measures, poor agreement was found between the BDI and HRSD in samples assessed on both. Overall, the results of study 2 indicated that conclusions concerning the relative efficacy of treatments within individual studies may be confounded with the definition of clinical significance employed. Consequently, the failure by primary researchers to employ a standard definition

of clinical significance risks that meta-analyses investigating the absolute or relative efficacy of depression treatments will be biased.

In study 3, the availability of individual patient data (IPD) meant that it was possible to investigate whether pre-treatment severity was predictive of recovery as defined by the Jacobson Method. Consequently, IPD meta-analyses employing hierarchical binary logistic regression with recovery status as the dependent and pre-treatment severity as the independent variable were undertaken. Separate analyses for the BDI and HRSD both controlled for study, treatment type (psychotherapy or medication) and gender. The results of the HRSD analysis showed that increasing pre-treatment severity predicted a reduced probability of recovery. However, the results of the BDI analysis showed that increasing pre-treatment severity predicted a reduced probability of recovery in females only. Pre-treatment severity on the BDI was not a significant predictor of outcome in males. Overall, the results of study 3 revealed that at lower severities, females were significantly more likely to recover than males of an equivalent pre-treatment severity on either measure. Only in severe cases was the probability of recovery no different between genders. The identification of a significant gender difference using this novel approach contrasts with previous research that identified no gender difference in response to psychotherapy (Parker et al., 2011). Consequently, the IPD meta-analytic approach may be a more powerful method by which to investigate factors that moderate outcome across depression treatment studies.

Chapter Two

The Nature of Major Depression

2.1 Introduction

It has been recognised for centuries that sadness and despair are a common experience for many people. Historical accounts indicate that the cause of severe mood disturbance was typically ascribed to physical illness for which the sufferer bore no responsibility. Symptoms of historical melancholia included extreme sadness, an inability to function and the frequent presence of delusions (Daly, 2007). Thus, historical melancholia may be a description of modern bipolar disorder or severe unipolar depression (Akiskal and Akiskal, 2007). The extreme nature of melancholia meant that its cause was attributed to an imbalance of the 'bodily humours' (Daly, 2007; Akiskal and Akiskal, 2007). However, historical accounts also describe less severe mood problems for which the sufferer was believed ultimately responsible. The 'sin' of acedia (Daly, 2007) originated in early Christian monastic settings and referred to a constellation of undesirable feelings and behaviours which interfered with devotional duties (Jackson, 1981). These were attributed to laziness or a 'lack of care' and included apathy, loss of hope, drowsiness and a desire to flee the monastery (LaMotte, 2007). However, acedia was not considered equivalent to normal sadness as the 4th century monk John Cassian described it as a 'dangerous foe' that was 'akin to sadness' (Daly, 2007; p34). These historical descriptions of the 'symptoms' of melancholia and acedia loosely correspond to those of major depression as defined by modern diagnostic systems.

This chapter describes major depression in terms of its diagnosis, epidemiology and the considerable burden it places on both the individual and wider economy. It will become apparent that major depression is a common but clinically heterogeneous disorder that is frequently comorbid with other disorders. Moreover, whilst many individuals with major depression will experience a single episode, a substantial minority will experience recurrent episodes. The heavy personal, social and economic burdens associated with major depression demand that effective treatments are available.

2.2 Diagnosing Major Depression

As there are no reliable physiological markers to denote the presence of major depression, current diagnostic methods rely on identifying psychological and behavioural symptoms (APA, 2000). The two major classificatory schemes by which major depression is diagnosed are the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM IV; APA, 1994) and the World Health Organisation's International Classification of Diseases (ICD-10; WHO, 1993). Because major depression is a highly recurrent disorder (Boland and Keller, 2008), both systems operationalise it in terms of the occurrence of a single 'depressive episode' (WHO, 1992) or 'major depressive episode' (MDE, APA, 2000). The diagnostic criteria for a depressive episode are similar in both systems. The DSM IV and ICD-10 both define recurrent depression as the occurrence of two or more episodes which are separated by at least 2 months where the criteria for a depressive episode are not met (APA, 2000; WHO, 1993). In DSM IV, the term Major Depressive Disorder (MDD) is used to denote the occurrence of one or more major depressive episodes and is thus synonymous with major depression.

In addition to being highly recurrent, major depression is also a clinically heterogeneous disorder (Rush, 2007). The diagnostic criteria of both the DSM IV and ICD-10 systems were designed to account for such heterogeneity. However, this means that depressed individuals with markedly divergent symptoms are assigned to the same diagnostic category (APA, 2000; Krueger et al., 2005). For example, two individuals diagnosed with a major depressive episode may both experience depressed mood and concentration difficulties. However, one individual may have accompanying symptoms of significant weight loss and insomnia, whilst the other experiences significant weight gain and hypersomnia. Because such differences may be important for the selection of appropriate treatment and thus prognosis (APA, 2000; WHO, 1992; Rush, 2007), both the DSM IV and ICD-10 systems enable the specification of depressive sub-type and episode severity (APA, 2000; WHO, 1992).

Whilst the DSM IV and ICD-10 diagnostic systems are very similar, sufficient differences exist which can make direct comparisons between them problematic. To illustrate, the degree of functional disability associated with the presenting problem cannot be used to support a diagnosis within the ICD-10 framework (WHO, 1992). In contrast, diagnosis within the DSM IV system explicitly requires that the disorder is sufficiently severe to cause clinically significant distress or impairment in social, occupational, or other important areas of functioning (APA, 2000). It is theoretically possible that individuals will meet the diagnostic criteria for a depressive episode according to ICD-10 but not DSM IV criteria as

they are not sufficiently distressed or impaired according to the latter system. However, Spitzer & Wakefield (1999) have argued that the DSM IV requirement of clinical distress is redundant for a diagnosis of MDD because it is highly unlikely that those meeting the DSM IV symptom criteria alone would not be distressed. This suggests that the ICD-10 and DSM IV systems will show high levels of agreement concerning individual diagnoses because their symptom criteria are very similar. However, because the focus of this thesis concerns the findings of psychological treatment studies that typically employ the DSM system, the DSM IV criteria for MDD will be presented in detail. Important differences between DSM IV and ICD-10 diagnostic criteria will be described where appropriate.

2.2.1 Diagnostic Criteria for Major Depressive Disorder

The current DSM IV (APA, 2000) is based on successive revisions of the DSM III (APA, 1980). The DSM III marked a radical departure from previous versions by providing explicit criteria by which to reach a diagnosis (Decker, 2007). By organising mental disorder in terms of prototypical symptom-based categories (Krueger et al., 2005) and avoiding theoretical issues concerning the aetiology of disorders, the DSM III led to both improved diagnostic reliability and a restoration of the scientific status of American psychiatry (Decker, 2007). Table 1 presents the current DSM IV diagnostic criteria for a major depressive episode.

According to DSM IV, the diagnosis of a major depressive episode requires that all criteria from A to E in Table 1 are met. An inspection of criterion 'A' shows that at least 5 of the 9 symptoms must be present nearly every day for at least two weeks and that one of these must be either depressed mood or a marked loss of interest or pleasure in most activities. The ICD-10 also requires that symptomatic criteria for a depressive episode are met for at least two weeks. However, the two systems use markedly different criteria to establish the presence of a depressive episode. The DSM IV requires that at least 5 of the criterion symptoms in Table 1 are present irrespective of episode severity. In contrast, the minimum number of symptoms required to meet diagnostic criteria in the ICD system varies according to episode severity. For example, a mild episode according to ICD criteria requires the presence of only four criterion symptoms, two of which must be typical in depression, i.e. depressed mood, loss of interest or increased fatigue (WHO, 1992).

However, a severe episode according to ICD criteria requires the presence of all 3 typical symptoms and at least four of seven additional symptoms. These are reduced concentration/attention, lowered self esteem/confidence, guilt, pessimism concerning the future, suicidality, sleep difficulties and diminished appetite (WHO, 1992). Thus, whilst

both systems use very similar symptomatic criteria[1], only the ICD system incorporates symptom count in defining the severity of depressive episodes. However, ICD-10 descriptions concerning the functional impact of mild, moderate and severe episodes correspond closely to those of DSM IV.

Table 1. DSM IV-TR Diagnostic Criteria For A Major Depressive Episode

A	Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.[1]
	1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful).[2]
	2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others)
	3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day.[3]
	4. Insomnia or hypersomnia nearly every day
	5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down)
	6. Fatigue or loss of energy nearly every day
	7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick)
	8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others)
	9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide
B	The symptoms do not meet criteria for a Mixed Episode
C	The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
D	The symptoms are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition (e.g., hypothyroidism).
E	The symptoms are not better accounted for by Bereavement, i.e., after the loss of a loved one, the symptoms persist for longer than 2 months or are characterized by marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation.

Notes:

1. Do not include symptoms that are clearly due to a general medical condition, or mood-incongruent delusions or hallucinations.
2. In children and adolescents, can be irritable mood.
3. In children, consider failure to make expected weight gains.

[1] In strict terms, ICD-10 excludes weight changes and increased appetite whereas DSM IV excludes pessimism.

Returning to DSM IV, a closer inspection of criterion 'A' in Table 1 reveals that individuals meeting diagnostic criteria for a depressive episode need not share any symptoms in common. In theory, two individuals may share none of the 9 symptoms in Table 1 because items 3 to 5 allow for increases *or* decreases in weight, sleep and psychomotor activity respectively. Nevertheless, researchers and clinicians have observed what appear to be relatively consistent constellations of depressive symptoms which may respond differently to treatment (Rush, 2007). Consequently, successive revisions of the DSM since version III have included specifiers that enable potentially important clinical characteristics of episodes to be recorded (APA, 2000). These episode specifiers are described below and concern symptom severity, remission status, chronicity and symptomatic features that may denote depressive sub-types.

2.2.2 DSM IV Episode Specifiers

Severity

The DSM IV categorises severity as mild, moderate or severe across three different domains. These are the number of criterion symptoms, their severity, and the degree of associated disability and distress. A mild episode has sufficient but perhaps no more symptoms than required for diagnosis. However, it must be associated with mild disability or reduced functionality that takes a substantial effort to overcome. A severe episode *without psychotic features* is characterised by the presence of a majority of criterion symptoms and evident occupational or social disability (e.g. inability to work). A severe episode *with psychotic features* is characterised by mood-congruent delusions or hallucinations. A moderate depressive episode is characterised as falling between mild and severe.

The utility of the three DSM IV severity domains in predicting a range of clinically important phenomena was investigated by Lux et al. (2010). They found that the three severity domains were not equally effective in predicting individual clinical phenomena (Lux et al., 2010). For example, symptom count was the strongest predictor of concurrent anxiety, whereas symptom severity was the strongest predictor of episode duration, and disability the strongest predictor of lifetime comorbid anxiety (Lux et al., 2010). Lux et al. (2010) concluded that depressive severity is a multifaceted and heterogeneous construct that cannot be fully captured according to any single domain described in the DSM IV. However, treatment studies that quantify depressive severity using symptom measures have shown that more severe patients typically require longer treatment (Shapiro et al., 1994), are less likely to benefit from non-specific treatment effects (Schatzberg and Kraemer, 2000) and make relatively poorer symptomatic improvements compared to less severe cases (Jarrett et al., 1991; Frank et al., 2011).

Remission status & chronicity

Remission is a vital concept in the assessment of treatment outcome, as it implies that an individual's functioning is no longer impaired (Keller, 2003; Boland and Keller, 2008). Remission refers to a level of symptomatic improvements sufficient to no longer meet diagnostic criteria (Romera et al., 2011; Keller, 2003). Remission according to the DSM IV (APA, 2000) may be full or partial. Full remission denotes a period of at least 2 months where significant symptoms of depression are absent. Partial remission refers to either (i) the absence of significant criterion symptoms for a period less than 2 months, or (ii) the presence of some criterion symptoms but the full criteria for a major depressive episode are no longer met (APA, 2000; p 412). Thus, irrespective of whether it is full or partial, the DSM IV defines remission as not meeting diagnostic criteria for an MDE for at least 2 months. DSM IV defines chronic depression as the failure to achieve remission over a period of 2 years or more (APA, 2000).

Depressive features

In order to increase diagnostic specificity within the clinically heterogeneous diagnostic category of major depressive disorder (Fink et al., 2007), the DSM IV includes depressive feature specifiers that identify potentially important symptom constellations (APA, 2000). Depressive features specifiers include melancholic features and atypical features, which for ease of presentation are described here as depressive sub-types. According to DSM criteria, the major feature of the melancholic sub-type is that the depressed mood is qualitatively different from that seen in non-melancholic depression; there is virtually no ability to experience pleasure and the depressed mood is not reactive to pleasurable stimuli (APA, 2000). The melancholic specifier also requires that three of the following five symptoms are present; mood worse in the morning, early awakening, weight changes, psychomotor changes, and excessive guilt. In contrast, the major feature of the atypical sub-type is that the depressed mood is reactive to pleasurable stimuli (APA, 2000). The atypical specifier also requires that two of the following four symptoms are present; increased appetite/weight, increased need to sleep (hypersomnia), feeling heavy or weighed down (leaden paralysis), and a long-standing pattern of pathological sensitivity to perceived rejection (APA, 2000). It must be noted that atypical features are not uncommon as they are estimated to occur in between 15 to 36 percent of depressed individuals (Cristancho et al., 2011). The term 'atypical' was originally used to differentiate between patients who responded significantly more favourably to monoamine oxidase inhibitors than imipramine (Thase, 2009).

The DSM IV specifications of melancholic and atypical features have been criticised as having limited utility. Firstly, where symptomatically distinct sub-groups are valid, they

ought to enable the identification of between group differences in either the course or treatment of depression (Rush, 2007). However, the stability of both DSM IV sub-types across separate episodes is poor (Melartin et al., 2004) and melancholic features are frequently concurrent with atypical features within the same individual (Angst et al., 2007). Moreover, after controlling for pre-treatment severity, Melartin et al. (2004) found no significant group differences between DSM IV melancholic and non-melancholic sub-types in treatment outcome.

However, a second criticism of both of these specifiers is that their respective DSM IV criteria do not adequately capture the symptomatology of melancholia or atypical depression on which they were originally based. This shortcoming may be responsible for study findings that have failed to support their clinical utility as the misclassification of non-melancholic (or non-atypical) patients will serve to confound results. For example, Taylor & Fink (2008) assert that melancholia is a well defined clinical syndrome that can be reliably diagnosed according to the following criteria: significantly reduced functioning due to an unremitting depressed mood, the presence of marked psychomotor disturbance and at least two vegetative symptoms from impaired sleep, appetite, libido or cognition (Taylor and Fink, 2008). Moreover, melancholic patients should demonstrate abnormally high cortisol levels or disturbed sleep patterns (Taylor and Fink, 2008). Consequently, Taylor & Fink (2008) argue that the current DSM IV melancholic features specifier is insufficiently stringent. Similarly, there is strong evidence for the existence of MDD with atypical features (Stewart et al., 2009). However, it has been proposed that the atypical features criteria of the DSM IV need revising, as they fail to reliably identify atypical cases (Thase, 2009). The most important shortcoming of the DSM criteria is that the presence of mood reactivity, which is the only obligatory feature for DSM atypical depression, is both redundant and does not discriminate between atypical and non-atypical cases (Thase, 2009).

In summary, the DSM IV melancholic and atypical depressive features specifiers were intended to enable clinicians to discriminate between patient sub-groups who might benefit from different treatment approaches. However, evidence has shown that they are of little prognostic value, possibly because the current DSM specifiers do not adequately reflect the criteria by which these sub-types were originally formulated. Consequently, until the controversy surrounding what constitute valid criteria for both are resolved, the melancholic and atypical sub-types may be better viewed as merely descriptive.

2.3 Course

Depression was once thought to be an acute and self-limiting disorder. However, it is now recognised that many remitted individuals will experience more than one episode and that the course of symptoms over time is highly variable. Early research into the course of MDD was characterised by the use of inconsistent terminology which hindered the identification of potentially important clinical information (Boland and Keller, 2008). In 1991 the MacArthur Foundation (Frank et al., 1991) recommended the use of standardised operational definitions of change points to describe the course of MDD within clinical trials. Their standardised definitions of remission, recovery, relapse and recurrence were summarised and updated in 2006 by the ACNP Task Force on Response and Remission in Major Depressive Disorder (Rush et al., 2006). The ACNP Task Force recommended that response should not be used in group comparison trials due to difficulties in its operationalisation (Rush et al., 2006). For example, whilst response is often defined as a minimum reduction of 50% on symptom measures in treatment trials, its clinical meaningfulness concerning individual patients is highly dependent on their pre-treatment severity score (Rush et al., 2006).

The ACNP Task Force defined remission as no longer meeting the criteria for an MDE according to DSM IV. However, in contrast to the DSM IV definition which requires 2 months, the ACNP definition requires only a 3 week period where no significant DSM IV criterion symptoms are present. Also, the ACNP criteria concerning what constitutes significant depressive symptomatology are more explicit than DSM IV descriptions. According to the ACNP Task Force remission requires that 3 or less DSM IV criterion symptoms are present over 3 weeks, none of which may be depressed mood or diminished interest. The ACNP Task Force provided definitions of recovery and relapse which are not used in DSM IV. Recovery is conceptualised as the point in time where an MDE is unlikely to occur in the near future (Rush et al., 2006; p1847) and is defined as a period of remission lasting more than 4 months. Thus, the ACNP definitions of remission and recovery differ only in terms of the length of time that a patient no longer meets DSM IV diagnostic criteria for an MDE. However, because biomarkers that reliably differentiate remission from recovery have not been identified, it is possible that the ACNP's proposed distinction between remission and recovery is not valid (Rush et al., 2006). Finally, the terms relapse and recurrence both refer to symptomatic worsening which means that an individual again meets the criteria for an MDE (Rush et al., 2006). However, the ACNP Task Force recommended that relapse be used to denote that diagnostic criteria are again met prior to recovery (i.e. during remission but before recovery), whereas recurrence be used to denote

the occurrence of a new major depressive episode after recovery has been established (Rush et al., 2006).

Prior to the publication of the ACNP Task Force recommendations in 2006 (Rush et al., 2006), individual studies typically employed idiosyncratic criteria to define what are nominally the same clinical outcomes (Keller, 2003). Consequently, remission, recovery, relapse and recurrence rates from such studies are likely to only approximate those that would be obtained according to ACNP definitions. This must be borne in mind for the majority of the research findings presented in the remainder of this chapter.

2.3.1 Duration of Untreated Episodes

There have been few naturalistic studies of the duration of untreated major depressive episodes for ethical reasons. However, prospective data suggest that the majority of cases will remit within one year and that the duration of episodes is higher in more severe cases. However, at 2 years it is likely that a substantial minority will meet criteria for chronic depression.

Prospective population based estimates for the duration of depressive episodes were obtained in the Netherlands Mental Health Survey and Incidence Study (NEMESIS, Spijker et al., 2002). Of 250 respondents who experienced a new episode according to DSM III-R criteria, the proportions recovered[2] were 50% at 3 months, 76% at 12 months and 80% at 21 months. Spijker et al. (2002) reported that higher severity or comorbid dysthymia predicted longer episodes, whilst recurrent depression predicted shorter episodes (Spijker et al., 2002). Posternak et al. (2006) found similar result for a sample of 130 non-chronically depressed patients who experienced a new episode over 15 years within the National Institute of Mental Health's Collaborative Program on the Psychobiology of Depression study (CDS, Katz et al., 1979). Of 84 individuals who did not receive any form of somatic treatment for a new MDE diagnosed according to the Research Diagnostic Criteria (RDC, Spitzer et al., 1978), the proportions recovered[3] were 38% at 3 months, 70% at 12 months and 75% at 2 years (Posternak et al., 2006). However, the CDS results failed to show that recurrent depression predicted shorter episodes (Solomon et al., 1997).

The NEMESIS results presented above do not account for treatment status. However, Spijker et al. (2002) found no significant difference in mean episode duration between those who did (67%) or did not receive treatment (33%). To explain this finding, Spijker et al.

[2] Defined as no or minimal symptoms over 3 months.
[3] Defined as no or minimal symptoms over 8 consecutive weeks.

(2002) suggested that treatment seekers were more likely to be severely depressed and would thus have experienced longer episodes had they not received treatment. The results of both the CDS and NCS-R studies appear to support this explanation. Firstly, non-treatment seekers[4] in the CDS achieved remission more rapidly than the sample as a whole which implies they had a better prognosis (Posternak et al., 2006). Secondly, higher severity in the NCS-R was predictive of longer episode duration (mild duration = 13.8 weeks; very severe duration = 23.1 weeks; Kessler et al., 2003).

Whilst it is possible that more severely depressed individuals were more likely to seek treatment and thus bias the results of the studies presented here, the overall results suggest that between a third to a half of cases will remit according to DSM IV criteria after 3 months and the majority by 12 months. However, approximately one fifth of cases will continue to be depressed at 12 months.

2.3.2 Recurrence

The onset of a first major depressive episode typically follows distressing life events such as bereavement, or divorce (APA, 2000). However, the onset of subsequent episodes are less likely to be preceded by an obvious cause (APA, 2000). The prospective CDS study (Katz et al., 1979) has provided important information concerning the naturalistic course of depression over 2 decades. The results indicate that recurrence is very common in patients who seek treatment for MDD and that the time between episodes typically decreases with increasing number of episodes. An important factor that may serve to both reduce the time to recurrence and increase the frequency of episodes is the persistence of residual depressive symptomatology during recovery. The experience of 3 or more major depressive episodes significantly increases the risk of recurrence.

The CDS results (Katz et al., 1979) showed that 22% of a sample of 141 non-dysthymic patients experienced a recurrence within the first year following recovery[5] (Keller et al., 1983). The risk of recurrence was highest immediately after the establishment of recovery but reduced consistently during follow-up (Keller et al., 1983). Over the longer term, recurrence rates at 5, 10, 15 and 20 years in the CDS were 60%, 75%, 87% and 91% respectively (Boland and Keller, 2008). However, an important finding was that the occurrence of 3 or more previous episodes predicted a significantly increased risk of recurrence which was estimated to increase by 16% following each episode (Solomon et al., 2000). In addition, whilst individuals did not demonstrate consistent time patterns between

[4] i.e. those who did not receive somatic treatments
[5] Defined as no or minimal symptoms over 8 consecutive weeks.

episodes, the overall results showed that the time between episodes decreased as the number of episodes increased. For example, the median time to recurrence following a first episode was 150 weeks, whereas it was 57 weeks following a fifth episode (Solomon et al., 2000). A consistent finding was that the rate and timing of new episodes was associated with the level of residual symptoms in recovered patients. Full recovery led to fewer recurrent episodes that were less frequent than recovery with residual symptoms. For example, recurrence rates in asymptomatic and symptomatic but recovered patients were 66% and 87% respectively; the mean time to recurrence for these groups were 180 and 33 weeks respectively (Boland and Keller, 2008).

2.4 Epidemiology of Major Depression

Surveys of the prevalence of psychiatric disorders have been undertaken since the Second World War. However, estimates of prevalence varied widely due to differences in methodology. Following the Second World War, the prevalence of MDD was typically assessed using screening instruments (Kessler et al., 2007). This method was severely limited as (i) screening instruments were prone to poor specificity or sensitivity leading to inaccuracy (ii) the use of different instruments between surveys makes comparing their results problematic. This has become less of an issue since the World Health Organisation commissioned the Composite International Diagnostic Instrument in the 1980s (CIDI, Kessler and Ustun, 2004) to compare psychiatric prevalence rates between countries according to standardised criteria (Kessler et al., 2007). The CIDI was based on the Diagnostic Interview Schedule (Robins et al., 1981) and was designed to be administered by lay interviewers. The CIDI was also designed to support psychiatric diagnoses according to both ICD and DSM criteria (Kessler and Ustun, 2004). However, the original version of the CIDI was not designed to capture detailed demographic and clinical data. This meant that countries could only be broadly compared in terms of overall prevalence rates (Kessler and Ustun, 2004).

The latest CIDI (version 3) was designed for the World Mental Health Survey Initiative (WMHS, Kessler, 1999) to facilitate the acquisition and comparison of psychiatric epidemiological data within participating countries (Kessler and Ustun, 2004). In addition to enabling the quantification of lifetime and 12 month diagnoses according to both DSM-IV and ICD–10 criteria, the CIDI-3 also includes items that assess severity, demographic, quality of life and disability data (Kessler and Ustun, 2004). Unlike previous versions, the CIDI-3 includes interview probe questions that increase the reliability of autobiographical

recall. These were designed to produce less biased estimates of prevalence and age of first onset data than previous versions (Kessler et al., 2010). However, any method that employs a retrospective approach remains likely to underestimate lifetime prevalence rates. For example, Wells & Horwood (2004) reported that only 44% of 25 year olds with a previous diagnosis of MDD recalled either of the key DSM IV symptoms of depression (items 1 & 2 in Table 1). Despite this potential limitation, the methodological rigour used to produce different translations of the CIDI-3 has led to it being described as 'state of the art' for comparing epidemiological findings across participating WMHS countries (Alonso and Lepine, 2007). Two large scale surveys within the WMHS framework have specifically examined the epidemiology of MDD. These are the European Study of the Epidemiology of Mental Disorders (ESEMeD, Alonso et al., 2002) and the American National Comorbidity Survey Replication Study (NCS-R, Kessler et al., 2003).

2.4.1 Prevalence

The NCS-R and ESEMeD surveys estimated that the 12 month prevalence of MDD according to DSM IV criteria is 6.6% in American and 4.1% in European adults (Alonso et al., 2004; Kessler et al., 2003). In terms of lifetime rates, 16.2% of Americans and 13.4% of Europeans will experience at least one depressive episode. The difference between these surveys may reflect that rates are genuinely higher in the American population. However, it is also possible that between-country difference concerning the stigma associated with mental disorder led to under-reporting across European countries as a whole (Bernert et al., 2009). However, both the ESEMeD and NCS-R lifetime prevalence rates are likely to be underestimates due to recall biases (Wells and Horwood, 2004) and the proportion of never-depressed individuals surveyed who will meet MDE diagnostic criteria in the future (Kessler and Wang, 2008).

The NCS-R results revealed that at least 13.1 million US adults met DSM IV criteria for a major depressive episode in the preceding year (Kessler et al., 2003). In terms of DSM IV symptomatology, the NCS-R results estimated that 10.4% of 12 month cases were mild, 38.6% moderate, 38% severe and 12.9% very severe according to the Quick Inventory of Depressive Symptomatology Self Report (QUIDS-SR; Rush et al., 2003; Kessler et al., 2003). Thus, a substantial proportion of 12 month cases demonstrated severe or very severe clinical symptoms in the NCS-R sample. Given that current demands for treatment exceed available resources, it would appear that the most efficient use of resources is to target only those who experience severe and persistent depression (Kessler, 2007). However, providing effective treatment for the large population of less severe cases could be an effective preventative strategy (Kessler, 2007).

2.4.2 Age Differences

Results from the ESEMeD study showed that the 12 month prevalence for any mental disorder is highest in the 18 to 24 year age group and lowest for individuals over 65 (Alonso and Lepine, 2007). Comparable results for the prevalence of MDE were found in the NCS-R where 12 month and lifetime rates in the youngest cohort (18 to 29 years) were significantly higher than those over 60 years (Kessler et al., 2003). However, the DSM IV employs hierarchical exclusion rules that typically prohibit a diagnosis of MDE in the presence of significant physical comorbidity (APA, 2000). The lower 12 month prevalence rate for older cohorts in the NCS-R may be artefactual, as higher levels of physical comorbidity in older adults may have precluded the diagnosis of a depressive episode (Kessler et al., 2010). To investigate this possibility, Kessler et al. (2010) re-analysed the WMHS data by omitting the hierarchical and organic exclusion rules normally employed. This meant that depression comorbid with a physical disorder was included in analyses.

The WMHS comorbidity study (Kessler et al., 2010) showed that the pattern of MDE prevalence by age reported in the NCS-R (Kessler et al., 2003) was similar to that seen in most developed countries. In Belgium, France, Germany, Israel, Italy, Japan, the Netherlands, New Zealand, Spain and the USA, the 12 month MDE prevalence rate was significantly lower in the oldest cohort compared to the youngest. Only Israel, Italy and Spain did not demonstrate significant age related differences in the 12 month prevalence of MDE. It appears that higher rates of physical comorbidity were not responsible for the lower rates of depression typically observed for older cohorts in developed countries (Kessler et al., 2010). An analysis across all WMHS developed countries showed that 12 month MDE prevalence for the oldest cohort was significantly lower than for the youngest cohort (Kessler et al., 2010). The WMHS 12 month MDE prevalence rate for each age cohort across developed countries was 7% (18 to 34 years), 6% (35 to 49 years), 5.1% (50 to 64 years) and 2.6% (65+ years). However, only Brazil reproduced the overall pattern observed across developed countries. A pooled analysis across Brazil, Colombia, India, Lebanon, Mexico, South Africa and Ukraine failed to demonstrate the existence of a 12 month MDE cohort effect across developing countries (Kessler et al., 2010).

In addition to providing data concerning age differences in the prevalence of MDE, the WMHS comorbidity study also enabled overall 12 month prevalence rates to be compared across countries. The results showed that overall estimated 12 month MDE prevalence rates across developed and developing countries were similar at 5.5% and 5.9% respectively (Kessler et al., 2010). However, individual countries showed marked variability in

prevalence rates which ranged from 2.2% (Japan) to 8.3% (USA) in developed countries, and from 4% (Mexico) to 10.4% (Brazil) in developing countries (Kessler et al., 2010).

The results of the WMHS comorbidity study suggest that the mean age of onset in younger cohorts is decreasing across both developed and developing countries (Kessler et al., 2010). Kessler et al (2010) proposed that these findings were plausible as the time interval between respondents' mean age and reported lifetime onset also increased with age. This result contrasted with previous findings showing that the mean age of onset across all age cohorts typically cluster in the ten years prior to interview due to retrospective reporting bias (Simon and Vonkorff, 1992). The WMHS comorbidity study results also revealed that the prevalence of severe episodes, as assessed using the QUIDS-SR, demonstrate an inverse relationship with age in developed countries as severe episodes were significantly less prevalent in the 65+ age group (19.6%) compared to younger cohorts (range 29.6% to 39.7%). This relationship was not seen for developing countries (Kessler et al., 2010). Finally, the WMHS comorbidity study results suggest that episode duration increases with age (Kessler et al., 2010). Mean episode duration in the 12 months prior to interview to significantly increased with age across both developed and developing countries. In developed countries, the mean episode for the youngest cohort lasted 25 weeks compared to 31 weeks for the oldest cohort (Kessler et al., 2010). However, these results may be confounded by treatment seeking differences between cohorts, as approximately 59% of respondents reported receiving some form of treatment.

2.4.3 Gender Differences

One of the most consistent epidemiological findings concerning MDD is that female prevalence rates are typically twice those seen in males (Boughton and Street, 2007). Both the ESEMeD and NCS-R studies found 12 month and lifetime MDD prevalence rates for females were approximately twice those for males.

Higher female prevalence is known to emerge in adolescence and continue into adulthood (Boughton and Street, 2007) although no significant gender differences have been found in terms of recurrence or chronicity (Kessler et al., 1993; APA, 2000). However, the results of the U.K.'s National Survey of Psychiatric Morbidity (NSPB; Bebbington et al., 2003) have shown that the preponderance of female depression disappears after the age of 55 due to a reduction in the prevalence of female depression. Boughton & Street (2007) reviewed numerous non-biological theories proposed to explain the higher rates of depression seen in females. Some theories propose, that higher levels of neuroticism or dependency in females increase the risk for depression, whilst others attribute differences to social restrictions

imposed by the female role. Alternatively, the construct of major depression may itself be biased towards identifying disorder in females (Boughton and Street, 2007). Irrespective of whether the latter proposal is true, it has been shown that the choice of measure used to quantify depression severity determines whether females are rated as more severely depressed than males (Salokangas et al., 2002; Sigmon et al., 2005).

However, whilst many factors are likely to contribute to gender differences in the prevalence of depression, there is increasing evidence that gender differences concerning emotional regulation are a key factor (Nolen-Hoeksema, 2012). Emotional regulation refers to activities that enable the individual to modify the nature of an emotional response (e.g. distraction, Nolen-Hoeksema, 2012). However, whilst females have been shown to employ a wider range of emotional regulatory behaviours than men (Tamres et al., 2002), it has been proposed that their greater tendency to ruminate on the causes and meaning of negative emotions places a higher proportion of them at risk of developing depression (Nolen-Hoeksema, 2012). Evidence that greater rumination in females may explain their higher risk for MDD has been provided in studies that show rumination to be predictive of higher depression scores (Nolen-Hoeksema, 2000; Nolen-Hoeksema and Aldao, 2011; Nolen-Hoeksema et al., 1997). Moreover, because it is associated with an increased risk for social phobia, generalised anxiety disorder and post-traumatic stress disorder (Nolen-Hoeksema, 2012), rumination is likely to be a key transdiagnostic risk factor for the development of several psychological disorders (Rachman, 1971; Nolen-Hoeksema, 1987; Morrow and Nolen-Hoeksema, 1990; Nolen-Hoeksema et al., 1993).

2.4.4 Comorbidity

Major depressive disorder is highly comorbid with psychological (Rush et al., 2005) and somatic disorders (Schmitz et al., 2007). In the NCS-R study 64% of 12 month MDD cases also met diagnostic criteria for another DSM IV 12 month disorder (Kessler et al., 2003). However, whilst MDD was highly comorbid with other psychological disorders, it only preceded other 12 month disorders in 12.6% of cases (Kessler et al., 2003). Estimates for the prevalence of major depression comorbid with physical disorders range from 5% to 10% in primary care settings, and from 8% to 15% in medical inpatient settings (Schmitz et al., 2007). Comorbid depression is associated with greater levels of disability and poorer prognosis for both psychological and physical disorders (Rush et al., 2005; Schmitz et al., 2007).

Where depression is comorbid with a physical disorder, the greatest impairments are seen in those who experience chronic physical problems. The Canadian Community & Health

Survey (Schmitz et al., 2007) revealed that the prevalence of functional disability in the 2 weeks prior to interview was significantly higher in respondents with chronic physical disorders and comorbid MDD (46%) compared to those with only chronic physical disorders (21%) or only MDD (27.8%). One of the most striking findings concerning the effect of comorbid depression and physical illness concerns cardiac mortality. In patients hospitalised for myocardial infarction, Lesperance et al. (2002) found a direct dose-response relationship between depressive symptomatology on the Beck Depression Inventory (BDI, Beck et al., 1996) and the risk of cardiac mortality during 5 year follow-up. Notably, the mortality rate in patients who scored 19 or more on the BDI was significantly higher than those scored less than 19 on the BDI after controlling for cardiac disease severity (Lesperance et al., 2002). These results suggest that comorbid depression is associated with increased mortality during recovery from myocardial infarction.

Where another psychological disorder is comorbid with MDD, episodes are typically more severe and last longer (Rush et al., 2005). As described earlier, there is evidence comorbid dysthymia increases the duration of depressive episodes (Spijker et al., 2002). However results from the naturalistic CDS study also indicated that comorbid panic (Coryell et al., 1988) or alcohol abuse (Mueller et al., 1994) reduce the likelihood of recovery from an MDE. Coryell et al. (1988) found that comorbid panic and MDD predicted significantly lower levels of recovery compared to non-comorbid cases (75% versus 86% respectively) over 2 years, whilst Mueller et al. (1994) found that comorbid alcoholism reduced the likelihood of recovery by 50% over an observation period of 10 years. However, neither of these studies controlled for treatment differences in their analyses. Nevertheless, they provide strong evidence that comorbidity serves to increase episode duration and suggests treatment efficacy will be lower in patients with comorbid conditions.

The moderating effect of comorbidity on treatment outcome has received relatively little attention (Carter et al., 2012; Hamilton and Dobson, 2002). However, there is consistent evidence that elevated anxiety symptomatology during an episode predicts poorer response to medication (Carter et al., 2012) and a lower probability of successful outcome following psychotherapy (Hamilton and Dobson, 2002). Given than anxiety disorders are highly comorbid with MDD, e.g. 57% of 12 month MDD cases met diagnostic criteria for at least one comorbid DSM IV anxiety disorder (Kessler et al., 2003), they are likely to be an important moderator of treatment outcome in MDD.

Finally, many previously remitted axis 1 disorders have not been identified as a risk factor for the development of a major depressive episode with the exception of early-onset simple

phobia and panic (Kessler & Wang 2008). However, Generalised Anxiety Disorder (GAD) has been identified as having the highest risk for the development of subsequent comorbid depression (Kessler & Wang 2008). The high levels of comorbidity between depression and anxiety disorders has been argued to be an artefact of changes in the diagnostic criteria for successive versions of the DSM which have allowed an increasing number of diagnoses to be made for the same individual (Kessler & Wang 2008). There have been suggestions that cases of comorbid anxiety and depression may stem from a common pathological process, and that the separation of the disorders in DSM III onwards has produced an artificial distinction for these patients (Frances, Manning et al., 1991). However, future research on the validity of differentiating between the two disorders is still required (Kessler & Wang 2008).

2.4.5 The Burden of Major Depressive Disorder

One of the most distressing aspects of mood disorders is the strong association with suicidal behaviour. Beautrais et al. (1996) reported that whilst 90% of patients hospitalised for attempted suicide met DSM III-R criteria for a psychiatric disorder, mood disorders specifically accounted for 80% of the attributable risk for serious suicide attempts (Beautrais et al., 1996) which themselves strongly predict completed suicide (Yoshimasu et al., 2008). Whilst it has been recommended that suicide prevention strategies should not focus solely on depression (Fleischmann et al., 2005), MDD itself is likely to be a major predictor of suicide as it accounted for approximately 28% of the attributable risk for suicide within the ESEMeD study (Bernal et al., 2007). It is estimated that up to 15% of severe MDD cases will die by suicide (APA, 2000). In addition to suicide, MDD is known to increase the risk of physical morbidity. For example, MDD has been shown to predict higher pain and mortality in medical inpatients (Herrmann et al., 1998; APA, 2000) and an increased likelihood of both admission to, and mortality in nursing homes (Onder et al., 2007; APA, 2000). In addition to poorer prognoses for cardiac patients with comorbid depression, MDD is itself a risk factor for the development of cardiac problems (Frasure-Smith and Lesperance, 2005).

Major depressive disorder is also a risk factor for a range of maladaptive behaviours. NCS-R data revealed that 45% of American respondents meeting DSM IV diagnostic criteria for substance use disorders in the previous 12 months also reported antecedent symptoms meeting criteria for an MDE (Kessler et al., 2003). This implies that depressive symptoms led to substance abuse in such cases. However, it cannot be ruled out that common factors lead to both disorders as the association between depression and substance abuse is complicated by the interaction between multiple factors (Swendsen and Merikangas, 2000).

When the onset of MDD occurs in adolescence it is associated with an increased risk of poor educational attainment, teenage pregnancy and impaired future marital relationships (Kessler and Wang, 2008). Within marital relationships, MDD is significantly associated with an increased risk of divorce due to impaired problem solving and communication (Davila et al., 2008). Moreover, where one partner has recovered from a depressive episode, the marital relationship may remain at risk as spousal negativity towards MDD has been shown to predict future episodes (Davila et al., 2008).

In addition to physical and behavioural burdens, MDD is costly to the wider economy. Major depression impairs work performance to a greater degree than arthritis, asthma, migraine, irritable bowel syndrome and hypertension (Kessler et al., 2008). Unsurprisingly, the economic impact of depression increases with increasing severity which leads to poorer work performance, increased risk of unemployment and the greater need for treatment (Birnbaum et al., 2010). The 1991 cost of treating MDD within the UK's National Health Service was estimated to be £417 million. However, the overall economic cost due to absence from work and premature mortality was far higher at nearly £3 billion (Churchill et al., 2001).

The importance of depression as a personal and economic burden is reflected in the World Health Organisation's projection that its contribution to the Global Burden of Disease (GBD) will rise from fourth place in 2001 to second place by 2020. In 2020 it is projected that the GBD associated with MDD will rank only behind that of ischemic heart disease. In developed countries it is projected to be the major burden of disease by 2020 (WHO, 2001). Thus, the identification and effective treatment of major depressive disorder is an increasingly pressing public health concern (WHO, 2001; WHO, 2008).

2.4.6 Treatment

Despite the high personal and economic costs associated with MDD, research indicates that depressed individuals show considerable delays in seeking treatment, that the recognition of depression is poor in treatment settings and that treatment is often inadequate.

The NCS-R provided data concerning the proportion of individuals with lifetime MDD who sought professional treatment (Wang et al., 2005). Treatment was defined in the NCS-R as any form of professional healing contact meaning that psychologists, counsellors, spiritual advisors and herbalists were included with conventional medical professionals (Wang et al., 2005). The NCS-R results showed that the vast majority (88%) of those with lifetime MDD sought some form of treatment for depressive symptoms. Several factors consistently

predicted the probability of initial treatment contact. Females and younger cohorts were more likely to seek treatment than males and older cohorts respectively. However, those with younger age at first onset were less likely to seek treatment than those with older age at first onset. Whilst 37% reported seeking initial treatment in the year following their first depressive episode, treatment seeking was typically delayed as the median delay was 8 years (Wang et al., 2005). Older cohort age and younger age at first onset predicted the longest delays in seeking initial treatment contact. Wang & Kessler (2005) suggested that the delays and lower treatment seeking rates associated with early age of first onset cases may have been due to poorer recognition of MDD symptoms in minors.

The results reported by Wang et al. (2005) were limited in that their analyses were unable to identify the proportions who actually received treatment. The World Health Organisation's Collaborative Study on Psychological Problems in General Health Care (CSPP, Sartorius et al., 1993) was specifically designed to investigate the detection and treatment of psychological disorders in primary care settings. The longitudinal CSPP study employed ICD-10 criteria to diagnose psychiatric disorders in a total sample of 26,422 adult patients across 15 sites worldwide. The CSPP results suggest that the identification of MDD is typically low in primary care settings as only 15% of those meeting ICD-10 criteria for major depression were correctly diagnosed. Of the remaining depressed individuals, 54% were identified as being psychiatric cases whilst 31% received no diagnosis (Lecrubier, 2007). The CSPP results also showed that patients in the youngest cohort were significantly less likely to be diagnosed with major depression than those in older cohorts. For example, only 43% of 18 to 24 year olds were correctly diagnosed with MDD compared to 59% of 25 to 44 year olds ($p < .05$; Lecrubier, 2007). The lower rate for the youngest cohort may have arisen because physicians are sometimes unwilling to diagnose a chronic mental disorder such as MDD in younger patients (Lecrubier, 2007). Finally, the CSPP results suggested that, even where correctly diagnosed, patients typically received inadequate treatment for depression from primary care physicians. However, treatment adequacy in the CSPP was assessed only in terms of whether patients received psychotropic medication (Lecrubier, 2007).

Data from the NCS-R (Kessler et al., 2003) enabled an assessment of the adequacy of both pharmacological and psychological treatments for MDD. Minimal treatment adequacy for MDD in the NCS-R was defined as receiving either, (i) 4 or more outpatient visits with a physician for pharmacological treatment over 30 days or more, (ii) 8 or more outpatient visits with any specialist provider of psychotherapy each lasting for 30 minutes or more (Kessler et al., 2003). The NCS-R results showed that 57% of 12 month MDD cases sought

help for emotional problems in the 12-months prior to interview. Of these, 90% were treated in healthcare settings and 55% of this sample were treated in specialist mental health settings (Kessler et al., 2003). The highest rate of minimally adequate treatment (64%) was seen in specialist mental health settings, where interventions were provided by psychiatrists, psychologists, counsellors or social workers (Kessler et al., 2003). The rate of minimally adequate treatment in general medical settings was 41% where treatments were provided by primary care physicians, other medical specialists, or non-specialist nurses (Kessler et al., 2003). Increasing severity according to the QUIDS-SR (Rush et al., 2003) and increasing number of comorbid DSM IV disorders both significantly predicted treatment seeking and treatment adequacy (Kessler et al., 2003). Finally, the NCS-R results revealed that of the entire sample meeting DSM IV diagnostic criteria for MDD only 21.7% received adequate treatment (Kessler et al., 2003).

2.5 Summary & Concluding Remarks

Major depressive disorder is a highly comorbid and recurrent disorder that affects twice as many females as males. Approximately 5% of adults in developed countries will meet diagnostic criteria for major depression each year and at least 10% will experience at least one episode in their lifetime. MDD is a major risk factor for suicide and a range of physical and behavioural sequelae that place a great burden on individuals and the wider economy. The burdens associated with MDD appear to be increasing in developed countries as younger cohorts demonstrate both the highest 12 month prevalence rates and most severe episodes. That the effective treatment of MDD is a pressing public health concern is highlighted by the World Health Organisation's prediction that it will form the major disease burden in developed countries by 2020. However, it appears that the majority of depressed individuals do not receive adequate treatment.

The apparently low rate of adequate treatment provision does not reflect attempts to improve the efficacy of depression treatments over recent decades. Indeed, a large body of primary research has been accumulated concerning the efficacy of specific psychological and pharmacological treatments. Recent attempts to summarise the results of numerous psychological treatment studies, and thus determine which treatments are the most effective, have relied on meta-analytic approaches. The next chapter will present an overview and critique of meta-analyses that have investigated the efficacy of psychological treatments for MDD.

Chapter Three

Meta-Analyses of Depression Studies: Overview and Critique

3.1 Introduction

Meta-analyses are increasingly used to summarise the results of primary treatment studies. However, whilst meta-analysis is capable of providing highly reliable empirical evidence concerning treatment efficacy, there are a range of methodological factors that can bias the results and in turn the conclusions drawn. To reduce the risk of bias within meta-analysis, the systematic review method has been developed. Adherence to the systematic review method ensures that potential methodological risk factors are minimised thus increasing confidence in meta-analytic results. The systematic review method is briefly described, along with an illustrative example of a major systematic review and meta-analysis of controlled trials on the efficacy of psychological treatments for depression (Churchill et al., 2001). However, the results of Churchill et al. (2001) demonstrate that the results of meta-analyses concerning the efficacy of psychological treatments are still at considerable risk of bias despite the employment of systematic review methods.

3.2 Evidence Based Movement in Healthcare

Since the 1980s, healthcare funding agencies have made increasing demands that the efficacy of treatments be demonstrated in empirical research (Niessen et al., 2000). The randomised controlled trial (RCT) is widely seen as the "Gold Standard" method to determine treatment efficacy (Pilling, 2008; Staines, 2007). The RCT is a trial in which subjects are assigned to one of two groups. One group is the experimental group and the other the control or comparison group. Both groups are assessed at pre and post-treatment and often followed up beyond the end of the treatment phase. The groups are compared to see if there are differences in outcome to determine which intervention was the most efficacious. Conducting an RCT requires the researcher to pay careful attention to the design and implementation of the trial to minimise potential confounds. By selecting patients according to pre-specified inclusion criteria and subsequent randomisation to treatment group, the RCT affords the best protection against threats to internal validity and increases confidence that interventions were responsible for any observed group differences in outcome (Howard et al., 1996; Chambless and Hollon, 1998). Despite their methodological rigour, the generalisability of RCT findings concerning psychological treatments have been criticised on the basis that included patients are not representative of those encountered in

applied clinical settings (Howard et al., 1996; Westen et al., 2004). However, where patients and treatments in research and applied settings are similar, RCT evidence will generalise to clinical settings (Wilson, 1998). In general, results of RCTs are considered to be the most reliable source of evidence concerning treatment efficacy and form a major component in the development of health policy and treatment guidelines (Oxman, 2004; NICE, 2009).

However, methodologically similar RCTs can produce conflicting findings (Chambless and Hollon, 1998). Studies may disagree on the relative efficacy of a specific treatment or whether a treatment is more efficacious than treatment as usual controls. Moreover, even where studies support the efficacy of a treatment, quantitative estimates of treatment effect can vary between studies which makes it difficult to estimate the average effect of treatment. This variability may arise for many reasons that include between-study differences in patient demographics, treatment implementation or the measures used to quantify symptomatic change. Consequently, valid methods of summarising the results of primary studies are required.

3.3 Methods for Summarising Research

Smith & Glass (1977) first used meta-analysis to summarise psychotherapy research findings in order to overcome the limitations of previous approaches. Previously, the expert literature review formed the basis by which the evidence base for treatment efficacy was synthesised. However, this method was criticised for two reasons: (i) subjective biases could lead to the exclusion of potentially important studies, (ii) the methods used to summarise the evidence lacked scientific rigour (Mullen and Ramirez, 2006). The 'vote counting' method was typically used to summarise the evidence concerning the efficacy of specific treatment approaches. In the vote counting method, the superior efficacy of treatment A over treatment B is confirmed when the majority of comparison studies find treatment A superior to treatment B (Andrews and Harvey, 1981; Mullen and Ramirez, 2006; Smith and Glass, 1977). However, the vote counting method cannot provide an estimate of the average effect of treatment across studies and actually suffers from reduced statistical power as the number of included studies is increased (Hedges and Olkin, 1980; Andrews and Harvey, 1981). These limitations led to the adoption of meta-analysis to summarise the evidence base for psychotherapy research (Andrews and Harvey, 1981; Smith and Glass, 1977). In contrast to vote counting methods, the statistical power to detect treatment differences in meta-analysis increases with the number of included studies (Mullen and Ramirez, 2006). The meta-analytic method is now briefly described, followed by a description of the history and methodological issues surrounding its use in the field of depression research.

3.3.1 Meta Analysis

Meta-analysis is a statistical method used to summarise either continuous or categorical outcome data from primary treatment studies. As currently practiced, the Centre for Reviews and Dissemination (CRD, 2009) state that it is necessary for primary studies to be methodologically similar and to compare the same types of treatment. Where primary studies include an untreated control group, the overall result of meta-analysis is an estimate of the mean efficacy of the active treatment. Where primary studies compare two active treatments, the overall results of meta-analysis is an estimate of their relative efficacy, that is, how much better on average one treatment is than the other. A fundamental assumption underlying the application of the majority of meta-analyses is that the variable treatment effect observed across primary studies are individual samples of the mean treatment effect in the clinical population of interest (Field, 2003). Such analyses are described as fixed-effects meta-analyses. An alternative approach is to assume that the results of included studies are sampled from differing clinical populations. Analyses that account for this latter assumption are used to find the mean treatment effect across populations and are termed random-effects meta-analyses (Field, 2003; CRD, 2009). However, for simplicity the following discussion will refer only to fixed-effects meta-analyses.

In order to calculate the mean effect of treatment across primary studies, it is necessary to combine their results in meta-analysis. However, individual studies often employ different outcome measures to compare treatment groups (e.g. treatment A & treatment B). Where categorical outcomes are employed in primary studies, the definition of what constitutes a clinically desirable outcome may vary. These studywise differences make pooling their results problematic. For example, for continuous outcomes it is unlikely that identical scores on different measures denote equivalent symptom severity (CRD, 2009). Consequently, the magnitude of the mean difference between treatments A and B will not be directly comparable between studies. This problem is overcome within meta-analysis by using a standardised dimensionless measure, or effect size, to compare treatment outcomes between the groups in individual studies (CRD, 2009).

The most commonly used effect size for continuous measures is the standardised mean difference (SMD, Nugent, 2006), a typical example of which is Cohen's d. This is calculated as the difference between the post-treatment mean scores between treatment groups A and B divided by their pooled pre-treatment standard deviation (Cohen, 1988). For example, where the mean of treatment type B is subtracted from the mean of treatment A, a Cohen's d of +1 indicates that the post-treatment mean for group A is one pooled standard deviation higher than that for group B. Where d = 0, there is no difference in the mean symptomatic change

between treatment groups A and B. It is generally assumed that such dimensionless SMD effect sizes are independent of the measure used to derive them (CRD, 2009; Smith and Glass, 1977). However, this assumption has recently been challenged as unjustified where there is a non-linear relationship across the range of possible scores between different outcome measures used across studies (Nugent, 2006). In such cases, failure to adjust individual effect sizes to account for the imperfect reliability of measures may lead to biased conclusions (Nugent, 2006; Nugent, 2009). Furthermore, as SMD effect sizes are dimensionless it may be difficult to interpret the clinical importance of any overall difference between treatments (CRD, 2009). Where primary studies have compared treatments using categorical outcomes, the odds (or risk) of achieving a clinically desirable outcome in each group is used to calculate an odds (or risk) ratio between treatment groups.

After effect sizes have been calculated for the treatments in each included study, they are combined to produce an overall effect size which reflects the average effect of treatment in the population (CRD, 2009). Weighting techniques are frequently used to ensure that effect sizes from smaller, and thus less precise studies, do not bias overall results (CRD, 2009). The overall effect size is then tested to determine whether its magnitude differs significantly from the predicted value according to the null hypothesis, i.e. treatments are equally effective. Where a significant overall effect size favours treatment over controls, the efficacy of treatment is supported. Where a significant overall effect size favours treatment A over treatment B, the relative efficacy of treatment A is superior to that of treatment B. In both cases, the overall effect size resulting from meta-analysis is a measure of how much better, on average, one treatment is relative to the other.

3.3.2 Meta-Analyses of Psychotherapy Studies

Smith & Glass (1977) performed an early meta-analysis to answer previous criticisms that psychotherapy was no more effective than the usual care received by patients (Eysenck, 1952). They avoided selection bias by including the results of nearly 400 studies that investigated the efficacy of individual and group based psychological treatments across a range of disorders (Andrews and Harvey, 1981). By pooling the results of studies that compared psychotherapy with untreated controls, Smith & Glass (1977) derived a significant SMD of 0.68 in favour of psychotherapy. They concluded that the average psychotherapy client was better off than 75% of those not receiving treatment. An additional analysis failed to demonstrate any difference between behavioural and non-behaviourally based therapies (Smith and Glass, 1977).

However, their methodology was criticised because Smith & Glass (1977) included studies that used non-clinical samples such as prisoners and college students. For such groups it was argued that the observed reductions in symptomatology may not be representative of actual

clinical outcomes (Andrews and Harvey, 1981). Thus, Smith & Glass's results may have overestimated the efficacy of psychological treatments in clinical patients (Andrews and Harvey, 1981). Andrews & Harvey (1981) re-analysed the dataset used by Smith & Glass but excluded non-clinical studies. Whilst their results confirmed that psychotherapy was more effective than no treatment, they were of limited utility to clinicians as they were based on studies varying widely in terms of diagnosis, therapy type and treatment duration (Andrews and Harvey, 1981). Consequently, Andrews & Harvey (1981) proposed that meta-analysis would be better suited to answering more specific questions concerning which treatments work best for specific psychological disorders. More recently, researchers have increasingly used meta-analysis to assess the efficacy of treatments for specific clinical populations.

3.3.3 Meta-Analyses of Psychotherapy Studies for Depression

Several well-known meta-analyses investigating the efficacy of psychological treatments for depression have been published. However, they have frequently reported conflicting results concerning the efficacy of specific psychological treatments for depression. Dobson (1989) reported that Cognitive Therapy (Beck et al., 1979) was superior to other psychological treatment approaches following a meta-analysis of 28 primary studies. However, Robinson et al. (1990) criticised Dobson's methodology on the basis that Dobson (1989) failed to use all available studies, had employed only one outcome measure - the Beck Depression Inventory (BDI, Beck et al., 1961), and failed to account for researcher allegiance biases (Berman et al., 1985).

In order to conduct the most comprehensive to-date assessment of psychotherapy treatment outcomes for depression, Robinson et al. (1990) included 58 studies in their meta-analysis. They found that psychotherapy in general was more effective than control conditions (wait-list, attentional control and pill-placebo) and that cognitive, behavioural and cognitive-behavioural approaches were superior to a broad category of 'general verbal therapy' which included psychodynamic, client-centred and interpersonal approaches (Robinson et al., 1990). However, once effect sizes were adjusted to control for researcher allegiance there were no significant differences between any type of psychotherapy (Robinson et al., 1990). Furthermore Robinson et al. found that studies using treatment manuals, therapist monitoring or a formal diagnosis of depression produced effect sizes that were no different from those that did not. Effect sizes were also found to be unaffected by the length or format of treatment, the type of outcome measure employed, initial symptom severity or the level of training of therapists (Robinson et al., 1990). However, in contrast to Dobson (1989), the individual study effect sizes included in meta-analysis by Robinson et al. (1990) were an average of all those reported by individual studies. This method was of questionable validity

and likely to be biased, as the mean effect size will have been based on some measures that were not correlated with symptomatic change (Nugent, 2009; Matt and Navarro, 1997).

The meta-analyses by Dobson (1989) and Robinson et al. (1990) were both at risk of producing biased results due to the inclusion of non-randomised studies (Gloaguen et al., 1998). However, the contrasting methods used to derive effect sizes from primary studies by these authors highlights that an agreed methodological approach to meta-analysis was lacking. The increasing influence of meta-analysis in healthcare research led to calls for the development and publication of recommended methodological approaches (Boissel et al., 1989). This was due to concerns that the widespread use of inappropriate methods and statistical techniques could lead to misleading conclusions within clinical research (Boissel et al., 1989). In addition, it was recognised that agreed methods were needed for the assessment of the methodological quality of studies, and, to deal with situations where the results of one or more included studies led to statistical heterogeneity (Boissel et al., 1989). Statistical heterogeneity refers to significant variation between the effect sizes of individual studies that cannot be ascribed to sampling error (CRD, 2009). Where identified, it may indicate that there are important clinical differences between patients' response to treatment across individual studies which raises the possibility that the overall results do not derive from a single population (CRD, 2009). However, statistical heterogeneity may also arise in analyses of the same clinical population due to methodological differences between included studies and suggests that the results of one or more studies are biased (CRD, 2009).

Gloaguen et al. (1998) published a meta-analysis which was "distinguished by its sophistication" (Wampold et al., 2002; p160). Gloaguen et al. included 48 randomised studies that had compared cognitive therapy (CT) with controls, antidepressant medication (ADM), behaviour therapy, or a combination of other psychotherapeutic approaches. Treatment outcomes following CT were superior to wait list/pill-placebo controls, ADM and all psychological approaches other than behaviour therapy. However, significant statistical heterogeneity was identified in the comparisons of CT with non-behavioural therapies. This suggested that outcomes for non-behavioural therapies were not equivalent between studies. According to Wampold et al. (2002), the superiority of CT to non-behavioural therapies in the presence of significant heterogeneity occurred because non-behavioural approaches in some studies were either not intended to be effective, or lacked common factors known to be therapeutic (Wampold et al., 2002). Wampold et al. (2002) referred to these as 'non-bona fide' treatments and defined as 'bona fide' those therapies that were tailored to the needs of individuals and provided in face-to-face situations by suitably qualified therapists.

Wampold et al. (2002) re-performed Gloaguen et al.'s (1998) meta-analysis but excluded studies that employed non-bona fide treatments. Wampold et al. (2002) required that bona fide therapies were either (i) previously known, (ii) described in sufficient detail to establish that psychologically 'active ingredients' were employed to modify specific psychological processes (Wampold et al., 2002). Their re-analysis confirmed Wampold et al.'s hypothesis that the superiority of CT to non-behavioural therapies would disappear following the removal of non-bona fide treatments. However, their additional hypothesis, that comparisons of CT with bona fide treatments would not be heterogeneous was not supported. Wampold et al. (2002) traced the source of statistical heterogeneity in their results to a single study that provided an extreme effect size favouring CT (McLean and Hakstian, 1979). This extreme effect size occurred because significantly fewer CT patients dropped out of treatment than in the comparison group, leading to a loss of randomisation between groups (Wampold et al., 2002). Wampold et al. (2002) justified removing the McLean & Hakstian study from their analysis on this basis and no longer found statistical heterogeneity for their results. However, this example reveals that the use of completer sample data has the potential to bias the results of individual studies, and thus meta-analysis, which has led to recommendations that only intention to treat (ITT) samples be used (CRD, 2009).

The examples presented thus far show that meta-analysis has increasingly been used to answer specific clinical questions. However, they also show that the results of meta-analyses may be unreliable due to biases arising from methodological factors such as the internal validity, definition of treatment type and the choice of outcome measures used in primary studies as well as the type of samples analysed. Because policy makers and practitioners demand the best available evidence concerning treatment efficacy, the systematic review methodology has been developed and disseminated in order to enable researchers to reliably identify, evaluate and summarise the results from all relevant studies that answer a specific review question (CRD, 2009; Higgins and Green, 2006). Indeed, whilst the U.K.'s National Institute for Health and Clinical Excellence (NICE) uses a range of methods to determine the efficacy of treatments, systematic reviews including meta-analysis of RCT data have been described as standing at the apex of the evidence hierarchy (Goldberg, 2006).

3.3.4 The Systematic Review Method

Systematic review refers to a method that attempts to identify all relevant empirical evidence in order to answer a specific research question (Liberati et al., 2009). By following explicit, pre-defined and reproducible methods, the systematic review has the potential to provide unbiased conclusions concerning treatment efficacy when combined with meta-analysis (Mullen and Ramirez, 2006; CRD, 2009). The key elements of systematic review that reduce

the risk of bias are the clear specification of the review's objectives and methods, a comprehensive search for eligible studies, the use of independent reviewers to assess the validity of included evidence and the systematic presentation of results (Liberati et al., 2009; CRD, 2009). The specification of a review's objectives includes the clinical population of interest, the interventions under comparison, the outcomes of interest and the design of studies to be included (CRD, 2009). Ideally, the search process should include both published and unpublished study evidence, as the inclusion of unpublished results can lead to significantly different conclusions to those based on published results alone (CRD, 2009; Pilling, 2008). Following search, reviewers independently determine the eligibility of potential studies according to pre-specified inclusion criteria. Where reviewers disagree concerning study eligibility, the methods used to resolve the issue of eligibility should be transparent. The use of independent reviewers reduces the risk that inappropriate studies are included or that appropriate studies are excluded. Following inclusion, the independent extraction of data from individual studies enables an unbiased assessment of whether their results are valid and thus suitable for inclusion in subsequent meta-analysis (CRD, 2009).

In comparison to earlier approaches, conducting a systematic review prior to meta-analysis has the potential to reduce the risk that estimates of treatment efficacy are biased. This is in part due to the use of pre-specified aims and methods which reduce the chance that individual subjectivity will influence the selection of studies for inclusion. Of equal importance is that adherence to published recommendations concerning the statistical process of meta-analysis will help identify results that are methodologically robust. However, adherence to systematic methods is no guarantee that meta-analysis can provide unbiased evidence concerning the absolute or relative efficacy of psychological treatments. This will be illustrated by presenting the post-treatment findings for Churchill et al.'s (2001) systematic review and meta-analysis that sought to determine the efficacy of brief psychological treatments in depressed adults.

3.3.5 Churchill et al.'s Systematic Review & Meta-Analysis

Churchill et al. (2001) conducted a systematic review that included a series of meta-analyses of results from controlled studies investigating the efficacy of psychological treatments in depressed adults. They included only studies where an explicit psychological model was employed in treatments lasting 20 sessions or less. Depression was operationalised in studies as an elevated symptom score on validated symptom measures (e.g. the BDI), an elevated symptom score in addition to clinical interview, or following diagnosis according to standardised diagnostic criteria (e.g. DSM IV). Where included studies operationalised depression as an elevated score on symptom measures there was considerable between-study variation in the minimum criterion score required for entry. For example, the minimum

criterion BDI score in four studies was less than 10 points, whilst six required that patients score between 10 and 17 points. According to Beck et al. (1988), scores of less than 10 represent minimal or no depression, whilst those between 10 and 18 represent mild to moderate depression. However, Beck et al.'s score criteria for the BDI were based on the symptoms observed in patients formally diagnosed with an affective disorder (Beck et al., 1988). Consequently, where studies used only elevated symptom scores to identify depression, it is inevitable that a substantial proportion would not have met a formal diagnosis of MDD (Coyne, 1994).

Sixty three studies were included in Churchill et al.'s meta-analysis following a search of both published and unpublished work. They tested four main hypotheses: (i) all variants of psychotherapy were superior to treatment as usual or wait-list controls (controls), (ii) cognitive and behavioural approaches (CBT) were superior to Interpersonal, Psychodynamic or Supportive therapies combined (non-CBT), (iii) individual therapy was superior to group therapy, (iv) CBT was superior to controls. The hypotheses were tested separately in ITT analyses for both continuous and categorical post-treatment outcomes (symptom severity and remission[6] respectively). In addition, Churchill et al. conducted post-hoc sensitivity analyses to determine whether the results were robust to changes in several study-level factors that they considered important *a priori*. Firstly, sensitivity analyses were conducted concerning the methodological quality of studies as assessed by the Quality Rating Scale (QRS, Moncrieff et al., 2001). The QRS enabled included studies to be rated on the descriptive adequacy concerning patient samples and the treatments employed, in addition to enabling an assessment of their internal validity. In brief, randomised studies employing blinded assessment on valid measures and comparing treatments via intention to treat analyses were deemed to possess high internal validity. Sensitivity analyses were also conducted to examine the effects of the mean depressive severity of patients in studies, the number of available therapy sessions in studies and the recruitment source of patients (i.e. whether patients came from clinical settings or were volunteers/responders to advertisements).

The results of Churchill et al.'s (2001) meta-analyses supported all four of their main hypotheses. However, their sub-analysis concerning study quality also revealed that the inclusion of low quality studies in meta-analysis had an unpredictable effect on conclusions. Whilst overall results based on all studies showed that psychotherapy in general was superior to controls, the mean symptomatic reduction between treated patients and controls

[6] Churchill et al. (2001) used the term recovery to denote the absence of clinically significant levels of depression at post-treatment. Remission is the more appropriate term according to ACNP recommendations (Rush et al., 2006).

demonstrated only a 'borderline difference' when analysis was restricted to the poorest quality studies (Churchill et al., 2001). Moreover, remission rates in poorer quality studies were the lowest in comparison to higher quality studies and the superiority of psychotherapy to controls became more pronounced as study quality increased (Churchill et al., 2001). In contrast, their overall finding that CBT was superior to controls revealed an opposite trend. Here, lower quality studies contributed higher remission rates and estimates of symptomatic reduction than those of higher quality (Churchill et al., 2001). This latter trend was evident in their comparison of remission rates for CBT versus non-CBT approaches, as the superiority of CBT to non-CBT approaches was lost in analyses that excluded poor quality studies (Churchill et al., 2001).

Churchill et al. (2001) concluded that the overall findings for their four main hypotheses were likely to be biased because the majority of studies demonstrated both low quality and low internal validity. In particular, the generally poor reporting of randomisation methods, the frequent use of antidepressants in psychotherapy arms, and potential researcher biases leading to the poor implementation of non-CBT approaches led them to conclude that the potential influence of bias on their results 'cannot be under-estimated' (Churchill et al., 2001, p94). However, based on the results of the remaining sub-analyses, they cautiously suggested that several study characteristics served to alter estimates of treatment efficacy. Studies that had included higher severity patients, clinical samples, or patients with a formal diagnosis of depression had consistently contributed lower estimates of treatment efficacy than those that had not. In addition their analysis suggested that overall treatment outcomes improved as the number of available therapy sessions increased (Churchill et al., 2001).

3.4 Summary & Concluding Remarks

Increasing demands for empirical evidence concerning the efficacy of psychological treatments have seen an increase in both the sophistication and employment of meta-analytic methods. In recent decades, meta-analysis has been increasingly focused on determining the efficacy of specific psychological treatment approaches for specific disorders. However, whilst the introduction and widespread adoption of the systematic review method has reduced the likelihood that meta-analytic results are unreliable due to the employment of sub-optimal methods, the risk of bias remains considerable. Churchill et al.'s (2001) results revealed that several methodological factors continue to undermine confidence that meta-analyses of psychological treatment studies are reliable. Foremost amongst these is the failure to either properly implement or report the methods by which patients are randomised to treatment condition in primary studies. In addition to the risk of bias posed by the

inclusion of non-randomised studies in meta-analysis, Churchill et al. (2001) showed that the inclusion of studies which included undiagnosed patients led to higher estimates of treatment efficacy than studies employing formal diagnostic procedures. Consequently, the inclusion of such studies in meta-analysis may serve to increase estimates of treatment efficacy thus reducing the generalisability of results as a substantial minority of included patients may not actually be depressed.

Given these caveats and the current importance of systematic review including meta-analysis to healthcare research, the first study of this thesis is itself a systematic review. Study 1 is a systematic review of published meta-analytic reviews that have investigated the efficacy of psychological treatments for adult MDD. However, to overcome the methodological difficulties identified by Churchill et al. (2001), all included studies in reviews were required to be RCTs where patients met formal diagnostic criteria for MDD. In addition, all treatments were required to be provided individually in order to increase the specificity of results. This approach allowed an assessment of what the best meta-analytic evidence tells us about the efficacy of psychological treatments for major depression.

Chapter Four

Study 1

A Systematic Review of Meta-Analyses Investigating Psychological Treatments for Major Depression

4.1 Introduction

The previous chapter presented an overview and critique of meta-analysis when used to synthesise the results of psychological treatment studies for MDD. Whilst some of the earlier methodological difficulties associated with this approach have been overcome by the adoption of the systematic review method prior to meta-analysis, there are still methodological factors which may undermine the reliability of review findings. The systematic review by Churchill et al. (2001) served as an example which revealed that the inclusion of non-randomised studies poses a major risk to the validity of meta-analyses within systematic reviews. Also, Churchill et al.'s (2001) results indicated that the inclusion of undiagnosed individuals may threaten the generalisability of results across clinical populations. Given that the results of meta-analysis are now a key element in the creation of clinical guidance, it is of vital importance to assess the quality of potentially influential reviews that have investigated treatments for depression.

The purpose of the present systematic review was to identify and examine published reviews where the efficacy of psychological treatments for MDD has been investigated via meta-analysis. Reviews were considered methodologically rigorous where included studies were randomised controlled trials (RCTs) that had investigated the efficacy of individual psychological treatments in patients who met standardised diagnostic criteria for MDD. In addition, psychological treatments were required to have been based on theoretical models of psychopathology. Consequently, approaches such as non-directive counselling were excluded. The identification of such meta-analytic reviews enabled an assessment of how effective current psychological treatments are for MDD. Furthermore, assessments of the methodological quality of included reviews enabled a determination of whether they were at risk of providing biased results both individually and across all reviews.

4.2 Method

4.2.1 Search Strategy

Meta-analytic reviews were searched for using the following databases: Cochrane DARE, Cochrane Database of Controlled Trials, Cochrane Database of Economic Evaluations, Cochrane Database of Systematic Reviews, Cochrane Database of Technology Assessments, PsychINFO (1967-2009), EMBASE (1980-2009), Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations (6/03/2009), Ovid MEDLINE(R) (1950-2009), SCOPUS (12/03/2009) and Web of Knowledge (SCI-EXPANDED 1945 to 11/03/2009; SSCI 1956 to 11/03/2009). An experienced reviewer (YD) designed the search filter template. Search terms included: depression, depressive, major depressive disorder, depressive disorder, dysthymia, psychotherapy meta-analysis, systematic review and possible variants for each database. The full search strategy is illustrated appendix A. References of the included reviews were searched to identify any additional eligible articles. No attempt was made to find eligible 'in press' articles. All databases were searched in the first two weeks of March 2009.

4.2.2 Eligibility Criteria

The eligibility criteria for reviews are shown in Table 2. No limitations were imposed on publication date, length of treatment or follow-up, outcome measures used to determine depressive symptom severity or numerical methods used for meta-analysis.

4.2.3 Selection of Meta-analytic Reviews

All identified articles were combined in a single database and duplicates removed. A single reviewer (MC) then excluded references which clearly did not meet the eligibility criteria based on title alone. If there was any doubt concerning an article's eligibility at this stage of the selection process the reference was retained and included in the abstract screening stage.

Two independent reviewers (MC/RD) then screened the abstracts of remaining articles to exclude ineligible articles. Articles which could not be excluded on the basis of title or abstract were obtained and assessed by two independent reviewers (MC/CH) using a purpose made screening tool (see appendix B). This tool operationalised the eligibility criteria and permitted each eligibility criterion to be recorded as present, absent or questionable for each meta-analytic review. The tool also included space to make comments, where it was not possible to fully assess a criterion from the manuscript alone, or where a minority of studies included in the review could result in the review's exclusion. Where reviewers failed to agree concerning the eligibility of reviews a third reviewer was consulted (PF).

Table 2. Eligibility Criteria for Included Meta-analytic Reviews

	Inclusion Criteria	Exclusion Criteria
Review Type	Systematic review, meta-analysis or review of randomised controlled trials Published in peer reviewed source	No English translation available Unpublished
Patient Type	Adults diagnosed with major depression according to a classificatory diagnostic scheme	Diagnosis based solely on screening instruments or where consistent methods were not described when using published criteria Depression treated specifically in the context of substance abuse, personality, psychotic or medical disorders Sub syndromal depression not meeting criteria for major depression
Treatment	Individual psychotherapy in at least one treatment condition based on theoretical model of psychopathology	Group therapy Computer administered therapy Self help interventions
Comparison Conditions	Treatment as usual, waiting list control, attentional control, psychotherapy, pharmacotherapy	
Outcomes	Meta-analytic estimates of therapeutic efficacy based on group-level data	Narrative review Reviews not using group as the unit of analysis

4.2.4 Data Extraction

Substantive data from included meta-analytic reviews were extracted using a modified version of the University of York's Centre for Reviews and Dissemination's abstract reporting format (CRD, 2009a). One reviewer (MC) extracted data which was checked and revised if necessary by another (CH) concerning:

- the authors' objectives
- search methods and included designs
- patient type, severity and duration of depression
- comparison conditions
- therapy types and diagnostic techniques used in studies
- setting and duration of therapy
- assessment points in time
- results from meta-analysis
- findings of any heterogeneity of results between included studies and whether they were accounted for
- authors conclusions & statements concerning implications for practice or research

The risk that any single review provided biased meta-analytic results was assessed using an instrument based on the University of Sheffield's School of Health & Related Research (ScHARR) Systematic Review Quality Appraisal guidance (University of Sheffield, 2009). Risk of bias data (formerly called, 'quality data' Liberati et al., 2009) were extracted by two independent reviewers (MC/YD). The quality of reviews was assessed in terms of the following questions:

- was the search process adequate?
- were eligibility criteria reported?
- included studies valid?
- was there an assessment of study quality?
- appropriate outcome measures used?
- methods of data extraction reported?
- appropriateness of any numerical synthesis and any sub-group analyses?
- presented numerical results appropriate?
- issues of generalisability addressed?

Answers to these questions were independently categorised as yes/no or partially with disagreements being resolved by discussion. Appendix C contains the composite instrument used to extract both types of data.

The risk that the results of reviews were biased was assessed in terms of within and across review risk of bias. Within review risk refers to the risk that an individual review provided biased results by virtue of its methodology. In contrast, across review bias refers to the risk that all reviews provided biased results as a consequence of shared methodological shortcomings. This approach to review bias was analogous to current recommendations for the reporting of risk of bias in single systematic reviews and meta-analyses of primary studies (Liberati et al., 2009). Within review risk of bias was assessed using the quality appraisal instrument for systematic reviews described above. Across review risk of bias was assessed within an iterative analysis of substantive and quality data. Important methodological issues that potentially affected all reviews were raised by some individual review authors. Where individual reviews provided insufficient information to assess across review risk of bias, the manuscripts of their included studies were consulted where possible.

4.3 Results

4.3.1 Review Selection & Objectives

Figure 1 presents an overview of the selection procedure. The independent assessment of substantive data from 107 full manuscripts identified 12 potentially eligible articles with a reviewer agreement rate of 94% (101 of 107 articles). It was found that some meta-analytic reviews contained results which were based on studies meeting our eligibility criteria whilst other results within the same review were not. In this case only the eligible results were extracted. Ten meta-analytic reviews were borderline cases for inclusion, five of which were included following further investigation and referral to the third reviewer PF (Casacalenda et al., 2002; Friedman et al., 2004; Leichsenring, 2001; Parker et al., 2008; Vittengl et al., 2007). The reasons for their inclusion are presented in appendix E. A search of the references of the seven included meta-analytic reviews failed to identify any further eligible reviews.

The objectives of included meta-analytic reviews are presented in Table 3. All compared the efficacy of psychotherapy with an alternative condition at post-treatment, with four providing results for follow-up outcomes (de Maat et al., 2006; Friedman et al., 2004; Leichsenring, 2001; Vittengl et al., 2007). Acute phase psychotherapy was compared in six reviews, whilst Vittengl et al. (2007) included comparisons of continuation phase psychotherapy (C-CT) with an alternative condition.

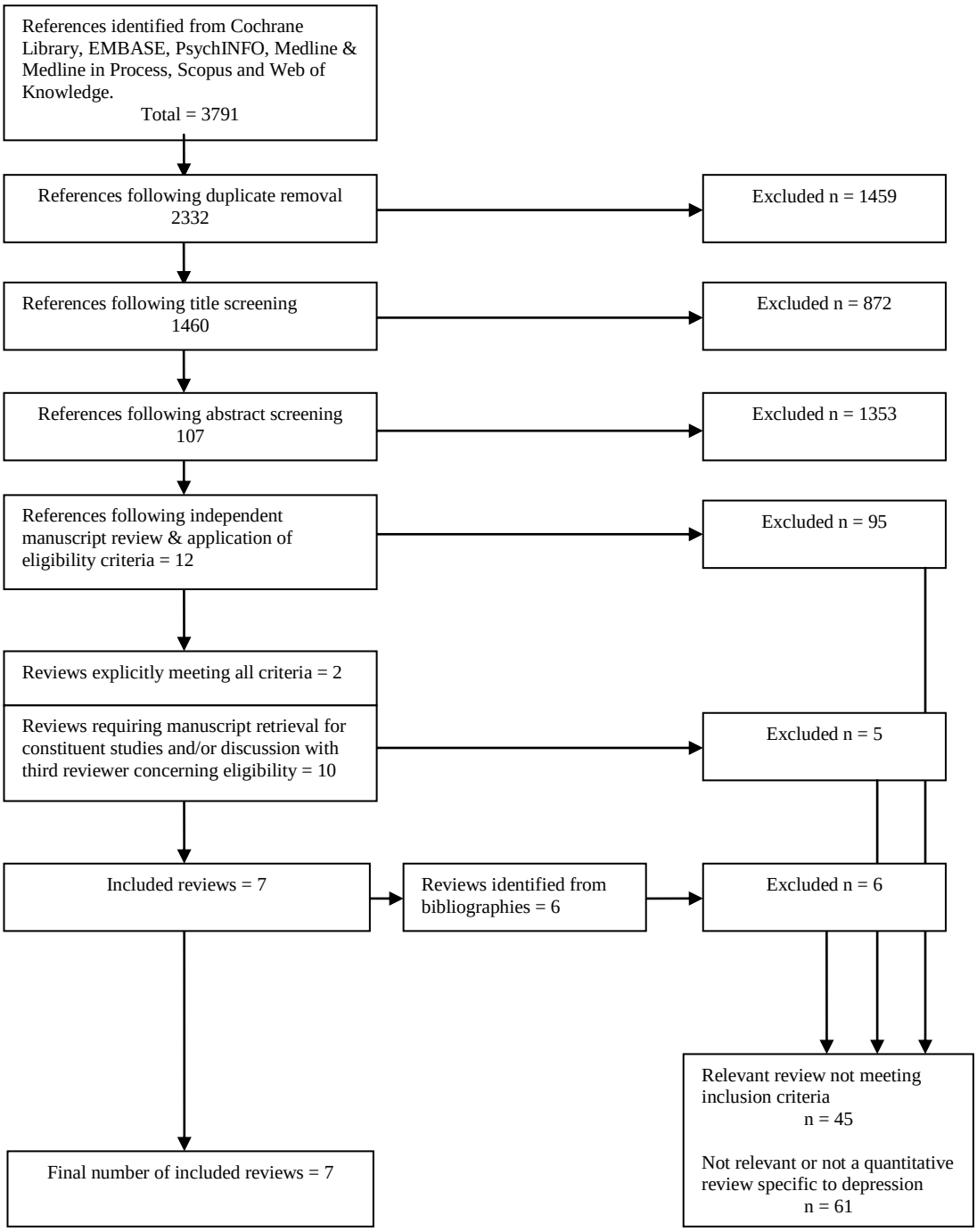


Figure 1. Selection of Eligible Meta-analytic Reviews

Table 3. Objectives of Included Meta-analytic Reviews.

Review	Objectives of the Review
Casacalenda et al. (2002)	To determine the percentages of patients achieving remission from depressive symptoms within randomised controlled trials that had directly compared psychotherapy, pharmacotherapy and control conditions.
de Maat et al. (2006)	To determine the relative efficacy of pharmacotherapy & psychotherapy assessed at treatment termination and at follow-up for clinically homogeneous patients. A secondary objective was to investigate the impact of dropout and the severity or chronicity of depression on outcomes.
de Maat et al. (2007)	To determine the relative efficacy of psychotherapy & psychotherapy combined with pharmacotherapy in the acute treatment of depression for clinically homogeneous patients. A secondary objective was to investigate possible differences in dropout rates between conditions and whether differences existed in prognosis for patients suffering differing severities or durations of depression.
Friedman et al. (2004)	To determine whether combined therapy was more efficacious than pharmacotherapy or psychotherapy alone as part of a broader review of the literature on combined treatment for major depressive disorder. Only studies which compared the same type of psychotherapy across treatment conditions were included.
Leichsenring (2001)	To directly compare the efficacies of Short Term Psychodynamic Therapy and CBT at post treatment and at follow-up for treatments lasting at least 13 sessions. The analysis used a range of psychosocial and psychiatric measures in addition to depression symptom measures.
Parker et al. (2008)	To challenge the findings of a previous meta-analysis by Gloaguen et.al (1998) where it was concluded that CBT for depression demonstrated superior post-treatment outcomes than pharmacotherapy. Only a sub-set of the original studies using the BDI and meeting stricter inclusion criteria than in the original meta analysis were used.
Vittengl et al. (2007)	To determine (i) the rate of relapse-recurrence amongst responders to acute phase cognitive therapy (A-CT) and (ii) whether A-CT reduced relapse-recurrence better than other acute phase treatments during follow-up. To determine whether continuation phase cognitive therapy (C-CT) reduced relapse-recurrence more than non-active control conditions or other active continuation phase treatments at the end of C-CT and during follow-up.

Table 4. Characteristics of Meta-analytic Reviews: Post-treatment Comparisons.

Review	Primary Comparison	Therapy Models	Setting	Diagnostic Criteria	Pre-treatment Severity		N_s	N_p	Maximum Therapy Sessions	Treatment Duration (weeks)	Primary Outcome
Casacalenda et al. (2002)	Psychotherapy vs ADM or controls	CBT IPT PST SWC	Outpatient	DSMIII DSMIII-R RDC	Min BD Min HRSD	14 to 20 13 to 14	6	883	6 to 20	10 to 34	Patients remitted
de Maat al. (2006)	Psychotherapy vs ADM	CT/CBT CBASP IPT	Outpatient	DSMIII-R DSMIV RDC	Min BDI Min HRSD	14 to 20 10 to 20	10	1233	16 to 24	8 to 20	Patients remitted
de Maat et al. (2007)	Psychotherapy vs combined therapy	CT/CBT CBASP STPP	Outpatient	DSMIII-R DSMIV RDC	Min BDI Min HRSD	14 to 20 12 to 20	7	903	16 to 24	8 to 20	Patients remitted
Friedman et al. (2004)	Psychotherapy vs combined therapy	CBASP PST CT	Outpatient	RDC Feighner DSMIV	Min HRSD	13 to 20	3	530	6 to 20	12	Effect sizes for: Symptom reduction Recovery
Leichsenring (2001)	CBT vs STPP	CBT STPP	Outpatient	RDC DSMIII Feighner	Min BDI Min HRSD	10 to 17 14	5	323	13 to 20	n/a	Patients remitted or improved
Parker et al. (2008)	Cognitive Therapy vs ADM	CBT	Outpatient/ Inpatient	RDC Feighner DSMIII-R	Min BDI Min HRSD	14 to 23 10 to 20	9	327	12 to 24	8 to 15	Effect sizes for: symptom reduction, patients responding
Vittengl et al. (2007a)	C-CT vs non-active controls	CT	Outpatient	DSMIII-R DSMIV	Min HRSD	12 to 16	4	234	6 to 10	35 to 52	Relapse/recurrence assessed at the end of C-CT

Key: ADM = antidepressant medication; BDI = Beck Depression Inventory; CBASP = cognitive behavioural analysis system of psychotherapy; CBT = cognitive behavioural therapy; C-CT = continuation phase cognitive therapy; CT = cognitive therapy; DSM = Diagnostic & Statistical Manual of Mental Disorders; Feighner = Feighner Diagnostic Criteria for Use in Psychiatric Research; HRSD = Hamilton Rating Scale for Depression; ; ITT = Intention to treat analysis; IPT = interpersonal psychotherapy; Min = Minimum; n/a = not available; N_p = total number of patients included in post-treatment analysis; N_s = maximum number of studies used in any post-treatment meta analysis; PST = problem solving therapy; RDC = research diagnostic criteria; STPP = short term psychodynamic psychotherapy; SWC = social work counselling.

Table 5. Characteristics of Meta-analytic Reviews: Follow-up Comparisons.

Review	Primary Comparison	Therapy Models	Diagnostic Criteria	Pre-treatment Severity at Start of Acute Treatment		N_s	N_p	Maximum Therapy Sessions	Treatment Duration (weeks)	Follow-up Period (weeks)	Primary Outcome
de Maat al. (2006)	Psychotherapy vs ADM	CT/CBT IPT	DSMIII-R DSMIV RDC	HRSD mild to moderate		6	231	20 to 24	8 to 20	52 to 104	Relapse
Friedman et al. (2004)	Psychotherapy vs combined therapy	CT	RDC DSM III	Min BDI Min HRSD	10 to 17 14	3	78	20 to 23	12 to 20	52 to 104	Relapse
Leichsenring (2001)	CBT vs STPP	CBT STPP	RDC DSMIII	Min BDI Min HRSD	10 to 17 14	4	270	16 to 20	n/a	52 to 104	Patients remitted or improved
Vittengl et al. (2007b)	C-CT vs non-active controls	CT	RDC DSMIII-R DSMIV	Min HRSD	12 to 16	5	232	10	20 to 35	69 to 312	Relapse/recurrence
Vittengl et al. (2007c)	A-CT vs Other depression specific psychotherapies	CT	RDC DSMIII DSMIII-R	Min BDI Min HRSD	20 14	4	194	8 to 20	16	52 to 104	Relapse/recurrence
Vittengl et al. (2007d)	A-CT vs ADM	CT	RDC DSMIII DSMIII-R DSMIV DSMIV-TR	Min BDI Min HRSD	20 to 21 12 to 21	7	344	20 to 24	8 to 16	52to 104	Relapse/recurrence
Vittengl et al. (2007e)	A-CT vs Combined therapy	CT	DSMIII DSMIV	Min BDI Min HRSD	20 to 21 14 to 21	3	136	20 to 24	8 to 12	52 to 104	Relapse/recurrence

Key: ADM = antidepressant medication; A-CT = acute phase cognitive therapy; BDI = Beck Depression Inventory; CBT = cognitive behavioural therapy; C-CT = continuation phase cognitive therapy; CT = cognitive therapy; DSM = Diagnostic & Statistical Manual of Mental Disorders; HRSD = Hamilton Rating Scale for Depression; IPT = interpersonal psychotherapy; Min = Minimum; n/a = not available; N_p = total number of patients included in follow-up analysis N_s = number of studies used in follow-up meta analysis; RDC = research diagnostic criteria; STPP = short term psychodynamic psychotherapy.

Note: all session and duration data refer to acute treatment studies providing data for follow-up comparisons except for Vittengl et al. (2007b) which refers to continuation treatments

4.3.2 Characteristics of Meta-analytic Reviews

In order to aid clarity Table 4 and Table 5 present the characteristics of included meta-analytic reviews for post-treatment and follow-up comparisons respectively. There are four references to Vittengl et al. (2007) in Table 5 as Vittengl et al. provided more than one type of treatment comparison at follow-up.

Treatment Comparisons

Post-treatment

Table 4 shows that three reviews compared psychotherapy with anti-depressant medication (ADM; Casacalenda et al., 2002; de Maat et al., 2006; Parker et al., 2008), two with psychotherapy plus ADM (combined therapy; de Maat et al., 2007; Friedman et al., 2004), and two with controls (Casacalenda et al., 2002; Vittengl et al., 2007). Specific psychotherapy models were compared with alternative treatments in three reviews (Leichsenring, 2001; Parker et al., 2008; Vittengl et al., 2007), whereas four reviews pooled psychotherapy models for their comparisons with alternative treatments (Casacalenda et al., 2002; de Maat et al., 2006; de Maat et al., 2007; Friedman et al., 2004). The post-treatment comparison for Vittengl et al. (2007a) in Table 4 refers to outcomes at the end of C-CT in patients who responded[7] to acute phase psychological treatments and were subsequently assigned to either a C-CT or untreated control group. Table 6 provides references to the studies used for post-treatment analyses in reviews.

Follow-up

Table 5 shows that four reviews made the same treatment comparisons at follow-up as seen at post-treatment in Table 4 (de Maat et al., 2006; Friedman et al., 2004; Leichsenring, 2001; Vittengl et al., 2007b). De Maat et al. (2006) and Leichsenring (2001) based their follow-up comparisons only on follow-up data that was available for the patients included in their post-treatment comparisons. However, both Friedman et al. (2004) and Vittengl (2007b) included follow-up data from studies that were not included in their post-treatment analyses. Vittengl et al. (2007) provided three additional comparisons for which there were no corresponding post-treatment results in Table 4. These compared A-CT with ADM (Vittengl et al., 2007d), combined therapy (Vittengl et al., 2007e), or other depression specific psychotherapies (Vittengl et al., 2007c). This was because the primary focus of Vittengl et al.'s review was to compare the efficacy of treatments in the prevention of relapse/recurrence - not the efficacy of acute treatments. The studies included for follow-up comparisons in reviews are referenced in Table 7.

[7] The criteria for response employed by original authors are not presented in this review.

Table 6 Studies Used for Post-treatment Comparisons in Reviews

	Rush, 1977	Hercegbaron, 1979	McLean, 1979	Blackburn, 1981	Hersen, 1984	Murphy, 1984	Beck, 1985	Roter-Zimmer, 1985	Thompson, 1987	Elkin, 1989	Hollon, 1992	Scott, 1992	Gallagher-Thompson, 1994	Shapiro, 1994	Murphy, 1995	Mynors-Wallis, 1995	Hautzinger, 1996	Schulberg, 1996	Blackburn, 1997	Jarrett, 1998	Jarrett, 1999	Jarrett, 2000	Keller, 2000	Mynors-Wallis, 2000	Jarrett, 2001	de Jonghe, 2004	Klein, 2004	DeRubeis, 2005
Casacalenda 2002		X								X		X				X		X			X							
de Maat 2006				X		X				X	X				X		X		X		X		X					X
de Maat 2007				X		X	X				X						X						X			X		
Friedman 2004						X					X												X	X				
Leichsenring 2001					X				X	X			X	X														
Parker 2008	X		X	X		X		X		X	X				X		X											
Vittengl 2007 (a)																				X		X			X		X	

For Vittengl et al (2007): a = C-CT vs controls

Table 7. Studies Used for Follow-up Comparisons in Reviews

	Blackburn, 1986	Simons, 1986	Gallagher-Thompson 1990	Shea, 1992	Evans, 1992	Jacobson, 1993	Gallagher-Thompson, 1994	Shapiro, 1995	Hautzinger, 1996	Fava, 1998	Gortner, 1998	Jarrett, 1998	Jarrett, 2000	Jarrett, 2001	Fava, 2004	Hollon, 2005	Segal, 2006
de Maat 2006	X	X		X	X				X							X	
Friedman 2004	X	X			X												
Leichsenring 2001			X	X			X	X									
Vittengl 2007 (b)										X		X	X	X	X		
Vittengl 2007 (c)				X		X		X			X						
Vittengl 2007 (d)		X		X	X				X				X			X	X
Vittengl 2007 (e)		X			X				X								

For Vittengl et al (2007):

b = C-CT vs non active controls
c = A-CT vs other depression specific psychotherapies
d = A-CT vs ADM
e = A-CT vs combined therapy

Diagnosis & Patient Samples

Table 4 and Table 5 both show that patients contributing to all meta-analyses in reviews were diagnosed with MDD using a variety of clinician rated diagnostic criteria. Post-treatment comparisons in six reviews were based on outpatients, whilst Parker et al. (2008) included a single study (Hautzinger, 1996) that had included inpatient data (Table 4). However, this study was used in only in one of four comparisons made by Parker et al. (2008) where inpatients contributed approximately 16% to both the sample of 181 CBT and 166 ADM patients. De Maat et al. (2006, 2007) also included Hautzinger et al. (1996) but extracted data for outpatients only.

Mean pre-treatment severity of patients in included studies

An examination of the included studies in reviews (Table 6 & Table 7) enabled the mean pre-treatment severity of patient samples used in meta-analysis to be compared between some reviews. De Maat et al. (2006, 2007) described the mean pre-treatment severity for each of their included studies. Because included studies used differing versions of the HRSD, de Maat et al. (2006, 2007) used a published algorithm to convert scores from differing versions to correspond with those of the 17 item HRSD. De Maat et al. (2006, 2007) concluded that the mean pre-treatment severity of patients across their studies fell within the mild to moderately depressed range (12 to 19.9 and 20 to 24.9 points respectively). Casacalenda et al. (2002) also reported mean HRSD scores ranging from 15.3 to 23.4 indicating mean severities in the mild to moderate range. However, Casacalenda et al. (2002) did not report which versions of the HRSD were used, and a mean score of 23.4 for Schulberg et al. (1996) suggested that approximately 50% of patients in this study may have been severely depressed according to de Maat et al.'s (2006, 2007) criteria.

Table 6 indicates that the majority of post-treatment studies in Friedman et al. (2004) and Parker et al, (2008) were included in de Maat et al. (2006). This suggested that the mean pre treatment severity of patients in these studies fell within the mild to moderately depressed range according to de Maat et al.'s (2006, 2007) criteria. Similarly, Table 7 indicates that all of the follow-up studies included by Friedman et al. (2004) and Vittengl et al. (2007e) were included in de Maat et al.'s (2006) follow-up analysis, again suggesting that their mean pre-treatment severity fell within the mild to moderate range. Finally, five out of seven studies in Vittengl et al. (2007d) were included by de Maat et al. (2006), suggesting that the majority of studies in Vittengl et al.'s (2007d) comparison had mean pre-treatment severities in the mild to moderate range.

Number of patients in meta-analytic comparisons

Table 4 and Table 5 show that the highest numbers of patients were included in meta-analytic comparisons of psychotherapy with ADM at both post-treatment (de Maat et al., 2006) and follow-up (Vittengl et al., 2007d). The lowest number of patients in post-treatment comparisons were found for comparisons of specific psychotherapy models (Table 4). However, this was not the case at follow-up where Vittengl et al. (2007d) incorporated the highest number of patients in a comparison of A-CT with ADM (Table 5).

Treatment Sessions & Overall Duration of Therapy

Post-treatment Analyses (Table 4).

Review meta-analyses pooled studies that varied considerably in terms of the maximum number of psychotherapy sessions available to patients. For example, Casacalenda et al. (2002) pooled studies where treatment ranged from 6 to 20 sessions. Acute treatment comparisons showed a wider range of maximum available treatment sessions than seen for continuation treatment. Original articles for the included studies in reviews were consulted which indicated that the median number of maximum available sessions for acute treatments was 20 across reviews. For continuation treatments the median number of maximum available sessions was 10.

The time period over which acute treatment sessions were provided also showed considerable variability except for Friedman et al. (2004) where all were scheduled for 12 weeks. The longest acute treatment duration of 34 weeks in Casacalenda et al. (2002) was due to the inclusion of Schulberg et al. (1996) where 16 weekly acute sessions were followed by four monthly continuation sessions. The highest treatment durations contributing to post-treatment analyses were seen for continuation treatments (Vittengl et al., 2007a) as continuation phase sessions were provided less frequently than for acute treatments (Table 5).

Follow-up Analyses (Table 5).

With the exception of Vittengl et al. (2007c), follow-up comparisons were based on studies that demonstrated a smaller range of available sessions than those used in post-treatment comparisons. The median values for available sessions and the duration of acute treatments were 20 and 16 weeks respectively. For C-CT the corresponding values were 10 sessions and 27 weeks (Vittengl et al., 2007b).

Definitions of Outcome

All meta-analytic reviews compared treatments in terms of categorical outcomes, the most common being remission at post-treatment and relapse[8] at follow-up (Table 8 & Table 9). Only Friedman et al. (2004) and Parker et al. (2008) presented effect sizes based on continuous measures i.e. symptom reduction (Table 4). The vast majority of categorical outcomes used in post-treatment comparisons were derived from studies that had effectively assessed remission in terms of a minimum severity score on either the Hamilton Rating Scale for Depression (HRSD, Hamilton, 1960) or Beck Depression Inventory (BDI, Beck et al., 1961). For example, most of Parker et al.'s (2008) definitions of 'response' were operationalised as a BDI score of less than 10 (Table 8). According to Beck et al. (1988) scores below 10 represent minimal or no depression in patients previously diagnosed with an affective disorder which indicates that the majority of outcomes in Parker et al. (2008) were estimates of remission. Similarly, the HRSD criteria used to define post-treatment recovery according to Friedman et al. (2004) were typically the same as those that de Maat et al. used to define remission (Table 8). The majority of categorical outcomes for follow-up comparisons were based on the identification of relapse/recurrence following a new MDE or retreatment for depression (Table 9). However, Table 8 and Table 9 reveal that the treatment comparisons in all reviews were based on diverse outcome definitions. For example, Table 8 shows that de Maat et al. (2006) included studies which defined remission as a criterion score of 6 or less on the HRSD whilst others used 7, 8 or 9. In addition, included studies could operationalise outcomes based on more than one criterion. For example, relapse was defined in one of Friedman et al.'s (2004) studies as a BDI greater than or equal to 16, or retreatment for depression (Table 9).

[8] This is better described as recurrence, however, the term will be retained to correspond with the definitions used in primary studies.

Table 8. Definitions of Post- treatment Outcome Used in Review Studies

Comparison	Review	Outcome Criteria	Definitions of Outcome
Controls	Casacalenda et al. (2002)	Remission - ITT	HRSD ≤ 6 or 7 Raskin Depression Scale ≤ 5
		Remission - Completer	HRSD ≤ 6 or 7
	Vittengl et al. (2007 a)	Relapse/recurrence	MDE or retreatment for depression. MDD and HRSD ≥ 16 for ≥ 2 visits
Psychotherapies	Leichsenring (2001)	Remission or improvement	HRSD ≤ 6 BDI ≤ 8 BDI & HRSD ≤ 10 SADS-Change, RDC
ADM	Casacalenda et al. (2002)	Remission - ITT	HRSD ≤ 6 or 7 Raskin Depression Scale ≤ 5
		Remission - Completer	HRSD ≤ 6 or 7
	De Maat et al. (2006)	Remission	HRSD ≤ 6, 7, 8 or 9 HRSD ≤ 9 and BDI ≤ 8 HRSD ≤ 9 and BDI ≤ 9
	Parker et al. (2008)	Response - ITT	BDI ≤ 7 or 9 BDI ≤ 9 after at least 12 sessions and 15 weeks of treatment.
		Response - Completer	BDI ≤ 9 BDI and HRSD ≤ 9 BDI ≤ 9 after at least 12 sessions and 15 weeks of treatment. 50% decrease in BDI (or HRSD) after maximum of 12 weeks. BDI ≤ 14 following at least 50% reduction in score.
Combined Therapy	de Maat et al. (2007)	Remission- ITT	HRSD ≤ 6, 7 or 8 BDI ≤ 10 BDI ≤ 9 and HRSD ≤ 9 BDI ≤ 8 and HRSD ≤ 9
	Friedman et al. (2004)	Recovery - Completer	HRSD ≤ 6, 7 or 8
		Recovery - ITT	HRSD ≤ 6 or 7

Key: ADM = antidepressant medication; BDI = Beck Depression Inventory; Completer = completer sample; HRSD = Hamilton Rating Scale for Depression; ITT = intention to treat sample; MDE = major depressive episode; MDD = major depressive disorder; RDC = Research Diagnostic Criteria; SADS = Schedule for Affective Disorders;.

See original reviews for references to instruments.

Table 9. Definitions of Follow-up Outcome Used in Review Studies

Comparison	Review	Outcome Criteria	Definitions of Outcome
Controls	Vittengl et al. (2007b)	Relapse/ recurrence	MDE. MDE or retreatment for depression
Psychotherapies	Leichsenring (2001)	Remission or improvement	BDI ≤ 8. LIFE-II, MDD. SADS-Change, RDC
	Vittengl et al. (2007c)	Relapse/ recurrence	MDE. BDI ≥ 16. BDI ≥ 16 or BDI ≥ 9 and retreatment for depression.
ADM	De Maat et al. (2006)	Relapse	BDI >15. Physician indicated need for treatment. Meeting RDC criteria for MDD for more than 2 weeks. Two BDI scores > 15 separated by 1 week. IDS > 29. Meeting criteria for MDD for more than 2 weeks or HRSD > 13.
	Vittengl et al. (2007d)	Relapse/ recurrence	MDE. BDI ≥ 16 for 2 weeks or more. IDSC ≥ 21 for 2 months or more. MDE or HRSD ≥ 14 for 2 weeks or more. MDE or retreatment for depression. MDE & BDI ≥ 15 or HRSD ≥ 16. Retreatment for depression or BDI ≥ 16.
Combined Therapy	Friedman et al. (2004)	Relapse	Retreatment for depression or BDI ≥ 16. BDI ≥ 16 for 2 weeks or more. Physician indicated need for treatment.
	Vittengl et al. (2007e)	Relapse/ recurrence	Retreatment for depression or BDI ≥ 16. BDI ≥ 16 for 2 weeks or more. IDSC ≥ 21 for 2 months or more.

Key: ADM = antidepressant medication; BDI = Beck Depression Inventory; HRSD = Hamilton Rating Scale for Depression; IDS = Inventory of Depressive Symptomatology; IDSC = Inventory of Depressive Symptomatology – Clinician Version; MDE = major depressive episode; MDD = major depressive disorder; RDC = Research Diagnostic Criteria; SADS = Schedule for Affective Disorders;

See original reviews for references to instruments.

4.3.3 Results of Meta-Analyses

Post-treatment Results

The post-treatment results for reviews' comparisons of psychotherapy with alternative treatments are presented in Table 10. The table shows that the majority of reviews presented overall meta-analytic results in terms of the clinical significance of treatments, i.e. rates of remission, response or relapse-recurrence (Casacalenda et al., 2002; de Maat et al., 2006; de Maat et al., 2007; Parker et al., 2008; Vittengl et al., 2007). Friedman et al. (2004) analysed categorical outcomes but presented overall comparisons in terms of Cohen's d. Significant statistical heterogeneity between the results of individual studies contributing to meta-analysis was found in one of six reviews (Parker et al., 2008).

Psychotherapy vs. Controls

Two reviews reported significantly better outcomes for psychotherapy in comparison to control conditions (Casacalenda et al., 2002; Vittengl et al., 2007). Casacalenda et al. (2002) reported that remission for acute psychotherapy was significantly higher than for control conditions in both ITT (47.9% vs. 27.7%) and completer analyses (59.5% vs. 24.6%). However, the superiority of psychotherapy to controls in their completer analysis was only identified following the removal of Herceg-Baron et al. (1979) where patient attrition was considered excessive (Casacalenda et al., 2002). Controls in Casacalenda et al. comparisons consisted of pill placebo (3 studies), treatment as usual (TAU, 2 studies) or 'supportive therapy'. The latter was intended as a non-treatment comparison condition for psychotherapy where patients could request one therapy session per month in addition to a scheduled monthly assessment (Herceg-Baron et al., 1979). Casacalenda et al. (2002) reported that approximately 45% of patients in the two TAU conditions received ADM. Vittengl et al. (2007) reported that the relapse/recurrence rate was significantly lower at the end of C-CT than that seen for non-treatment controls (12% vs. 38% respectively).

Psychotherapy vs. Psychotherapy

A single review (Leichsenring, 2001) reported no significant difference between the efficacies of acute STPP and CBT in terms of remission or improvement (Table 10).

Psychotherapy vs. Medication

Three reviews reported no evidence for the superiority of psychotherapy or ADM in terms of remission or symptom reduction (Casacalenda et al., 2002; de Maat et al., 2006; Parker et al., 2008). Parker et al. (2008) attributed the single significant result that favoured CBT over ADM in Table 10 to bias arising from significantly higher attrition rates in ADM patients. However, Parker et al. (2008) identified significant statistical heterogeneity between the results of individual studies for all four of their meta-analytic comparisons. Table 10 reveals

that pooled quantitative estimates of psychotherapeutic efficacy showed considerable variation between reviews. For example, the estimated ITT psychotherapy remission rates for Casacalenda et al. (2002) and de Maat et al. (2006) were 47.9% and 37.9% respectively, whilst the ITT relative risk of remission in Parker et al. (2008) and de Maat et al. (2006) were 1.795 and 0.91 respectively.

Psychotherapy vs. Combined therapy

Two reviews reported an advantage for combined therapy over psychotherapy with no evidence of heterogeneity between included studies (de Maat et al., 2007; Friedman et al., 2004). However, only de Maat et al. (2007) reported the significance of their results which showed significantly higher pooled ITT remission rates for combined therapy (46%) compared to psychotherapy alone (34%). However, de Maat et al. (2007) performed sensitivity analyses which revealed that combined therapy was superior to psychotherapy only for chronically depressed patients who were moderately depressed at pre-treatment. This will be further described in the discussion.

Follow-up results

Follow-up results are presented in Table 11. Two reviews presented categorical results (de Maat et al., 2006; Vittengl et al., 2007), and two presented results in terms of symptomatic reduction (Friedman et al., 2004; Leichsenring, 2001). Only the comparison of C-CT with non-active controls (Vittengl et al., 2007b), reported on additional treatment during the follow-up phase. No review identified statistically significant heterogeneity between the results of their included studies.

Psychotherapy vs. Controls

A single review (Vittengl et al, 2007b) reported that relapse rates were significantly lower in C-CT patients than in no-treatment controls during follow-up (40% versus 73% respectively over a mean of 153 weeks).

Psychotherapy vs. Psychotherapy

Two reviews failed to identify the superiority of CT over STPP (Leichsenring, 2001) or other depression specific psychotherapies (Vittengl et al., 2007c) at follow-up. Relapse rates during follow-up (mean = 92 weeks) were 25% and 29% for CT and other depression specific psychotherapies respectively (Vittengl et al., 2007c).

Psychotherapy vs. ADM

Two reviews reported that psychotherapy was superior to ADM in the prevention of relapse during follow-up ranging between 52 to 104 weeks (de Maat et al., 2006; Vittengl et al., 2007d). Table 11 shows that the relapse rate for psychotherapy patients of 27% in de Maat

et al. (2006) was numerically lower than the 39% seen for Vittengl et al. (2007d). However, relapse for ADM patients in both studies was similar at approximately 60%.

Psychotherapy vs. Combined therapy

One review (Vittengl et al., 2007e) reported no difference in relapse between CT and combined therapy during a mean follow-up period of 61 weeks (33% and 39% respectively). Friedman et al. (2004) reported a Cohen's d of 0.12 that favoured combined therapy but failed to indicate its significance level.

Summary

Where reviews made the same treatment comparisons they reached the same conclusions. The reviews indicated that psychotherapy is more efficacious than no treatment, but that psychotherapies do not differ at post-treatment and follow-up. Comparisons between psychotherapy and ADM indicated equivalent efficacy at post-treatment, but that psychotherapy is more effective by follow-up. The combination of psychotherapy and ADM appears to be more efficacious at post-treatment but not at follow-up.

Table 10. Post-treatment Comparisons with Psychotherapy

Comparison	Review	Outcome	Sample	N_S	Results		C.I (95%)
Controls	Casacalenda (2002)	Remission	ITT	6	Psychotherapy Controls	47.9% 27.7%**	37.8 – 57.9 15.7 – 39.7
		Remission	Completer$^{\Omega}$	2	Psychotherapy Controls	59.5% 24.6%**	n/a
	Vittengl (2007a)	Relapse/ Recurrence over a mean of 41 weeks	Unclear	4	C-CT Controls AUC	12% 38% 0.61*	n/a n/a 0.53 – 0.68
STPP vs CBT	Leichsenring (2001)	Remission or Improvement	Unclear	5	Cramer's Φ	0.08	n/a
ADM	Casacalenda (2002)	Remission	ITT	6	ADM Psychotherapy	46.2% 47.9%	37.6 – 54.8 37.8 – 57.9
		Remission	Completer$^{\Omega}$	2	ADM Psychotherapy	61.8% 59.5%	n/a
	De Maat (2006)	Remission	ITT	10	ADM Psychotherapy Relative Risk	34.8% 37.9% 0.91	n/a n/a 0.79 – 1.06
	Parker (2008)	Response	ITT Completer	5 7	Relative Risk	1.795†*$^{\partial}$ 1.11†	n/a
		Symptom reduction	ITT Completer	5 5	Cohen's d for BDI	-0.353† -0.173†	-0.81 – 0.10 -0.64 – 0.29
Combined Therapy	De Maat (2007)	Remission	ITT	7	Psychotherapy Combined therapy Relative Risk	34% *** 46% 1.32 $^{\Delta}$	1.12 – 1.56
	Friedman (2004)	Symptom reduction	Completer	2	Cohen's d for BDI	+0.1$^{\P\Delta}$	n/a
		Recovery	Completer ITT	3 2	Cohen's d for HRSD	+0.69$^{\Delta\P}$ +0.24$^{\Delta\P}$	n/a

Key: ADM = antidepressant medication; AUC = Area under the curve; BDI = Beck depression inventory; CBT = cognitive behavioural therapy; Completer = completer analysis; C-CT = continuation phase cognitive therapy; C.I. = 95% confidence interval; HRSD = Hamilton rating scale for depression; ITT = intention to treat analysis; n/a = not available; N_S = number of studies used in analysis; STPP = short term psychodynamic psychotherapy. Ω = Following removal of Herceg-Baron et al (1979).

Δ = favoured combined therapy
¶ = unknown significance level
∂ = favoured CBT
† = significant heterogeneity between included studies

* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$

Table 11. Follow-up Comparisons with Psychotherapy

Comparison	Review	Outcome	N_S	Follow-up (weeks)	Results		C.I (95%)
Controls	Vittengl (2007b)	Relapse/recurrence over a mean of 153 weeks	5	69 – 312	C-CT Controls AUC	40% 73% 0.64*	n/a 0.57 – 0.72
STPP vs CBT	Leichsenring (2001)	Remission or improvement	4	26 - 104	Cramer's Φ	0.12	n/a
	Vittengl (2007c)	Relapse/recurrence over a mean of 92 weeks	4	52 – 104	CT Other PT AUC	25% 29% 0.50	n/a n/a 0.42 – 0.58
ADM	De Maat (2006)	Relapse	6	52 - 104	ADM Psychotherapy Relative Risk	57% 27% 0.46***	n/a n/a 0.33 – 0.65
	Vittengl (2007d)	Relapse/recurrence over a mean of 68 weeks	7	52 - 104	ADM CT AUC	61% 39% 0.61*	n/a n/a 0.53 – 0.67
Combined Therapy	Friedman (2004)	Symptom reduction	3	26 - 104	Cohen's d	- 0.12¶Δ	n/a
	Vittengl (2007e)	Relapse/recurrence over a mean of 61 weeks	3	52 - 104	CT CT plus ADM AUC	33% 39% 0.51	n/a n/a 0.42 – 0.61

Key: ADM = antidepressant medication; CT = cognitive therapy; CBT = cognitive behavioural therapy; C-CT = continuation phase cognitive therapy; C.I. = 95% confidence interval; N_S = number of studies used in analysis; Other PT = other depression specific psychotherapies.

Δ = favours combined therapy
¶ = unknown significance level

* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$

For Vittengl et al. (2007): b = C-CT vs non active controls
c = A-CT vs other depression specific psychotherapies
d = A-CT vs ADM
e = A-CT vs combined therapy

Table 12. Within Review Risk of Bias Data

Review[α]	Search Period	Databases Searched	Reviewers Assessed Validity of Studies?	Sample Analysed
Casacalenda (2002)	until 2000	Medline PsychINFO	yes	ITT Completer
de Maat (2006)	1980 to 2005	Medline EMBASE CCTR CDRP PsychINFO	yes	ITT
de Maat (2007)	1980 to 2005	Medline EMBASE CCTR CDRP PsychINFO	yes	ITT
Friedman (2004)	1967 to 2002	Medline PsychINFO	no	ITT Completer
Leichsenring (2001)	1966 to 1998	Medline Psychlit	yes	Unclear
Parker (2008)	1977 to 1996 [β]	Medline [β] EMBASE	no	ITT Completer
Vittengl (2007)	until 2006	Medline PsychINFO	no	Unclear

Key: CCTR = Cochrane Controlled Trials Register; CDRP = Cochrane Database of Reviews and Protocols.

α = first author only
β: studies originally identified in Gloaguen et al. (1998)

4.3.4 Assessments of Review Bias

Within Review Risk of Bias

The risk that meta-analyses provide unreliable results has been reduced by the introduction of systematic methods to identify and assess eligible studies as valid for inclusion. Important factors affecting the validity of eligible studies include inadequate randomisation methods, and between-group differences in patient attrition or treatment integrity (Perepletchikova and Kazdin, 2005). Where results are combined from studies that differ widely for these factors, it is likely that the results of meta-analysis will biased (CRD, 2009). It is therefore essential that studies of poorer quality are identified in order that they be excluded or investigated concerning their potential influence on meta-analytic results via sensitivity analysis (CRD, 2009). An evaluation of substantive and quality data revealed that reviews differed on factors known to increase the risk of providing biased results.

No review was identified as a systematic review by its authors and there was considerable variation in the reporting of methodological details. Our quality of review instrument showed that only two reviews provided adequate detail concerning their method of data extraction (de Maat et al., 2006, 2007). However, all reviews clearly described their eligibility criteria, included appropriate studies and based outcome comparisons on widely used symptom rating scales or a diagnosis of depression. There was no evidence that the pooling of included studies in meta-analysis in any review was inappropriate according to our quality appraisal instrument. However, it was found that Casacalenda et al. (2002) did not adhere to recommended methods for pooling individual study effect sizes. Only Leichsenring (2001) and Friedman et al. (2004) failed to provide confidence intervals for their main results, with the latter providing no indication of their statistical significance. All reviews addressed relevant issues concerning the generalisability of their results. Additional within review risk of bias data is presented in Table 12.

Search

Six reviews searched electronic databases as their primary source (Table 12). The seventh (Parker et al., 2008) was a replication of a previous meta-analysis with only the studies identified in the previous work assessed for eligibility (Gloaguen et al., 1998). De Maat et al. (2006, 2007) specified a priori that included studies must be published after 1980, the year the DSM III was published. Friedman et al., (2004) and Leichsenring (2001) gave no reason for their earliest search date. Table 12 shows that two reviews had no earliest date

limitation for the publication of included studies (Casacalenda et al., 2002; Vittengl et al., 2007). Two reviews required that studies be published in English (Casacalenda et al., 2002; Friedman et al., 2004). Four reviews (de Maat et al., 2006; de Maat et al., 2007; Parker et al., 2008; Vittengl et al., 2007) included one study not published in English (Hautzinger et al., 1996). No review included unpublished studies nor tested for publication bias. Whilst no review performed a search that corresponded to the methods recommended for a full systematic review, the detail of reporting and comprehensive search undertaken by de Maat et al. (2006, 2007) suggested that their results were the least likely to be affected by search bias.

Eligibility criteria

All reviews required that included studies were RCTs comparing outcomes for adults diagnosed with depression according to a classificatory diagnostic scheme. Only de Maat et al. (2006, 2007) required that independent reviewers agree for study inclusion. All reviews reported additional eligibility criteria which are presented along with their rationale in Table 13.

Reviews' assessments of study validity

Four reviews reported assessments of the validity for their included studies (Table 12). However, assessment according to published standards was used in only three (Casacalenda et al., 2002; de Maat et al., 2006; de Maat et al., 2007). De Maat et al. (2006, 2007) required that eligible studies meet published quality criteria in terms of randomisation, reporting of attrition and the use of blinded outcome assessments. Also, patients in all included studies were required to receive equivalent amounts of treatment contact to minimise performance bias (de Maat et al., 2006; 2007). Only de Maat et al. (2006, 2007) required that eligible studies use methods to ensure that medication was administered at a therapeutic dose. However, despite providing the clearest descriptions concerning study validity, de Maat et al. (2006, 2007) did not report that the validity of studies was assessed beyond this requirement. Casacalenda et al. (2002) reported that a post hoc assessment showed that all their studies had used blinded assessments, however, they reported that 50% of studies had failed to provide an adequate description of the randomisation process.

None of the remaining reviews reported assessments of study quality according to published standards. However, Table 6 shows that six of eight studies in Parker et al. (2008) and three of four in Friedman et al. (2004) were included in Maat et al.'s (2006, 2007) post-treatment

comparisons and thus met published standards. Similarly, Table 7 reveals that all follow-up studies in Friedman et al. (2004) and Vittengl et al. (2007e) met de Maat et al.'s validity criteria, as did the majority of those included in Vittengl et al. (2007d).

Table 13. Additional Eligibility Criteria for Studies Included in Reviews

Review	Eligible studies required to:	Rationale:
Casacalenda et al. (2002)	Compare treatment with controls.	To answer criticisms that previous meta-analyses included studies with no empirical evidence of treatment efficacy (Klein, 2000).
	Provide remission rates by treatment.	To provide clinically relevant results.
de Maat et al. (2006, 2007)	Meet methodological quality criteria of the Cochrane Collaboration. Include only psychiatric outpatients. Ensure adequate medication. Use formal psychotherapies lasting less than 6 months.	To reduce the methodological and clinical heterogeneity of studies included in previous meta-analyses.
	Provide remission rates by treatment.	To provide clinically relevant results.
Friedman et al. (2004)	Use same psychotherapy in both treatment conditions.	None given.
Leichsenring (2001)	Include 20 or more patients in treatments.	Increased statistical power.
	Provide 13 or more sessions of psychotherapy.	To ensure adequate psychotherapy provided.
Parker et al. (2008)	Compare CBT as stand-alone treatment with ADM.	To perform a re-analysis of Gloaguen et. al.'s (1998) meta-analysis by excluding potentially confounding studies where some CBT patients received ADM.
	Report outcomes for the BDI alone.	BDI used by Gloaguen et al. (1998).
Vittengl et al. (2007)	Provide follow-up data for CBT versus other treatments. Provide relapse/recurrence rates for responders to acute treatment.	To provide clinically useful follow-up data for both acute and continuation phase CBT.

Synthesis methods

Estimates of treatment efficacy can be based on those patients who complete a predefined adequate course of treatment (completer analysis) or on all patients who commence treatment (intention to treat analysis, ITT). Meta-analysis of individual study ITT data provides protection against the possible emergence of non-randomisation bias due to differential attrition between treatments groups (CRD, 2009). Table 12 shows that three reviews presented post-treatment comparisons based on both completer and ITT data, and two for ITT data alone. It was unclear which types of sample were used by Leichsenring (2001) and Vittengl et al. (2007) as well as for all follow-up comparisons. This made it impossible to assess the potential impact of patient attrition on these comparisons.

The reviews used differing methods to synthesise study data. Four used a fixed effects model (Casacalenda et al., 2002; de Maat et al., 2006, 2007; Friedman et al., 2004), whilst Parker et al. (2008) and Vittengl et al. (2007) used a random effects model. It was unclear which model was used by Leichsenring (2001). Casacalenda et al. (2002) was the only review that pooled study effect sizes without first weighting them to account for differences in study size. This raised the possibility that their results were biased by large observed treatment differences which are more likely in smaller studies (CRD, 2009). All reviews except Casacalenda et al. (2002) tested for significant heterogeneity between individual study effect sizes. Appendix D presents further details of the synthesis methods used in reviews.

Summary

De Maat et al. (2006, 2007) adhered to systematic methods and presented the lowest overall risk of bias according to our review appraisal instrument. The lack of information provided by the five remaining reviews, and differences in their methods, made it difficult to determine their relative risk of providing biased results. For example, the risk of bias in Casacalenda et al. (2002) was reduced by their post hoc appraisal of study validity and the inclusion of only blinded assessments. Whilst this was a methodological strength, as unblinded assessments are at greater risk of bias than blinded (Jadad et al., 1996; Lynch et al., 2010), the risk of bias was increased by the possibility of poor randomisation in 50% of studies and their unconventional method of synthesis. In contrast, Parker et al. (2008) used conventional synthesis methods but did not report assessments of study validity, nor sought to identify all eligible studies. However, the majority of studies in post-treatment comparisons of psychotherapy with ADM in Parker et al. (2008) were valid according to criteria used by de Maat et al. (2006). It was found that follow-up comparisons were at greater risk of bias than those for post-treatment due to a lack of information concerning the type of samples used for synthesis. The corresponding lack of information provided by

Leichsenring (2001) and Vittengl et al. (2007) for post-treatment outcomes, and Friedman et al.'s (2004) failure to report the significance of their findings suggests that these reviews provided the highest risk of bias.

Across Review Risk of Bias

It was described earlier that no review searched for unpublished studies nor tested for publication bias. In addition, an examination of substantive and quality data indicated that all reviews were at a risk of bias due to further common methodological limitations. Individual reviews pooled studies that showed considerable variability concerning pre-treatment severity, duration of psychotherapy and definition of remission. Of equal importance was the finding that the integrity of psychological treatments in all studies in every review could not be assured.

Treatment integrity

In order that valid conclusions be made concerning treatment efficacy, it is essential that treatment is provided as intended (Perepletchikova, 2009). Where the integrity of treatment is in doubt, it is not possible to be confident that observed treatment differences are due to differences in the treatments themselves. Confidence in the integrity of psychotherapy is increased where it can be shown that therapists (i) adhere only to the principles specific to the psychotherapy under investigation and, (ii) are competent in the use of these principles (Perepletchikova et al., 2007; Westen et al., 2004). Whilst treatment manuals specify which techniques may and may not be used, failure to monitor therapist performance during psychotherapy risks that proscribed techniques are used, or that prescribed techniques are provided improperly. The issue of treatment integrity was raised by some review authors which led to further scrutiny here.

Concern was expressed by Friedman et al. (2004) and Vittengl et al. (2007) that the integrity of psychotherapy was likely to have been inconsistent between the included studies in their reviews. In addition, de Maat et al. (2006, 2007) reported that included studies had ensured the integrity of ADM treatments but did not report the same for psychological treatments. An examination of the original manuscripts of studies included by de Maat et al. (2006, 2007) revealed that some provided no, or little information concerning the methods used to ensure the integrity of psychological treatments. For example, Murphy et al. (1995) reported that psychotherapists were required to demonstrate competence in the provision of CBT prior to the study, whereas, Blackburn et al. (1997) described only that therapists were

extensively trained. In contrast, de Rubeis et al. (2005) reported therapists' experience, that the least experienced had received training judged to establish their competence in CBT, and that all had followed standard procedures for the provision of CBT. However, none reported that the integrity of psychotherapy was assessed during treatment. Thus, the risk remained across all reviews that their results were biased by the inclusion of studies that had failed to ensure the integrity of psychological treatments.

Treatment duration

In general, reviews pooled results from studies that varied considerably in terms of treatment duration, and the number and frequency of psychotherapy sessions (see Table 4). The pooling of results for psychotherapies across a range of treatment durations is problematic. For example, psychotherapy provided for 16 weeks may produce much higher remission rates than if provided for 8 weeks. If so, any pooled estimate of treatment effect based on both these durations is not representative of either. This may appear unimportant in comparisons of relative efficacy because individual studies compare treatments over the same time interval. However, the inclusion of shorter studies that compare psychotherapy with ADM in analyses could bias results in favour of the latter treatment, as medication is known to produce more rapid symptomatic reduction during early treatment (Watkins et al., 1993; Elkin et al., 1989). The potentially more rapid onset of ADM efficacy may favour its provision over psychotherapy in the shorter term. However, the results of this review suggest that acute psychotherapy provides better protection against relapse than acute ADM during follow-up.

It was implicit in the examples above that the frequency of psychotherapy sessions was the same across different treatment durations. However, the frequency of psychotherapy sessions was also highly variable between included studies in individual reviews. For example, the study by Hautzinger et al. (1996) provided 3 sessions of CBT per week for 8 weeks and was included in three reviews (de Maat et al., 2006; de Maat et al., 2007; Leichsenring, 2001). This contrasts with Blackburn et al. (1981) where 23 sessions of CBT were provided over 20 weeks and was included in three reviews (de Maat et al., 2006; de Maat et al., 2007; Parker et al., 2008). The difference in duration of 12 weeks between these studies is considerable despite patients receiving over 20 sessions in both. This result shows that the intensity of treatment could vary for specific therapy types. If it is the case that the intensity of psychotherapy sessions is a determinant of treatment efficacy, then the inclusion of studies that vary on this factor may confound the results of meta-analysis.

Pre-treatment severity

In addition to a diagnosis of MDD, the vast majority of included studies required that patients meet a minimum criterion score on a symptom severity measure prior to study entry. The range of minimum severity scores in review studies was presented in Table 4 and Table 5. There was considerable variation in the minimum criterion for study entry in all but one review comparison (Vittengl et al., 2007c). For example, Table 4 shows that the minimum criterion score for the HRSD ranged from 10 to 20 for the studies in Parker et al. (2008). According to the American Psychiatric Association, an HRSD score of 10 is classified as mild, whereas, a score of 20 is classified as severe depression (Kriston and von Wolff, 2011). This indicates that the mean pre-treatment severity of patients varied considerably across included studies in the majority of reviews.

This is a matter for concern as decreasing pre-treatment severity is associated with increased response to placebo (Schatzberg and Kraemer, 2000; Fournier et al., 2010). Moreover, in the absence of untreated control data, it is not possible to estimate the proportion of patients whose symptomatic change may actually be attributed to treatment in individual studies (Klein, 1996). Consequently, where reviews made direct treatment comparisons, lower severity studies would be more likely than higher severity studies to produce results that were confounded with placebo response. Thus, there was a risk in reviews that compared the relative efficacy of treatments, that a large proportion of patients in low severity studies did not remit as a direct result of treatment. If so, this would serve to overestimate the efficacy of treatments and likely obscure any treatment differences that may exist in more severely depressed samples.

Definitions of treatment outcome

An inspection of Table 8 reveals that the definitions of remission employed across included studies in reviews showed considerable variation. Similarly, Table 9 shows that definitions of relapse were also highly variable across studies. The use of differing outcome definitions between included studies in reviews is a severe limitation which will be described concerning post-treatment outcomes. The same issues affect analyses of follow-up data. Table 8 shows that the majority of included studies defined remission in terms of a minimum criterion score on the BDI or HRSD (or both). For both measures, lower scores represent a more stringent definition of remission than do higher scores. For example, Zimmerman et al. (2004b) reported that 6.8% of patients scoring 10 or less on the 17-item HRSD still met DSM IV diagnostic criteria for MDD compared to 3.4% for a criterion of 7 or less, and none for 3 or less. However, Table 8 reveals that included studies in individual reviews

frequently used different criterion scores to define remission on the HRSD. For example, the HRSD criterion for studies included in de Maat et al. (2006) ranged from 6 points or less to 9 points or less. Thus, studies using less stringent definitions will have contributed higher remission rates to de Maat et al.'s overall analysis than those using more stringent definitions. Table 8 shows similar results for the BDI, where the minimum BDI criterion ranged from 7 to 10 points across studies included in reviews.

The absence of a consistent definition of what constitutes a clinically significant outcome for each of these measures is problematic for several reasons. Firstly, as definitions become less stringent, it becomes less likely that patients have actually remitted from depression. Consequently, where studies use a range of idiosyncratic outcome definitions, it is unclear to what degree the overall clinical significance rates provided by meta-analysis actually represent remission. Furthermore, where categorical outcomes are used in meta-analysis, it cannot be assumed that the relative efficacy of depression treatments will be invariant as the stringency of definitions change. Thus, it is possible that the results of individual meta-analyses will be biased by larger studies where the stringency of the criterion score used to define remission inadvertently favour a specific treatment type. Finally, it is unclear to what degree pooling the results of studies that have used different measures (e.g. the BDI, HRSD, or both) risks that the results of meta-analyses are biased (Nugent, 2009). However, they will certainly be less precise than results based on studies that employ a standard definition of remission on the same outcome measure (Matt and Navarro, 1997).

4.4 Discussion

This systematic review identified seven meta-analytic reviews which summarised the results of psychological treatment efficacy studies for major depressive disorder. The eligibility criteria ensured the inclusion of only randomised controlled trials that examined treatment effects in samples meeting a formal diagnosis of major depression. The requirement that psychological treatments were based on theoretical models of psychopathology meant that potentially *non bona fide* treatments were excluded. Whilst none of the reviews were described as systematic reviews, they provided the best meta-analytic evidence concerning the current efficacy of individually provided psychological treatments for depression. However, reviews varied with respect to the risk of producing biased results. Examination of substantive and quality data indicated several methodological factors that risked introducing bias into all reviews. Nevertheless, where reviews made the same treatment comparisons they reached the same overall conclusions. The overall conclusions that may be drawn

across the results of included reviews will be discussed whilst bearing in mind individual review risk of bias. Following this, factors which may have biased all reviews are discussed, with a final discussion concerning the limited utility of heterogeneity tests to reveal such sources of bias.

4.4.1 Conclusions Based on the Meta-analytic Results of Reviews

The varying degree by which reviews adhered to the systematic review methodology made it difficult to assess to what degree they risked providing biased results. The results indicated that, irrespective of the quality of individual reviews, follow-up comparisons were at greater risk of bias than post-treatment comparisons due to uncertainties concerning the nature of the samples used in analyses. It was unclear whether follow-up analyses were based on all patients who entered treatment, all those who completed treatment, or only those who remained in contact with investigators during follow-up. Moreover, in all but one analysis (Vittengl et al., 2007b), it was unclear whether patients received treatment for depression during follow-up.

Comparing psychotherapy types & establishment of psychotherapeutic efficacy

Two reviews concluded that psychotherapy was superior to controls at post-treatment Casacalenda et al., 2002; Vittengl et al., 2007a) and one at follow-up (Vittengl et al., 2007b). Two reviews that compared the relative efficacy of specific psychotherapies failed to find any difference between therapy types at post-treatment (Leichsenring, 2001) and follow-up (Leichsenring, 2001; Vittengl et al., 2007c). The results of these reviews suggest that, overall, *bona-fide* psychotherapies were superior to controls and were equally effective in the treatment of depression. However, only Casacalenda et al. (2002) clearly reported the type of samples used in analyses which meant that the comparisons made by Leichsenring (2001) and Vittengl et al. (2007) were at greater risk of bias. Nevertheless, the comparison of C-CT with untreated controls by Vittengl et al. (2007b) was at lowest risk of bias for all follow-up comparisons across reviews, as it was the only one where patients were guaranteed to receive no treatment.

The overall ITT analysis by Casacalenda et al. (2002) showed that post-treatment remission was 47.9% , 46.2% and 27.7% in psychotherapy, ADM and control samples respectively. This suggests that remission in approximately 50% of those receiving an active treatment may have been due to placebo effects. However, it was not possible to quantify this proportion, as Casacalenda et al.'s inclusion of treatment as usual samples in their analysis meant that approximately 23% of their control sample received antidepressant medication. Nevertheless, the proportion remitting due to placebo effects was likely to have been close to

Casacalenda et al.'s control sample rate of 27.7%, as Posternak & Miller (2001) have indicated that up to 20% of patients included in depression treatment studies may remit in the absence of treatment. The evidence also suggests that continuation phase psychotherapy provides significant protection against future depressive episodes compared to acute phase psychotherapy alone. Vittengl et al. (2007) found that both post-treatment (12% vs 38%) and follow-up (40% vs 73%) relapse rates were significantly lower in C-CT samples than in samples who received only acute phase psychotherapy. Thus, the most reliable evidence showed that that 73% of patients receiving acute phase psychotherapy relapsed over a follow-up period of 153 weeks or approximately 3 years (Vittengl et al., 2007b).

Comparison of psychotherapy with medication

The overall results of three reviews (Casacalenda et al., 2002; de Maat et al., 2006; Parker et al., 2008) provided strong evidence that the efficacies of psychotherapy and medication were no different at post-treatment. That neither broad treatment class was superior following the acute treatment of major depression was supported by one of the highest quality reviews according to our appraisal instrument (de Maat et al., 2006). Moreover, whilst Casacalenda et al. (2002) and Parker et al. (2008) demonstrated a greater risk of bias, they reached the same conclusion in analyses that included different studies to those of de Maat et al. (2006). Parker et al. (2008) and Casacalenda et al. (2002) shared one third and two thirds of included studies in common with de Maat et al. (2006) respectively. However, the overall remission rate for psychotherapy of 47.9% reported by Casacalenda et al. (2002) was markedly higher than the potentially more reliable 37.9% reported by de Maat et al. (2006). Whilst the difference will in part have originated in the inclusion of different studies, it is possible that Casacalenda et al.'s (2002) higher rate also resulted from their failure to use weighting in analyses. Thus, the best available evidence from de Maat et al. (2006) suggests that less than 40% of patients who started psychotherapy for major depression remitted by the end of treatment.

In terms of follow-up, the results of the high quality review by de Maat et al., (2006) were similar to those of Vittengl et al. (2007). Both reviews reported that approximately twice as many ADM as psychotherapy patients relapsed over a 1 to 2 year period following treatment. De Maat et al. (2006) reported an overall relapse rate of 27% and 57% for psychotherapy and ADM respectively, whilst the corresponding figures for Vittengl et al. (2007d) were 39% and 61%. Thus, evidence from two reviews indicated that where psychotherapy successfully lead to remission, it provided longer lasting benefit than discontinued medication. However, the conclusions reached by de Maat et al. (2006) and Vittengl et al. (2007) were not wholly independent as the majority of included studies in

both reviews were the same (Table 7). This, and uncertainty concerning the nature of the samples used in follow-up comparisons suggests that the conclusions reached in these reviews were at risk of bias.

Comparison of psychotherapy alone & combined with medication

Two reviews published the statistical significance levels for their comparisons of psychotherapy with combined therapy (de Maat et al., 2007; Vittengl et al., 2007). The high quality review by de Maat et al. (2007) concluded that combined therapy was superior to psychotherapy at post-treatment. This conclusion was based on an ITT analysis where overall remission for combined therapy was 12% greater than for psychotherapy alone (46% versus 34% respectively). In contrast, over a 1 to 2 year follow-up period, Vittengl et al. (2007) found no significant difference in relapse rates between CT plus medication and CT alone (39% versus 33% respectively). However, uncertainty concerning the type of samples employed in analysis and the possibility that patients received treatment during follow-up meant that Vittengl et al.'s results were at risk of bias. Consequently, the most reliable conclusion concerning combined therapy and psychotherapy is that combined therapy was superior to psychotherapy at post-treatment.

Summary

The results of reviews that closely adhered to systematic review methods were at least risk of leading to biased conclusions. In terms of post-treatment outcome, the evidence strongly suggests that there was no difference between the efficacy of psychotherapy and ADM (Casacalenda et al., 2002; de Maat et al., 2006; Parker et al., 2008). There was also high quality evidence suggesting that combined therapy was superior to psychotherapy alone (de Maat et al., 2007). However, this finding will be discussed further in a subsequent section of this discussion, as it did not apply to all patients included in de Maat et al.'s analysis. There was also tentative evidence from one review that psychotherapy was superior to controls at post–treatment (Casacalenda et al., 2002).

In terms of follow-up, the most reliable evidence indicated that continuation phase psychotherapy provided greater protection against relapse compared to acute psychotherapy alone (Vittengl et al., 2007b). In addition, evidence from one of the highest quality reviews also indicated that psychotherapy was associated with a significantly lower probability of relapse compared to medication (de Maat et al., 2006). However, it must be borne in mind that uncertainty concerning the nature of samples meant that the follow-up results for all reviews were at a greater risk of bias than were post-treatment results.

4.4.2 Risk of Bias Across Reviews

Whilst our review quality appraisal instrument indicated that the results of some reviews were at less risk of bias than others, an examination of substantive review data revealed several factors that may have biased the results of all reviews. It was shown that no review searched for unpublished studies, nor checked for publication bias. Consequently, were unpublished studies to have been included, it is possible that the result of individual reviews would have been different. However, irrespective of whether unpublished studies were included or not, four additional factors were identified which risked that reviews provided biased results.

Firstly, the integrity of psychological treatments may have been inconsistent across studies included in reviews. An examination of original manuscripts for the studies included in de Maat et al. (2006, 2007) revealed that they varied considerably in terms of reporting the methods by which treatment integrity was assured. Whilst the detail of reporting for primary studies may not have reflected the efforts made to ensure or assess treatment integrity, it is likely that the results of some were based on poorly implemented psychological treatments. Indeed, Bhar & Beck (2009) have argued that the majority of studies used in recent comparisons of CBT with STPP have not adequately implemented procedures that ensure the integrity of either treatment type. Consequently, when the results of meta-analyses that include such studies find no difference between CBT and STPP they are at best ambiguous (Bhar and Beck, 2009). Moreover, it is possible to speculate that ensuring high levels of treatment integrity for medications is typically easier to achieve than for psychological treatments in comparison studies. If so, then the inclusion of studies that have poorly implemented psychotherapy in meta-analysis may be responsible for the frequent finding that psychotherapy and ADM are no different at post-treatment.

It was also found that the overall duration of psychotherapy and number of sessions available to patients varied considerably between studies included in reviews. In addition, the intensity of psychotherapy typically showed marked variability between included studies in reviews. That is, the average number of sessions per week in some included studies was much higher than in others. Such variability in the overall duration and timing of treatment sessions meant that psychological treatments with potentially different efficacies were combined as ‘psychotherapy’. Evidence that treatment duration is correlated with outcome for specific psychotherapy models was provided by Shapiro et al. (1994; Shapiro et al., 2003). By comparing outcomes for cognitive behavioural or psychodynamic-interpersonal therapy provided over 8 or 16 weeks, Shapiro et al. (2003) concluded that longer treatment was more beneficial for the majority of depressed patients. Moreover, whilst overall

symptomatic reduction on the BDI was no different between 8 and 16 week samples, symptomatic reduction at 16 weeks was significantly greater than at 8 weeks in patients categorised as severely depressed (BDI > 27). The variation in treatment duration across studies in reviews makes it difficult to interpret their results. For example, whilst Parker et al. (2008) compared the post-treatment efficacies of CBT and ADM, CBT was provided over a range of 12 to 24 sessions in studies lasting between 8 to 15 weeks. Given that the onset of ADM efficacy may be more rapid than that of CBT (Watkins et al., 1993; Elkin et al., 1989), to what duration and intensity of CBT did Parker et al.'s finding of no difference between treatments best refer? More importantly, perhaps, is the possibility that the inclusion of shorter versions of established psychotherapies in future meta-analyses could lead to biased overall conclusions that medication is more effective than psychotherapy across all treatment durations. The best evidence from the reviews presented here indicated that psychotherapy was as effective as medication at post-treatment and was better at preventing relapse at follow-up. The intensity of psychotherapy over typical treatment durations and the rapidity of onset of treatment efficacy are areas which warrant further research.

A third factor that was highly variable between included studies in reviews was the mean pre-treatment symptom severity of patient samples. If it is generally the case that an individual's level of pre-treatment severity significantly predicts treatment outcome, then the inclusion of studies that vary widely on this factor makes interpretation of review conclusions problematic. Again, to what severity of depression do the results of meta-analysis apply? In addition, it may be important that patient severity is balanced across treatment groups within individual studies. An examination of original manuscripts showed that the randomisation process in some of the included studies in de Maat et al. (2006) stratified patients by pre-treatment severity to ensure the equivalence of treatment groups on this variable (e.g. Blackburn et al., 1981). However, where primary studies did not use Blackburn et al.'s approach, group equivalence could not be guaranteed solely on the basis of non-significant differences between group means. For example, following an examination of original study data, DeRubeis et al. (1999) revealed that significantly more severely depressed patients (HRSD $\geq$ 20) were entered into the ADM arm of Murphy et al. (1984) according to the BDI. Consequently, because placebo effects are more marked in less severely depressed samples (Schatzberg and Kraemer, 2000; Fournier et al., 2010), there was a potential source of bias favouring CBT over ADM in Murphy et al. (1984). Whilst such bias ought to be random in nature and thus be cancelled out where meta-analyses contain many studies, the highest number of studies in any review was 10. Consequently,

without stratified randomised allocation by pre-treatment severity it is possible that statistically influential studies will bias the results of meta-analysis.

Finally, individual studies included in reviews used idiosyncratic definitions of remission which meant that the stringency by which remission was defined was variable. Thus, some studies will have underestimated, whilst others overestimated, the proportion of patients who achieved remission. Consequently, it is unclear to what degree the overall rates reported by reviews actually represented remission. Moreover, whilst this lack of clarity concerning the clinical significance of reported outcomes is itself undesirable, the variation in stringency between the included studies in reviews raised the possibility that review conclusions were biased. It is possible that the relative efficacy of treatments were confounded with choice of outcome measure and remission criterion employed in some studies. Again, where the results of a statistically influential study are biased, it is possible that the overall results of meta-analysis will also be biased. Unfortunately, this problem cannot be overcome by using continuous data effect sizes to compare treatments, as Churchill et al. (2001) revealed that these correlate poorly with clinically significant outcome.

In order to reduce the risk of bias in meta-analyses seeking to compare treatments in terms of remission, it is necessary that included studies employ an empirically-based standard definition of clinical significance which best represents remission. The Jacobson method of clinical significance is ideally placed to do this and is described in the next chapter. However, a limitation of the Jacobson method is that normative data for outcome measures and individual patient data (IPD) from primary studies are required. Consequently, the Jacobson method cannot be used for conventional meta-analysis where the results are based on summary data from already published studies.

The utility of heterogeneity testing for biased results

Heterogeneity testing is used in meta-analysis to identify whether the observed variation between individual studies' effect sizes is greater than would be expected due to measurement error (CRD, 2009). A significant result may indicate that more than one population has been included in meta-analysis, or that the effect sizes[9] of one or more included studies may be biased. However, heterogeneity tests suffer from low power where overall information is sparse, or where greater than 50% of included information derives from a single study (Hardy and Thompson, 1998). Given that relatively few studies were included in review analyses, it was unlikely that heterogeneity tests would have been able to

[9] Based on either continuous or categorical outcomes.

identify studies that were biased due to the factors discussed above. Indeed, no evidence of significant heterogeneity was found between studies in de Maat et al.'s primary comparison of combined therapy with psychotherapy (de Maat et al., 2007). However, they conducted a sensitivity analysis[10] which revealed that combined therapy was superior to psychotherapy only in patients with moderately severe *and* chronic depression (de Maat et al., 2007). Their overall primary comparison was biased by a single study of chronic depression which contributed 44% of the data to the analysis (Keller et al., 2000). Were de Maat et al. (2007) to have ignored the clinical variability between studies and relied solely on heterogeneity testing, their primary conclusion risked being interpreted as applicable to all depressed patients included in their review.

In addition, the findings of the present review suggest that the interpretation of significant statistical heterogeneity is virtually impossible in conventional meta-analyses of depression treatment studies. To illustrate this point, imagine that several studies comparing CBT with pill placebo are included in meta-analysis. Furthermore, suppose that heterogeneity testing has shown that the effect size favouring CBT in one study (study A) is significantly greater than that of remaining studies. This situation could arise for several reasons: (i) CBT in study A may have been more efficacious than in other studies due to higher levels of treatment integrity (ii) patients in study A were more severely depressed than in remaining studies; thus a comparatively low placebo response may have led to CBT appearing more efficacious than in remaining studies (iii) CBT in study A was provided more frequently than in remaining studies which potentially increased its efficacy over remaining studies, (iv) the definition of remission used in study A inadvertently resulted in a significantly larger effect size favouring CBT than seen across remaining studies. Taken together, these examples show that the interpretation of significant heterogeneity between the outcomes of depression treatment studies is virtually impossible within conventional meta-analysis as currently practised.

Whilst the conclusions of this review may be limited by including only published reviews in English, it is unlikely that there were relevant unpublished reviews. A major limitation is that initial assessments of review quality were based on the details provided in review manuscripts. Thus, it is possible that space limitations imposed by publishers meant that our conclusions do not reflect the quality of research undertaken by reviewers. Finally, we did not investigate the affiliations nor sources of funding of review authors as a potential source of bias.

[10] Not reported in this review

4.5 Summary & Concluding Remarks

The increasing use and influence of meta-analysis as a method to summarise the results of psychotherapy trials within a systematic review makes an investigation of the potential problems with the approach timely. The best evidence from included reviews suggested that 38% to 48% of patients who start individual psychotherapy will remit by the end of treatment. However, remission in approximately half of these patients may be due to placebo effects. There was strong evidence that the efficacies of psychotherapy and ADM do not differ at post-treatment and limited evidence that psychotherapy is superior to ADM at preventing relapse. Nevertheless, approximately 70% of those who remit following psychotherapy will relapse over the next three years.

However, confidence in these conclusions is undermined by several important methodological factors that may bias the results of all meta-analyses of depression treatment studies. Foremost is the likelihood that the integrity of psychological treatments was sub-optimal in some of the primary outcome studies. Nevertheless, were it the case that all treatments were properly provided, there were studywise variations concerning the timing of psychotherapy sessions, the definition of outcome, and the average pre-treatment severity of patients that were potential sources of bias. These factors may have reduced the validity of reviews' conclusions because it cannot be guaranteed that meta-analysis can control for individual study bias (Matt and Navarro, 1997).

Finally, irrespective of the risk that they led to biased conclusions, the use of idiosyncratic outcome definitions of treatment efficacy compromised the conclusions that could be drawn. A standardised operational definition of the clinical significance of treatment is needed to allow a clearer assessment of the absolute efficacy of psychological treatments for depression. Study two will attempt to address this issue and is presented in chapter 6 following a description and critique of the Jacobson method.

Chapter Five

A Review and Critique of the Jacobson Method Approach to Clinical Significance

5.1 Introduction

In study 1, the majority of reviews based their meta-analyses on the remission rates reported by individual studies. That categorical remission data was used in many of the meta-analyses in study 1 reflects an increasing acceptance that overall treatment comparisons using standardised mean differences are difficult to interpret (CRD, 2009). Where empirical evidence is required to support the use of one treatment in preference to another, it is essential that clinical significance rates are included in meta analyses. This is because significant between-treatment differences in the magnitude of change over the course of treatment may be of little clinical relevance (Chambless and Hollon, 1998). It is not enough to know that treatments differ statistically, it is also important to know whether they differ in a clinically meaningful way. The clinical relevance of psychotherapy outcome research has been greatly enhanced by supplementing inferential statistics with reports of the clinical significance of treatment effects. Clinical significance attempts to capture whether therapy has produced meaningful change and has been operationalised in several ways. One method that has been widely applied across treatment approaches and psychiatric disorders is the empirically derived approach of Jacobson and colleagues (Jacobson et al., 1984; Jacobson and Revenstorf, 1988; Jacobson and Truax, 1991). Whilst alternative methods exist for determining the clinical significance of individual outcomes in treatment studies, the 'Jacobson method' is increasingly popular (Ogles et al., 2001) and has been recommended as the method of choice (Lambert and Ogles, 2009). This chapter reviews the development of clinical significance methodology from its origins in applied behaviour analysis through to Jacobson's last conceptualisation. The strengths and weaknesses of the Jacobson method are then discussed.

5.2 The Development of Clinical Significance

In the late 1950s, concerns that psychotherapy was ineffective saw an increase in the number of treatment studies attempting to show that psychotherapy was superior to no-treatment (Kiesler, 1966). However, early studies typically suffered from low internal validity due to

factors such as biased sampling, poor specification of treatments and the failure to differentiate between psychotherapy models (Bergin, 1966; Garfield, 1981). Nevertheless, in 1966 Bergin identified 7 studies whose methods were sufficiently sound to draw important conclusions concerning psychotherapy research and practice. None of Bergin's included studies demonstrated that treatment was superior to no-treatment in terms of mean improvement on outcome measures. However, because treatments typically produced greater variability on outcome measures than observed in no treatment, Bergin concluded that the degree of both improvement and deterioration was more marked in treated groups. Thus, inferential statistics had failed to identify important differences between treatment and no treatment groups in the 7 studies, This suggests that inferential statistics alone, provide a limited assessment of treatment efficacy (Bergin, 1966), because within group variability is disregarded (Garfield, 1981; Barlow, 1981; Hugdahl and Ŏst, 1981).

More recently, concerns have been raised that the lack of influence of clinical research on clinical practice may be ascribed to the use of traditional methodologies and a reliance on inferential statistics (Barlow, 1981; Westen et al., 2004; Boisvert and Faust, 2006). One concern is that, although inferential statistics can reveal the relative values of treatments under comparison, no conclusions concerning the absolute value of those treatments may be drawn. This is because a statistically reliable result may have little clinical relevance. For example, a patient may make statistically significant improvement on measures of symptomatic state, yet still be considerably impaired in everyday functioning. Another concern that may still limit the influence of research findings in clinical settings is the patient uniformity myth (Kiesler, 1966) which assumes that patients with the same diagnosis will respond similarly to a particular treatment (Westen et al., 2004). However, whilst two individuals may share the same diagnostic category, the difference between their symptoms may be far more notable than the similarities. For example, where two treatments are equally efficacious according to mean comparisons, it is possible that one treatment produces high levels of improvement for a minority of patients whilst the other produces minimal improvement in most patients (Hugdahl and Ŏst, 1981). Thus, without knowing the proportion of patients who benefit, remain unchanged or deteriorate, it is very difficult for clinicians to generalise from the research study to clinical practice.

Early proposals for evaluating the individual effects of treatment originated in the field of applied behaviour analysis. Risley suggested that interventions should be evaluated in terms of both experimental and therapeutic criteria (Risley, 1970; cited in Kazdin and Kazdin, 1977). The experimental criterion concerns whether or not the intervention was responsible for the behaviour change. For example, in an applied behavioural intervention, the

experimental criterion is satisfied when the experimental variable is shown to reliably control the emergence of a desired behaviour (Baer et al., 1968). The therapeutic criterion concerns whether the behaviour change is meaningful to the client. This can be readily applied in situations where the presence or absence of behaviours denotes success. For example, where treatment eliminates self-injury in a person with autism, it clearly meets a therapeutic criterion of no self-injury. However, should treatment result in only a 50% reduction in the number of self-injurious episodes, it does not meet the therapeutic criterion when the outcome is defined in all or nothing terms, despite the possibility that a reduction in self injurious behaviour may represent a significant improvement in wellbeing. This illustrates that when symptoms remain, defining what is a meaningful or clinically significant outcome is problematic (Kazdin and Kazdin, 1977).

A potential solution to this problem lay in the concept of social validation (Wolf, 1978). In an attempt to operationalise the social benefits of treatment, Wolf (1978) argued that the effects of behavioural interventions according to objective measures should be compared with their effects as judged by consumers. The impetus to socially validate treatment efficacy led to the development of empirical procedures that can determine whether clinically significant change has occurred (Kazdin and Kazdin, 1977). According to the social validation approach, treatment efficacy can be assessed by comparing the behaviour of treated patients with well functioning peers (social comparison), or, by the subjective evaluation of individuals in everyday contact with the patient. The development of clinical significance methods has largely drawn on the social comparison method because it can address problems associated with the therapeutic criterion where symptoms remain. However, the validity of the social comparison method is highly dependent upon the normative reference group. Kazdin & Kazdin (1977) stressed that normative data needs to be obtained from a population which is similar to the patient in all but dysfunctional behaviour. "The level of behaviour of the peers who did not warrant or receive treatment can serve as the criterion by which the success or clinical importance of treatment is evaluated. If treatment has effected marked changes in behaviour, the client's performance should fall within the normative level of his peers" (Kazdin and Kazdin, 1977, pp 431-432). This notion lies at the heart of Jacobson and colleagues' (Jacobson et al., 1984; Jacobson et al., 1999; Jacobson and Truax, 1991) approach to clinical significance.

An increasing recognition of the need to report psychotherapy research findings in a more clinically meaningful way, specifically to increase their relevance to clinical practice, has resulted in a variety of operational definitions of clinical significance. However, some definitions are somewhat arbitrary, such as that of 'response' which is often defined as a

50% reduction in pre-treatment score on outcome measures (Hiller et al., 2012). Alternatively, where definitions are clinically relevant they are essentially subjective, e.g. the recommended criterion score of 7 or less for remission on the Hamilton rating scale for depression (Rush et al., 2006). Clearly, a standardised approach is required to overcome the methodological and interpretative difficulties associated with idiosyncratic definitions of clinical significance. In the 1980s Jacobson, Follette & Revenstorf (1984) argued that (i) the use of idiosyncratic definitions of clinical significance by researchers represented a rather limited advance in psychotherapy outcome research and (ii) an agreed and valid method of determining the clinical significance of treatment effects was required that would permit between and across study comparisons that was applicable across a wide range of psychiatric disorders.

5.3 The Jacobson Approach to Clinical Significance

The Jacobson approach is based on the premise that definitions of clinically significant change should incorporate the concept of a return to normal functioning. "Clients entering therapy are viewed as part of a dysfunctional population and those departing from therapy as no longer belonging to that population" (Jacobson and Truax, 1991, p 13). Two criteria are used to determine whether clinically significant change has taken place: (i) patients receiving treatment should move from a theoretical dysfunctional population to a functional population on symptom measures, and (ii) the change must be statistically reliable. Movement into the functional distribution is determined by establishing a cut-off point beyond which it is more likely that the patient's post-treatment symptom score belongs to the functional rather than the dysfunctional population. Reliability is assessed using the reliable change index (RCI) appropriate to specific outcome measures. Comparing an individual's pre- to post-treatment change score with the RCI ensures that the observed change score is genuine and not due to measurement error.

5.3.1 Operational Definition of Clinical Significance

Cut-off points

Jacobson et al. (1984) proposed three methods to determine whether an individual's level of functioning falls within the functional distribution following treatment. Each method relies on the creation of cut-off points on the target variable chosen to index the clinical problem. The cut-off points are: *(a),* the patient's level of functioning falls outside the range of dysfunctional distribution, defined as two standard deviations beyond the mean in the direction of functionality; *(b),* the patient's level of functioning falls within the range of the

normal population defined as falling within two standard deviations of the mean of the functional or normal population or *(c)*, the patient's post-treatment score is more likely to be drawn from the functional distribution than the dysfunctional distribution.

Reliable change index

The Reliable Change Index (RCI) is used to account for the less than 100% reliability of psychometric instruments and ensures that the magnitude of change is statistically reliable. If the RCI is greater than 1.96, then the change is considered significant at the .05 level and reliable. The original method for calculating the RCI was amended by Christensen & Mendoza (1986) as the original RCI was based on the standard error of measurement surrounding a single true score. However, because the RCI is used to judge the reliability of change as quantified by two scores, Christensen & Mendoza (1986) proposed that the standard error of difference should be used to calculate the RCI. Accordingly, Jacobson & Revenstorf (1988) adopted this amendment and recommend its use. It is important to note that the RCI is not itself a measure of clinical significance - it only denotes that the observed degree of symptomatic change is greater than that to be expected by measurement error alone.

5.3.2 Guidelines for Choosing Cut-off Points

Using hypothetical examples, Jacobson & Truax (1991) outlined when each cut-off point is appropriate. It is important to bear in mind that each cut-off point will give different estimates of clinical significance. For overlapping distributions cut off point *a* is the most stringent, cut-off point *b* the most lenient and cut-off point *c* occupies an intermediate position. Cut-off point *c* is strongly recommended if appropriate normative data exits for both the functional and dysfunctional distributions. It is the least arbitrary method as it is based on the relative probability of a patient's post-treatment score belonging to either the functional or dysfunctional distribution. This method provides the most accurate estimate of a return to normal functioning as a direct comparison is made with a patient's well functioning peers.

If data for a normative sample is not available, cut-off point *a* is the only alternative. This is the most stringent of the cut-off points for overlapping distributions. The major limitation of cut-off point *a* is that it is less valid than cut-off point *c* as the well functioning population is not taken into account. A further point is that the more overlap between the two distributions the more stringent *a* becomes relative to *c*. Cut-off point *b* can only be used when normative data exists. It provides the most lenient cut off point when the distributions are overlapping and it would seem ill advised to use this cut-off point in this situation. However, for non-

overlapping distributions, *b* is the most stringent cut-off point. Indeed, Jacobson & Truax (1991) argue that in the case of non-overlapping distributions only *b* ensures that a patient has entered the functional distribution. As *b* is solely determined from normative data, the cut-off point will not vary from study to study.

The Jacobson approach can assign patients to one of four outcome categories namely, (i) recovered, (ii) improved, (iii) unchanged and (iv) deteriorated. The recovered category refers to patients who have demonstrated both a statistically reliable improvement and whose post-treatment score falls within the functional range. Improved refers to patients who have demonstrated a reliable improvement in symptom score but have failed to enter the functional range. Unchanged refers to patients whose symptom scores have not reliably changed, whilst deteriorated patients have demonstrated a reliable worsening of their symptoms.

5.4 Critique of the Jacobson Approach

Clinically significant change in Jacobson's terms involves becoming a member of a functional or normal population. However, normative data is often unavailable for measures (Lambert and Ogles, 2009) and it is not always clear how a functional or normative sample should be defined. Jacobson & Revenstorf (1988) stated that an ideal normative sample would not include subjects who were dysfunctional but should include outliers if those individuals were not seeking therapy. This definition only partially addresses the complexity of defining normative samples, as receiving treatment is imperfectly correlated with being dysfunctional. For example, in the National Comorbidity Survey (Kessler et al., 1994), more than 60% of individuals meeting a lifetime psychiatric disorder had not received professional treatment. This means that the majority of individuals diagnosable with a psychiatric disorder would be included as outliers in normative samples according to Jacobson & Revenstorf's (1988) recommendation. However, irrespective of their reasons for not seeking treatment, such individuals are dysfunctional according to objective assessment. Where normative samples include such individuals, the means of the normal and dysfunctional populations will be closer than would be the case if they were excluded (Saunders et al., 1988). Consequently, the use of treatment seeking as a criterion to exclude individuals from normative samples will downwardly bias the amount of symptomatic change required to achieve clinically significant change (Saunders et al., 1988).

Estimates based on populations which include symptomatic individuals provide a less stringent test of a return to normal functioning than a group comprised entirely of asymptomatic individuals. However, this problem is overcome where studies have derived normative population estimates from asymptomatic samples (e.g. Ogles et al., 1995). This approach does not conflict with the initial recommendations of Jacobson & Revenstorf (1988) which Tingey et al. (1996a) criticised as being too vague for the operationalisation of normative samples. Follette & Callaghan (1996) pointed out that the major purpose of the Jacobson methodology is to define a clinically significant outcome in terms of what patients may reasonably expect from therapy. Thus, individual researchers must decide on what type of normative data to employ in analyses, depending on its availability and intended use (Follette and Callaghan, 1996). This approach allows the Jacobson method to be applied in situations where it is unlikely that treatment will return the client to normal functioning; investigators may quantify clinical significance by comparing treatment with the normative population of patients who have previously received the most effective treatment to date (Follette and Callaghan, 1996). Thus, it is clear that investigators should carefully describe the normative reference group used in analyses in order that the psychotherapy field may draw informed conclusions concerning treatment efficacy (Saunders et al., 1988). However, one limiting factor in employing the Jacobson approach is the lack of suitable normative data for many relevant measures of psychopathology (Lambert and Ogles, 2009).

Another area of concern relates to the conceptualisation of distinct normative and dysfunctional populations by Jacobson et al. (1984). Both Wampold & Jenson (1986) and Hollon & Flick (1988) argued that discrete functional and dysfunctional distributions typically do not exist on symptom measures. Rather, the scores of both groups form a continuum and scores for the dysfunctional group occupy one tail of a single distribution. Accordingly, the derivation of cut-scores according to the Jacobson approach was deemed inappropriate (Wampold and Jenson, 1986). A further criticism was that even where such discrete distributions exist, variations in the mean level of dysfunctional severity between individual treatment studies would make comparison of their results difficult because each study would produce different estimates for the cut-points *a* and *c* (Hollon and Flick, 1988). A proposal to overcome these difficulties (Hollon and Flick, 1988) was that the dysfunctional population should be ignored and that clinical significance should be determined by assessing how much closer an individual's score has moved towards the general population mean following treatment. However, Hollon & Flick's recommendation to use unscreened and demographically representative normative samples represents a major threat to the notion that clinically significant change equates with a return to normal functioning, as such reference groups will contain individuals with notable levels of

psychopathology. Indeed, it has been estimated that up to 20% of the general population suffer from emotional disorders (Saunders et al., 1988). Another major limitation of Hollon & Flick's approach is that it provides no empirically based method capable of categorising whether a patient has entered the normative range as the degree of improvement required for clinical significance is arbitrary (Hollon and Flick, 1988). Jacobson & Revenstorf (1988) rejected Hollon & Flick's methodological criticisms by pointing out that within any distribution there are two distinct groups, (i) those who actively seek or receive treatment and (ii) those who do not. If such distinct groups exist, a cut off point could be established where there is equal probability of an individual being a member of either group.

The two-criterion approach of the Jacobson method has been criticised for being too conservative which leads to two problems. First, mildly symptomatic individuals with pre-treatment scores below the cut-off point can never make clinically significant change, only reliable improvement. Second, a severely symptomatic individual could demonstrate vast symptomatic improvement yet not reach the cut-off point. This means that they will be classified as having made reliable improvement rather than having made clinically significant change. In an attempt to address these issues, Tingey et al. (1996a) suggested using adjacent samples to distinguish between asymptomatic, mildly distressed, moderately distressed, and severely distressed levels of clinical significance. Whilst this would provide greater detail concerning client change, one limitation of Tingey et al.'s approach is that it requires more normative data than Jacobson et al.'s approach. However, a major limitation of Tingey et al.'s approach is that the typically poor validity of factors used to define adjacent samples means that it is less precise than the Jacobson method (Martinovich et al., 1996) and is likely to be clinically meaningless (Follette and Callaghan, 1996). That the Jacobson method is too conservative and might be abandoned by some researchers led Follette & Callaghan to propose that such researchers must be willing to state that " We have abandoned the goal of returning clients to normal functioning" (p140, Follette and Callaghan, 1996). However, it would appear that abandoning this goal is never justifiable as shown by the work of Lovaas who developed behavioural treatments for autistic children over a 30 year period (Lovaas, 1993). Following intensive treatment over 2 years, Lovaas showed that 47% of children achieved normal intellectual and educational functioning in contrast to 2% of controls.

There are several remaining methodological issues pertinent to the Jacobson approach. The Jacobson approach assumes that both functional and dysfunctional populations are normally distributed. However, many instruments used in clinical research have restricted ranges and therefore skewed distributions which may lead to errors in the calculation of cut-point *c*

(Jacobson and Revenstorf, 1988; Tingey et al., 1996b). The problem of skewed data is particularly evident when psychopathology measures are used in well functioning samples (Seggar et al., 2002; Martinovich et al., 1996). However, no empirical research has been conducted to examine to what degree this problem affects the precision of cut-off scores. Another issue that has concerned researchers is that the most recent formulation of the Jacobson approach (Jacobson et al., 1999) was not designed to control for regression to the mean which may reduce accuracy (Lambert and Ogles, 2009). Several researchers have modified the original formula for the RCI in order to account for this. For example Hsu (1989) modified the formula to include estimates of the mean and standard deviation of the population which scores would be expected to regress towards, whereas Speer (1992) used the reliability of the outcome measure to reduce pre-treatment scores toward the pre-treatment mean. However, Atkins et al. (2005) performed simulations that compared such modifications to the original RCI proposed by Jacobson et al. and found that there were no practical differences between their results when the reliability of outcome measures is high.

Researchers have also attempted to examine whether the four statistically defined Jacobson outcome categories are noticeably different according to patients (Ogles et al., 2001). For example, Ankuta & Abeles (1993) found that in a sample of outpatients with varied diagnoses, self-reported satisfaction following therapy was significantly higher in recovered than unchanged patients (as assessed on the SCL -90-R; Derogatis, 1983). These findings were later supported by Lunnen & Ogles (1998) who found that patients who made any reliable improvement (e.g. recovered or improved) demonstrated higher levels of perceived change and therapeutic alliance than unchanged or deteriorated patients. They concluded that the RCI was an effective index of symptomatic improvement but not deterioration (Lunnen and Ogles, 1998). Taken together, the results of these studies provided empirical evidence that both the recovered and improved categories of the Jacobson approach may be valid indicators of change that is meaningful to patients. However, this was not the case for deterioration, as Lunnen & Ogles (1998) found that unchanged and deteriorated patients were indistinguishable.

Finally, where multiple measures have been used to assess outcome, it has been found that recovery on one measure does not guarantee recovery on another. For example, Ogles et al. (1995) used the Jacobson approach to determine the clinical significance of outcomes in the Treatment of Depression Collaborative Research Program (TDCRP, Elkin et al., 1989). It was found that BDI and HRSD recovery rates showed considerable differences in completer

samples. For example, in the CBT group, the BDI recovery rate[11] was 28% whilst the HRSD rate was 45% (Ogles et al., 1995). Such differences could have arisen because some patients were already within the functional range on the BDI at the start of treatment. However, it is more likely that the rate difference between two measures assessing the severity of depression arose because each measure taps into different facets of the same construct (Jacobson and Revenstorf, 1988). Such findings indicate the desirability of reporting clinically significant outcomes for more than one measure and highlight the importance of developing a consensus on valid and appropriate measures for specific clinical populations. However, this objective has not been fully realised as researchers are free to use whichever measures they prefer and frequently fail to properly implement the Jacobson method (Ogles et al., 2001).

5.5 Summary & Concluding Remarks

The assessment of the clinical significance of treatment should utilise a methodology that is rigorous, objective and provides rigorous and non-ambiguous outcomes to providers and users of healthcare. The Jacobson approach fulfils this criterion and represents a meaningful and appropriate way of assessing change following treatment. However, the approach does have several limitations and each requires a resolution.

First, until adequate normative samples exist, there will be limitations of the applicability of the optimal Jacobson approach employing criterion 'c'. However, there are difficulties associated with obtaining normative data for many primary measures as they are not applicable to well functioning samples. Nevertheless, the Jacobson approach provides the alternative 'a' and 'b' criteria which, though not optimal, still provide an empirical basis by which to quantify clinical significance. Second, there has been relatively little empirical investigation into the validity of the four treatment outcome categories which can be derived from the Jacobson approach. For example, no studies appear to have investigated the concurrent validity of recovery according to the Jacobson method with diagnostic status following treatment. It would be reasonable to expect that recovered patients no longer meet diagnostic criteria and that unchanged patients to continue to do so. Certainly if there were no differences between these treatment outcome categories in terms of diagnostic status, then the clinical utility and validity of the Jacobson approach would be seriously undermined. Third, different outcome measures will result in somewhat different

[11] These rates were based on comparisons with screened normative samples. The BDI and HRSD results for comparisons with unscreened samples also showed considerable differences

proportions of patients being allocated to each of the treatment outcomes. In their reanalysis of TDCRP data, Ogles et al. (1995) showed that for the CBT group recovery rates according to the HRSD were approximately twice those for the BDI. These results show that the Jacobson approach will produce widely differing estimates of treatment efficacy, depending on the outcome measure employed. Thus, it is important that researchers reach consensus on which measure, or combination of measures, should be used to quantify clinically significant change.

Finally, the Jacobson approach has been criticised as being too stringent. Indeed, psychotherapy looks far less effective if clinical significance is used as the index of efficacy. For example, Jacobson, Wilson & Tupper (1988) found that although exposure treatments for agoraphobia were significantly better than control conditions, only 27% of clients achieved clinically significant change. This highlights the fundamental advantage of the Jacobson approach over standard inferential statistics; clinicians, researchers and patients will have an extremely clear idea as to whether a treatment works in terms of the probability that an individual receiving this treatment will make a return to normal functioning. A standardised approach to clinical significance provides a meaningful baseline by which to judge improvements in efficacy over time, thereby allowing healthcare providers and purchasers to determine whether novel pharmacological and psychological interventions represent a clinical advance. This form of benchmarking strategy allows efficient between and within study comparisons to be made, which is a fundamental component of effective evidence based practice.

Chapter Six

Study 2

Investigating Depression Treatment Outcomes Using the Jacobson Method of Clinical Significance

6.1 Introduction

Accurate estimates of treatment efficacy are fundamental to evidence-based medicine. However, the primary outcome studies that contributed to each meta-analysis included in the systematic review in study1 frequently used different methods to quantify treatment efficacy. These primary treatment studies operationalised treatment efficacy on the basis of the proportion of patients achieving remission according to post-treatment and follow-up scores on either a self report measure, the Beck Depression Inventory (BDI) or on a clinician rated measure, the Hamilton Rating Scale for Depression (HRSD). Pooling the results from such studies in meta-analysis is problematic. Because each measure assesses different aspects of depressive symptomatology, correlations between BDI and HRSD scores may be as low as .54 in depressed samples (Steer et al., 1987). Also, in patients assessed on both measures, symptomatic improvement is typically greater according to the HRSD than the BDI (Uher et al., 2008; Lambert et al., 1986). These factors raise the possibility that a proportion of patients categorised as remitted in studies using the HRSD would not be remitted according to the BDI and vice versa. Consequently, it is difficult to reach a balanced appraisal of the absolute and relative efficacy of interventions for major depression.

Even when studies used the same outcome measure to define remission between study variation was evident. For example, in the meta-analysis by de Maat et al. (2006), the stringency used to define remission on the HRSD in studies ranged from a score of less than 7 to a score of less than 10 points. As highlighted in study 1, the use of idiosyncratic definitions across studies is problematic because; (i) treatment differences observed between individual studies may be confounded with the definition of remission employed, (ii) published results of such studies may over or underestimate treatment efficacy. An alternative method of defining treatment efficacy in controlled evaluations of treatments for MDD is required. As detailed in the previous chapter, the Jacobson ‘c’ method (Jacobson

and Truax, 1991) provides a standardised approach that can be applied across studies to provide an index of both the relative and absolute merits of treatments.

The primary aim of this chapter was to use the Jacobson method to quantify recovery in published randomised controlled trials for the treatment of depression. A major advantage of the Jacobson 'c' method is that recovery rates are based on the proportions of patients who reliably return to the normative range on measures and are thus an estimate of absolute recovery. An additional advantage of the Jacobson method is that the identification of reliable symptomatic change enables patients to be allocated to one of four treatment outcomes; (i) recovered, (ii) improved, (iii) no change and (iv) worse. However, the Jacobson method required that individual patient data (IPD) was made available by study authors. Data were obtained for published studies where outcomes had been assessed using the BDI and/or HRSD. These measures were chosen as they were the most commonly used in the studies included in the meta-analyses in chapter 4.

There were two secondary aims to this study. Firstly, to compare the published clinical significance rates of studies with recovery as determined by the Jacobson method. It was hypothesised that both the relative and absolute published efficacies of treatments could differ markedly from those based on the Jacobson method. Secondly, to determine the level of agreement concerning recovery between the BDI and HRSD in samples that had been assessed on both measures. Given that Jacobson recovery represents a return to the normative range in terms of depressive symptomatology, it was hypothesised that there would be high levels of agreement between the BDI and the HRSD.

The results will firstly describe the studies that were used to assess the clinical significance of depression treatments according to the Jacobson method. Following this, the Jacobson clinical significance findings in studies are presented separately for the BDI and HRSD. Next, the published clinical significance results of individual studies are compared with the corresponding Jacobson method recovery rates. The final section of the results will examine the level of agreement for Jacobson recovery between the BDI and HRSD in the same patient sample.

6.2 Method

6.2.1 Search for Studies & Obtaining Individual Patient Data

The following sources were searched for references to studies that employed psychological treatments for depression:

- Reference sections of the reviews identified in study 1.
- A published database of 115 randomised controlled trials investigating psychological treatments for depression (Cuijpers et al., 2008).
- A database containing references to 149 controlled studies of psychotherapy for depression from the Free University of Amsterdam (Downloaded from http://www.psychotherapyrcts.org 19th November 2009).
- Electronic databases: SCOPUS, Web of Science, & OVID (final search 29th January 2010).
- References contained in Appendix 17b of, "Depression: the treatment and management of depression in adults." (NICE, 2009).

Studies were required to be methodologically similar to those included in the systematic review. Eligible studies had to meet the following criteria:

- Adult patients diagnosed with major depressive disorder (MDD) via structured clinical interviews according to DSM III, DSM III-R or DSM IV diagnostic criteria. Studies involving older adults, or studies treating depression in the context of substance abuse, personality disorder, psychotic or medical disorders were excluded.
- Face to face individual psychotherapy provided in at least one treatment condition with or without follow-up assessment. Preventative, maintenance, and therapies not based on a theoretical model of depression were excluded.
- Comparison conditions were treatment as usual, wait list control, attentional control, psychotherapy, pharmacotherapy, pill placebo.
- Studies were randomised controlled trials published in English from 1990.
- Depressive severity was assessed using the BDI and/or HRSD at both pre and post-treatment

A flow chart depicting the selection of studies is presented in Figure 2. The titles of all identified references were used to screen out articles that obviously did not meet the inclusion criteria. Following an examination of the abstracts for the 282 remaining references, the full text of 51 were obtained to determine eligibility. Figure 2 indicates the

reasons why 34 of the 51 studies were ineligible. The final decision concerning the eligibility of individual studies was reached following discussion with PF.

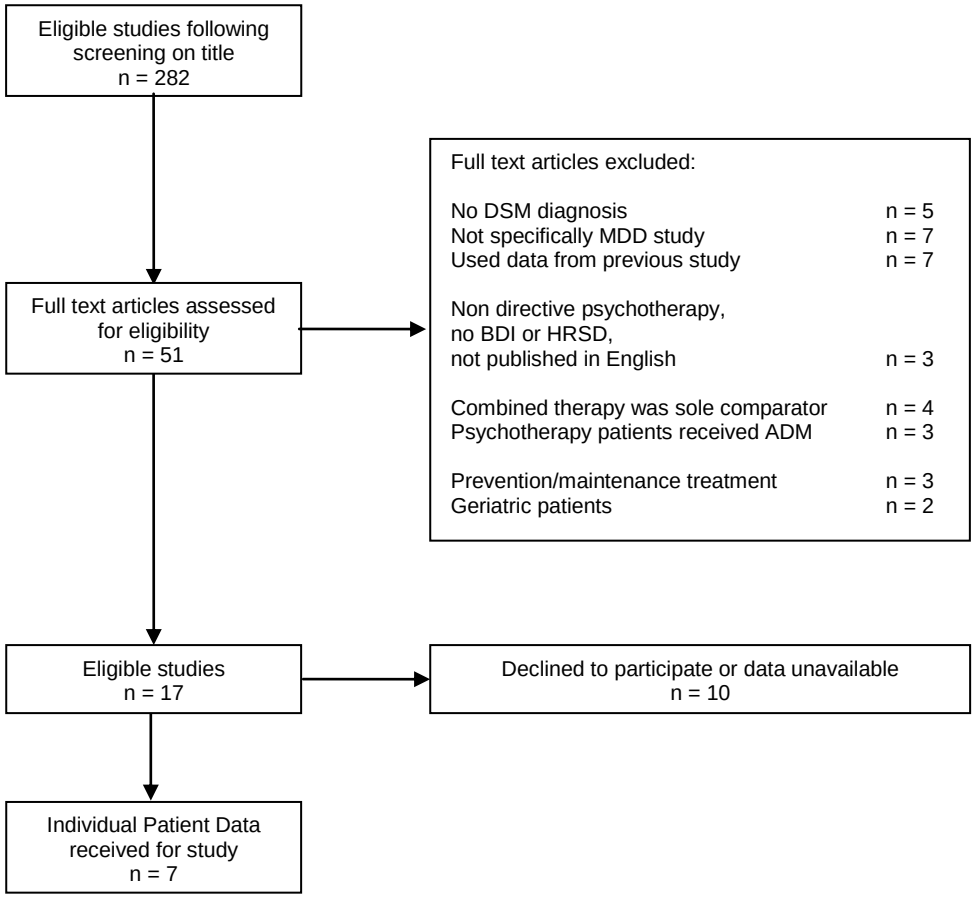


Figure 2. Identification of Eligible IPD Studies.

The authors of the 17 eligible studies were contacted via email in order to explain the purpose of the study and to request anonymous patient data. If authors did not reply after 2 weeks they were contacted two more times over the following 3 months. One author made no reply to requests for data, four did not wish to cooperate and four reported that original patient data was no longer available. One willing author was unable to participate due to the disruption caused by a large earthquake.

The individual patient data specifically requested from authors were:

- pre and post-treatment outcome data for the BDI (Beck et al., 1961) and/or HRSD (Hamilton, 1960)
- treatment type
- completer status
- number of treatment sessions
- number of previous episodes of depression
- concurrent dysthymia
- sex and age

IPD was obtained for the following 7 studies: Constantino et al. (2008), David et al. (2008), Dekker et al. (2008), DeRubeis et al. (2005), Jacobson et al. (1996), Jarrett et al. (1999), and Salminen et al. (2008).

6.2.2 Determining Jacobson Clinical Significance Criteria for the BDI & HRSD

Calculating clinical significance according to the Jacobson 'c' method, required that both the reliability and distribution of scores in non-depressed samples were known for measures. Consequently, electronic databases and review bibliographies were searched to obtain reliability and normative data for the BDI and HRSD. However, approximately 30% of U.S. adults have been shown to meet DSM IIIR criteria for psychiatric disorder in any year (Kessler et al., 1994). Thus, it is likely that the mean scores observed in the general population for both the BDI and HRSD will be elevated by a substantial minority of undiagnosed psychiatric 'cases'. In order to ensure that the Jacobson criteria for both measures represented psychiatric wellbeing, normative data were obtained only from studies where individuals had been screened for psychological disorder.

For the BDI a test-retest reliability of .81 (n = 74) was obtained from (Hatzenbuehler et al., 1983). The normative range for the BDI was obtained from the asymptomatic sample reported in Seggar et al. (2002) (mean = 2.88, SD = 2.44, n = 81). Whilst version one of the BDI was used in studies (Beck et al., 1961), two versions of the HRSD were used. Despite this, a single reliability and normative range were used to calculate the Jacobson clinical significance criteria for both HRSD versions. The reasons for this are described below.

Four studies (Dekker et al., 2008; Jacobson et al., 1996; Salminen et al., 2008; DeRubeis et al., 2005) used the 17-item version of the HRSD. However, DeRubeis et al. (2005) modified the original to make it sensitive to changes in atypical depressive symptomatology. These

minor modifications enabled *increases* in sleep, appetite or weight to be scored in contrast to the original version. Consequently, typically depressed patients would score the same on either version, whereas, atypical patients could score higher on the modified version (personal communication). It was assumed that the reliability of the modified version in DeRubeis et al. (2005) would not differ substantially from the original 17-item version as the modifications were minor. The 21-item version of the HRSD was used by Jarrett et al. (1999) which consists of the 17-item version plus four additional items (diurnal variation, derealisation & depersonalisation, paranoid symptoms and obsessional symptoms; Hamilton, 1960). Because Williams et al. (1988) found a difference of only .01 between the reliabilities of the 17- and 21-item versions in the same sample, it was assumed that the reliability for the 17-item version was appropriate for Jarrett et al.'s (1999) results. Consequently, a reliability of .85 for the 17-item version of the HRSD reported by Akdemir et al. (2001) was used to calculate Jacobson clinical significance criteria in all studies. This was the Pearson's correlation between independent raters in a sample of 93 depressed patients over a retest interval of 5 days (Akdemir et al., 2001).

The normative range for the HRSD was derived from a sub-set of control studies reviewed by Zimmerman et al. (2004a). Eight studies were identified where healthy controls had been screened for psychological disorder (Atmaca et al., 2002; Fassino et al., 2002; Grundy et al., 1996; Lanquillon et al., 2000; Rehm and O'Hara, 1985; Rubin et al., 2002; Wahby et al., 1990; Williams et al., 1991: cited Zimmerman et al., 2004a). However, the HRSD version in these studies was varied as 3 used the 17-item version, 3 used the 21-item version and one used a 24-item version. The version used in one study was unknown. It was decided to combine the results of the 8 studies as Zimmermann et al. (2004a) found no difference between the mean scores of controls in studies that used different versions of the HRSD. Consequently, the means and standard deviations of the 8 studies were weighted by sample size to produce a mean HRSD score of 2.80 and standard deviation of 1.60.

The distribution of scores seen in depressed samples for each measure was obtained from the pre-treatment scores of patients in the IPD studies. The data used to calculate the Jacobson's clinical significance criteria for both measures are summarised in Table 14.

Table 14. Data Used to Determine Jacobson Clinical Significance Criteria for the BDI & HRSD

Symbol	Definition	BDI	HRSD
M_1	Mean of depressed sample	28.52[a]	20.59[b]
S_1	Standard deviation of depressed sample	8.74	4.44
M_2	Mean of asymptomatic sample*	2.88[c]	2.80[d]
S_2	Standard deviation of asymptomatic sample*	2.44[c]	1.60[d]
r_{xx}	Reliability of scale	0.81[e]	0.85[f]
S_E	Standard error of measurement for scale	3.81	1.72
S_{diff}	Standard error of difference score	5.39	2.43

* Asymptomatic samples were screened to exclude psychiatric cases.
a Comprises all available pre-treatment scores (n = 499) from 5 IPD studies using the BDI.
b Comprises all available pre-treatment scores (n = 651) from 5 IPD studies using the HRSD.
c Seggar et al. (2002).
d Weighted result of 8 screened control studies (n = 399) in Zimmerman et al. (2004a).
e Hatzenbuehler et al. (1983).
f Akdemir et al. (2001).

Jacobson's 'c' for each measure was calculated using the formula:

$$c = \frac{S_1M_2 + S_2M_1}{S_1 + S_2}$$

According to the values in Table 14 Jacobson's 'c' for the BDI and HRSD were determined to be 8.48 and 7.51 respectively. Consequently, the respective cut-off points for recovery on the BDI and HRSD were deemed to be a score of 8 or less and 7 or less respectively.

The reliable change index (RCI) for each measure was calculated using the formulae:

$RCI = (X2 - X1)/S_{diff}$

where $S_{diff} = \sqrt{(2S_E^2)}$

and $S_E = S_1\sqrt{(1- r_{xx})}$

for the BDI	for the HRSD
$S_E = 8.74\sqrt{(1- 0.81)} = 3.809$	$S_E = 4.44\sqrt{(1- 0.85)} = 1.719$
$S_{diff} = \sqrt{2(3.8096)^2} = 5.387$	$S_{diff} = \sqrt{2(1.7196)^2} = 2.432$

An RCI greater than 1.96 is required for reliable change at the 5% level. Thus, the minimum reliable change in score from pre to post-treatment on the BDI was 11 points on the BDI and 5 points on the HRSD.

6.2.3 Data Analytic Strategy

Comparisons of mean pre-treatment severity between studies

Received IPD were used to calculate the mean pre-treatment severity of patients within each study for both the BDI and HRSD. Significant between-study differences in pre-treatment severity were then examined using separate analyses of variance (ANOVA) for both measures.

Determining Jacobson clinical significance rates

The clinical significance of individual patient outcomes was determined according to the Jacobson criteria for the BDI and/or HRSD. Only post-treatment data were analysed as insufficient data were provided for follow-up outcomes. One of four mutually exclusive clinical significance categories was assigned to each patient: (i) recovered, (ii) improved, (iii) no change, and (iv) worse. Where IPD was missing for patients, the clinically significant outcome assigned was 'no change'. Individual outcomes for each study were then used to calculate the percentage of patients in treatments occupying each of the four clinical significance categories. Whilst improvement is a clinically desirable outcome, it was not used to assess treatment efficacy. This was because the improved category in individual studies included unrecovered patients with an unknown range of symptom severity. Consequently, improvement rates were an unsuitable measure of clinical significance for both within- and between-study treatment comparisons. Jacobson recovery status was used to investigate the relative efficacy of treatments in each study via goodness of fit testing.

6.3 Results

6.3.1 Study Characteristics

All studies were based on outpatients attending for treatment. The majority of studies (5/7) used DSM IV criteria to diagnose MDD, the remaining two used DSM III criteria (Jacobson et al., 1996; Jarrett et al., 1999). Of the five studies that used the HRSD to assess depressive symptomatology, four reported that raters were blinded to patients' treatment condition. The exception was Salminen et al. (2008) who reported that HRSD raters were not blinded. Six of the seven studies reported that psychotherapy was manualised. Again, the exception was Salminen et al. (2008) who reported that STPP was not manualised. Only three studies clearly reported treatment adherence checks (David et al., 2008; Dekker et al., 2008; Jacobson et al., 1996). Selected characteristics of the 7 included studies are presented in Table 15. Studies varied considerably in terms of treatment type, number of patients, duration of psychotherapy, and the operational definitions of clinically significant outcomes.

Treatment groups were stratified on potential moderators of outcome during the randomisation process in four studies (David et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999).

Treatment comparisons

Three studies compared alternative types of psychotherapy (Constantino et al., 2008; David et al., 2008; Jacobson et al., 1996); two compared psychotherapy with antidepressant medication alone (Dekker et al., 2008; Salminen et al., 2008); and two included a pill placebo control arm in addition to ADM and psychotherapy (DeRubeis et al., 2005; Jarrett et al., 1996). The placebo condition in DeRubeis et al. (2005) was terminated after 8 weeks. The majority of studies used CBT where it was provided according to the principles outlined by Beck et al. (1979). However, Table 15 reveals that no study was a direct replication of any other. For example, whilst CBT was compared directly with ADM in 3 studies (David et al., 2008; DeRubeis et al., 2005; Jarrett et al., 1999) none used exactly the same type of medication. Two studies compared CBT with selective serotonin re-uptake inhibitors (David et al., 2008; DeRubeis et al., 2005), whilst Jarrett et al. (1999) used the monoamine oxidase inhibitor, Phenelzine.

Minimum intake severity

All studies specified a minimum level of depressive symptoms to be eligible for inclusion. However, studies used different measures and different levels of severity to define eligibility. For example, Table 15 shows that eligibility was assessed using only the BDI in a single study (Constantino et al., 2008) and that four used only the HRSD (Dekker et al., 2008; DeRubeis et al., 2005; Jarrett et al., 1999; Salminen et al., 2008). Two studies assessed intake severity using both the BDI and HRSD (David et al., 2008; Jacobson et al., 1996).

Table 15 shows that the minimum BDI severity score was 20 points in all studies using the BDI (Constantino et al., 2008; David et al., 2008; Jacobson et al., 1996). However, the minimum HRSD severity score differed across studies. Four studies used a minimum HRSD score of 14 (David et al., 2008; Dekker et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999). In contrast, DeRubeis et al. (2005) and Salminen et al. (2008) used a minimum HRSD score of 20 and 15 points respectively. Finally, only two studies examined the stability of depressive symptomatology prior to starting treatment (DeRubeis et al., 2005; Jarrett et al., 1999). DeRubeis et al. (2005) excluded patients who failed to meet the severity criterion at both screening and baseline assessments which were separated by at least 7 days. Jarrett et al. (1999) excluded those who initially met the severity criterion at the screening phase but failed to do so after a 14 day non-treated interval.

Number of sessions & duration of psychotherapy.

Psychotherapy sessions in all studies lasted between 50 to 60 minutes. However, there was considerable variation between studies in terms of the frequency of sessions. Table 15 shows that a full course of psychotherapy lasted for 20 sessions in four studies (David et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999) and for 16 sessions in two studies (Constantino et al., 2008; Salminen et al., 2008). The 8 sessions provided by Dekker et al. (2008) did not represent a full course of psychotherapy. Table 15 also shows that the duration of a full course of psychotherapy ranged from a minimum of 10 weeks (Jarrett et al., 1999) to a maximum of 16 weeks (DeRubeis et al., 2005; Jacobson et al., 1996; Salminen et al., 2008). Completer status was defined as receiving 12 sessions or more in David et al. (2008) and Jacobson et al. (1996) and 5 sessions in Dekker et al. (2008). The remaining studies did not define completer status in terms of sessions.

Table 15 reveals that duration or session data alone poorly represented studywise differences concerning the timing of therapeutic interventions. Some studies reported providing a fixed number of sessions per week for the duration of therapy (Dekker et al., 2008; Jarrett et al., 1999; Salminen et al., 2008), whereas others reported more frequent sessions at the beginning of treatment (Constantino et al., 2008; DeRubeis et al., 2005). Whilst no information was available concerning the timing of treatment in 2 studies (David et al., 2008; Jacobson et al., 1996), bi-weekly session were provide at some stage as both provided 20 sessions in less than 20 weeks.

Table 15. Characteristics of the Seven IPD Studies

Study	Treatment	Number Starting Treatment	Attrition (percent)	Minimum Intake Severity	Number of Sessions & Duration of Psychotherapy	Randomisation; Sample stratified on:	Criterion for Clinical Significance	IPD data received
Constantino et al. (2008)	CBT	11	27.3	BDI $\geq$ 20	16 sessions over 13 weeks	not reported.	BDI $\leq$ 15	BDI
	ICT	11	0		(6 bi-weekly then 10 weekly)		& reliable change[α]	
David et al. (2008)	Fluoxetine	55	14.0	HRSD $\geq$ 14	20 sessions over 14 weeks	no of previous episodes,	HRSD $\leq$ 6	BDI
	CBT	57	10.7	BDI $\geq$ 20	(frequency not reported)	sex, marital status,	& no MDD	
	REBT	44	8.8			dysthymia.		
Dekker et al. (2007)	Venlafaxine	44	4.5	HRSD $\geq$ 14	8 sessions[β] over 8 weeks	not reported.	none given	HRSD-17
	SPSP	59	8.5		(weekly)			
DeRubeis et al. (2005)	Paroxetine	120	15.8	HRSD $\geq$ 20	20 sessions over 16 weeks	no. of previous episodes,	HRSD $\leq$ 7	HRSD -17
	CBT	60	16.7		(8 bi-weekly then 12 weekly)	sex.		
	Placebo[γ]	60	n/a					
Jacobson et al. (1996)	AT	43	11.6	HRSD $\geq$ 14	20 sessions over 16 weeks	no. of previous episodes,	BDI $\leq$ 8	HRSD-17
	BA	56	12.5	BDI $\geq$ 20	(frequency not reported)	sex, marital status,	& no MDD	BDI
	CBT	50	6.0			dysthymia, severity.		
Jarrett et al. (1999)	Phenelzine	36	25.0	HRSD $\geq$ 14	20 sessions over 10 weeks	marital status,	HRSD $\leq$ 9	HRSD -21
	CBT	36	13.9		(bi-weekly)	length of current episode.		BDI
	Placebo	36	63.9					
Salminen et al. (2008)	Fluoxetine	25	24.0	HRSD $\geq$ 15	16 sessions over 16 weeks	not reported.	HRSD $\leq$ 7	HRSD -17
	STPP	26	19.2		(weekly)			BDI

Key: AT = Coping with Automatic thoughts; BA = Behavioural activation; BDI = Beck Depression Inventory; CBT = Cognitive behavioural therapy; HRSD = Hamilton Rating Scale for Depression; ICT = Integrative cognitive therapy; IPD = Individual patient data; n/a = not applicable.

α: Reliable change determined using the Jacobson method;
β: Interim outcomes: study was first 8 weeks of a longer study;
γ: Placebo condition terminated at 8 weeks;

Randomisation & sample stratification

No study adequately described the randomisation process according to CONSORT recommendations (Begg et al., 1996). The random allocation sequence used to assign patients to treatment condition was generated or implemented was not described by any study. Only Jarrett et al. (1999) reported that patient allocation was undertaken by an independent statistician. Table 15 shows that four studies stratified samples to balance groups on potential prognostic factors and that no study stratified samples using exactly the same set of factors (David et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999).

Published clinical significance criteria for post-treatment outcomes

Table 15 presents the published clinical significance criteria that formed the basis for comparisons between published and the Jacobson method post-treatment results. Only Dekker et al. (2008) did not use an index of clinical significance to compare treatments. The published criteria for Jarrett et al. (1999) correspond to the findings in their published abstract. Overall, studies demonstrated considerable variability in the definition of clinical significance; two compared treatments in terms of response (Constantino et al., 2008; Jarrett et al., 1999) and four compared treatments in terms of remission (David et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Salminen et al., 2008).

Constantino et al. (2008) used Jacobson and Truax's (1991) clinical significance method to quantify response rates according to the BDI. Response was defined as a score of 15 or less which included patients who were either 'non-distressed' (BDI = 0 to 9) or 'minimally distressed' (BDI = 10 to15) (Constantino et al., 2008). However, Constantino et al. provided limited information concerning the application of the Jacobson method to their data and did not report the minimum reliable score change required for the BDI. Response in Jarrett et al. (1999), was defined as a post-treatment HRSD score of 9 or less.

The definition of remission in DeRubeis et al. (2005) and Salminen et al. (2008) required only that patients meet a criterion score on the HRSD. However, DeRubeis et al. (2005) also employed an algorithm[12] to ensure that patients who demonstrated consistent remission during the final weeks of treatment were not excluded due to any transient symptom worsening at the final (post-treatment) assessment. Finally, Table 15 shows that the definition of remission in David et al. (2008) and Jacobson et al. (1996) required that patients no longer met diagnostic status in addition to meeting a criterion score on a symptom measure; the former study used the HRSD, the latter used the BDI.

[12] This was omitted from the table to simplify presentation.

Individual Patient Data received for outcome measures

Table 15 shows the measures for which IPD was received. All authors provided IPD for either the BDI and/or HRSD, treatment group, completer status and sex. Two studies provided outcome data for the BDI, (Constantino et al., 2008; David et al., 2008), two for the HRSD (Dekker et al., 2008; DeRubeis et al., 2005) and three provided outcome data for both measures (Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). As David et al. (2008) provided data for the BDI but published remission rates only for the HRSD, a direct comparison of published and Jacobson method results was not possible. Only one study (Salminen et al., 2008) failed to provide any last observation carried forward data (LOCF) for patients who dropped out of treatment. Two studies provided IPD for patients assessed at follow-up (Jacobson et al., 1996; Salminen et al., 2008). Both provided IPD for 12 month follow-up assessment. Six of the seven studies provided data for a complete course of treatment. Only interim results were obtained for Dekker et al. (2008) as the treatments provided after the first 8 weeks no longer met the inclusion criteria of this study (Dekker, personal communication). Finally, authors did not consistently provide IPD concerning the number of treatment sessions received, number of previous episodes of depression, dysthymia or age.

As described earlier in section 6.2.2 the 17-tem and 21-item versions of the HRSD were used in studies. Consequently, the maximum possible symptom score on the HRSD differed across studies. Three studies (Dekker et al., 2008; Jacobson et al., 1996; Salminen et al., 2008) used the 1967 version of the 17-item HRSD which provided a maximum score of 52 points (Hamilton, 1967). DeRubeis et al. (2005) used the 1960 version of the 17-item HRSD (Hamilton, 1960) which provided a maximum score of 50 points. The modifications for atypical depression made by DeRubeis et al. (2005) described in section 6.2.2 did not alter the maximum score from 50, but did allow atypical patients to score higher than on the 1960 version of the 17-item HRSD. Finally, Jarrett et al. (1999) used the 1960 version of the 21-item HRSD which consists of the 17-item version plus four additional items (Hamilton, 1960). Consequently, the maximum possible score for the HRSD version used by Jarrett et al. (1999) was 62 points. However, 3 of the items (derealisation & depersonalisation, paranoid symptoms and obsessional symptoms) are rarely endorsed by patients (Hamilton, 1967).

Pre-treatment severity differences between studies

Table 16 presents the mean pre-treatment scores of the total sample in studies according to received IPD. Only DeRubeis et al. (2005) published a pre-treatment mean score on the HRSD (mean 23.4, s.d.= 2.9) that differed to that in the table (mean 23.9, s.d.= 3.4). The

difference occurred because DeRubeis et al. (2005) provided IPD obtained during the screening phase but published results for the baseline assessment 1 week later (personal communication).

Significant differences between the mean pre-treatment scores in studies were identified via one way analyses of variance (ANOVA), for both the BDI ($F_{4, 494} = 11.09$, $p < .0001$) and HRSD ($F_{4, 646} = 88.42$, $p < .0001$). As post-hoc testing showed that the variances in BDI studies and HRSD studies were not homogeneous Tamahane tests were used to investigate significant differences between the pre-treatment means in studies (Tamahane: BDI: $F_{4, 494} = 16.42$, $p < .001$; HRSD: $F_{2,646} = 2.44$, $p = .046$).

Table 16. Total Sample Mean Pre-treatment Severity in IPD Studies*

Study	BDI	(s.d.)	HRSD	(s.d.)
Constantino et al. (2008)	29.1	5.6	-	-
David et al. (2008)	30.9	10.5	-	-
Dekker et al. (2007)	-	-	20.2	3.7
DeRubeis et al. (2005)	-	-	23.9	3.4
Jacobson et al. (1996)	29.4	6.6	18.5	4.1
Jarrett et al. (1999)	25.6	8.0	17.5	3.2
Salminen et al. (2008)	23.8	6.6	18.6	3.2
All studies	28.5	8.7	20.6	4.4

* Calculated using IPD.

Post hoc tests identified two groups of studies whose BDI pre-treatment mean scores did not significantly differ ($p < .05$). The first group consisted of Constantino et al. (2008), David et al. (2008) and Jacobson et al. (1996). The second group consisted of Jarrett et al. (1999) and Salminen et al. (2008). The BDI pre-treatment means of both David et al. (2008) and Jacobson et al. (1996) were significantly higher than those of both Jarrett et al. (1999) and Salminen et al. (2008). However, the BDI pre-treatment mean of Constantino et al. (2008) was only significantly higher than that of Salminen et al. (2008) and was no different to that of Jarrett et al. (1999).

In terms of the HRSD, post hoc tests ($p < .05$) revealed that the pre-treatment mean score in DeRubeis et al. (2005) was higher than that of all the remaining studies (Dekker et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). Pre-treatment means did

not differ between Jacobson et al. (1996), Jarrett et al. (1999) and Salminen et al. (2008). However, the pre-treatment mean in Dekker et al. (2008) was significantly higher than that of both Jacobson et al. (1996) and Jarrett et al. (1999) but no different to that of Salminen et al. (2008).

6.3.2 Jacobson Clinical Significance Rates for the BDI & HRSD in IPD Studies

Complete Jacobson clinical significance rates for each of the seven studies are presented in Table 17 and Table 18 for the BDI and HRSD respectively. With the exception of Dekker et al. (2008) the results are presented for post-treatment outcomes. Also, insufficient follow-up data was received to warrant an analysis and is not presented. The results for both ITT and completer samples revealed that the BDI did not categorise any patient as 'worse' in contrast to the HRSD. An examination of IPD for the completer sample in Jarrett et al. (1999) confirmed that 2 patients who were unchanged on the BDI were worse according to the HRSD. The tables show that, overall, the majority of patients receiving an active treatment were categorised as improved or recovered following a full course of treatment for both measures and sample type. Patient attrition typically led to higher recovery rates in completer samples than in ITT samples. However, an analysis of IPD indicated that attrition was not confined to those patients who were unchanged or became worse during treatment.

BDI rates

In terms of ITT samples, BDI results in Table 17 show that the percentage recovery rate for active treatments varied widely between studies. Recovery rates ranged from only 27.3% for ICT in Constantino et al. (2008) to 62% for CBT in Jacobson et al. (1996). However, Table 17 also shows that in most studies the majority of patients in active treatments were typically improved or recovered at post-treatment. Salminen et al. (2008) was the only exception where 52% of ADM and 53.8% of STPP patients were unchanged (i.e. 'no change'). In addition, active treatments were associated with greater overall improvement (i.e. improved or recovered) than controls, as placebos in Jarrett et al. (1999) demonstrated both the lowest recovery rate (22%) and highest rate of 'no change' (66.7%). Finally, recovery across all active treatments was 48.1% .

In terms of completer samples, an examination of Table 17 shows that recovery rates for active treatments demonstrated a similar range to those observed in ITT samples. Recovery rates ranged from only 25% for CBT in Constantino et al. (2008) to 63.8% for CBT Jacobson et al. (1996). However, with the exception of the CBT group in Constantino et al. (2008), recovery in completer samples was higher than that in ITT samples. In contrast to

the ITT sample, the majority of patients had either improved or recovered in all active treatment groups as the percentage of 'no change' in the ADM and STPP groups in Salminen et al. (2008) were now 36.8% and 42.9% respectively. The 63.9% attrition rate for placebos in Jarrett et al. (1999) resulted in a 61.5% recovery rate that was higher than that of the CBT group . Recovery across all active treatments was 53.0%.

HRSD rates

Comparing the results of Dekker et al. (2008) with the remaining studies in Table 18 is problematic as they corresponded only to 8-week outcomes. Consequently, they are not considered here as post-treatment results.

In terms of ITT samples, the results for active treatments in Table 18 show that the percentage range of post-treatment recovery varied widely between studies. Post-treatment recovery ranged from 38.3% for CBT in DeRubeis et al. (2006) to 66.0% for CBT in Jacobson et al. (1996). As seen for the BDI, the majority of patients receiving an active treatment were either improved or recovered at post-treatment. However, unlike the BDI, the results in Table 18 show that the HRSD categorised a small percentage of patients as 'worse' at post-treatment (DeRubeis et al., 2005; Jarrett et al., 1999) and 8 weeks (Dekker et al., 2008). The highest overall rate of worsening was 11.9% for the SPSP group in Dekker et al. (2008) whilst the highest post-treatment rate was 8.3% for both the CBT and placebo groups in Jarrett et al. (1999). In common with BDI results, active treatments were associated with greater overall post-treatment improvement than controls, as placebos in Jarrett et al. (1999) demonstrated both the lowest recovery rate (22%) and highest rate of 'no change' (55.6%). However, the 8 week recovery rates for ADM (11.4%) and SPSP (8.5%) in Dekker et al. (2008) were substantially lower than the post-treatment rate for placebos in Jarrett et al. (1999). Finally, recovery across all active treatments was 50.7%.

Table 17. Percentage of Patients Allocated to four Categories of Clinical Significance using Jacobson Criteria for the BDI at Post-treatment.

Study	Treatment	Intention to Treat Analysis (ITT)					Completer Analysis				
		n	Worse	No Change	Improved	Recovered	n	Worse	No Change	Improved	Recovered
Constantino et al. (2008)	CBT	11	-	36.4	27.3	36.4	8	-	37.5	37.5	25.0
	ICT	11	-	9.1	63.6	27.3	11	-	9.1	63.6	27.3
David et al. (2008)	ADM	57	-	19.3	35.1	45.6	49	-	14.3	34.7	51.0
	CBT	56	-	23.2	30.4	46.4	50	-	20.0	30.0	50.0
	REBT	57	-	15.8	42.1	42.1	52	-	15.4	40.4	44.2
Jacobson et al. (1996)	AT	43	-	25.6	18.6	55.8	38	-	15.8	21.1	63.2
	BA	56	-	16.1	26.8	57.1	49	-	16.3	22.4	61.2
	CBT	50	-	20.0	18.0	62.0	47	-	17.0	19.1	63.8
Jarrett et al. (1999)	ADM	36	-	36.1	13.9	50	27	-	18.5	18.5	63.0
	CBT	36	-	36.1	16.7	47.2	31	-	32.3	16.1	51.6
	Placebo	36	-	66.7	11.1	22.2	13	-	23.1	15.4	61.5
Salminen et al. (2008)	ADM	25	-	52.0	16.0	32.0	19	-	36.8	21.1	42.1
	STPP	26	-	53.8	7.7	38.5	21	-	42.9	9.5	47.6
All Active Treatments*		464	-	26.1	25.9	48.1	402	-	20.4	26.6	53.0

Key: AT = Coping with automatic thoughts; BA = Behavioural activation; BDI = Beck Depression Inventory; CBT = Cognitive Behavioural Therapy; ICT = Integrative Cognitive Therapy; REBT = Rational Emotive Behavioural Therapy; STPP = Short Term Psychodynamic Psychotherapy.

* excluding placebo

Table 18. Percentage of Patients Allocated to four Categories of Clinical Significance using Jacobson Criteria for the HRSD at Post-treatment.

Study	Treatment	Intention to Treat Analysis (ITT)					Completer Analysis				
		n	Worse	No Change	Improved	Recovered	n	Worse	No Change	Improved	Recovered
DeRubeis et al. (2005)	ADM	120	1.7	16.7	37.5	44.2	101	-	9.9	37.6	52.5
	CBT	60	1.7	13.3	46.7	38.3	50	-	10.0	46.0	44.0
Jacobson et al. (1996)	AT	43	-	18.6	18.6	62.8	38	-	10.5	18.4	71.1
	BA	56	-	17.9	23.2	58.9	49	-	16.3	24.5	59.2
	CBT	50	-	20.0	14.0	66.0	47	-	14.9	14.9	70.2
Jarrett et al. (1999)	ADM	36	-	27.8	22.2	50.0	27	-	14.8	18.5	66.7
	CBT	36	8.3	22.2	22.2	47.2	31	6.5	19.4	22.6	51.6
	Placebo	36	8.3	55.6	13.9	22.2	13	-	15.4	23.1	61.5
Salminen et al. (2008)	ADM	25	-	40.0	8.0	52.0	19	-	21.1	10.5	68.4
	STPP	26	-	34.6	19.2	46.2	21	-	19.0	23.8	57.1
Dekker et al. (2008)$^{\alpha}$	ADM	44	6.8	52.3	29.5	11.4	42	4.8	52.4	31.0	11.9
	SPSP	59	11.9	59.3	20.3	8.5	54	13.0	55.6	22.2	9.3
All Active Treatments*		452	1.3	20.6	27.4	50.7	383	0.5	13.6	27.7	58.2

Key: AT = Coping with automatic thoughts; BA = Behavioural activation; CBT = Cognitive Behavioural Therapy; HRSD = Hamilton Rating Scale for Depression; SPSP = Short Psychodynamic Supportive Psychotherapy; STPP = Short Term Psychodynamic Psychotherapy.

α: 8 week results.

* excluding placebos & 8 week results for Dekker et al. (2008).

In terms of completer samples, the HRSD results in Table 18 show that post-treatment recovery for active treatments demonstrated a similar range to those observed in ITT samples. Post-treatment recovery rates ranged from 44.0% for CBT in DeRubeis et al. (2005) to 71.1% for AT in Jacobson et al. (1996). An examination of the results in Table 18 shows that post-treatment recovery in completer samples was higher than in ITT samples and that the proportion of patients categorised as 'worse' was reduced. Again, the majority of patients in active treatments were either improved or recovered at post-treatment. In contrast to the results for post-treatment outcomes the majority of ADM and SPSP patients in Dekker et al. (2008) were unchanged at 8 weeks. As seen for the BDI, the 63.9% attrition rate for placebos in Jarrett et al. (1999) resulted in a higher recovery rate in placebos (61.5%) compared to CBT (51.6%). Recovery across all active treatments was 58.2%.

Post-treatment Jacobson recovery rates by treatment class

Table 17 and Table 18 reveal that post-treatment recovery rates for similar classes of active treatment varied considerably between studies in both ITT and completer samples. The 8 week results for Dekker et al. (2008) are not described. Psychotherapy type was classed as either CBT according to Beck et al. (1979), or non-CBT, whilst all medications were classed as ADM.

BDI recovery

In terms of ITT samples, Table 17 shows that ADM recovery on the BDI ranged from 32% in Salminen et al. (2008) to 50% in Jarrett et al., 1999. CBT recovery ranged from 36.4% in Constantino et al. (2008) to 62% in Jacobson et al. (1996). Recovery for non-CBT psychotherapies ranged from 27.3% for the ICT group in Constantino et al. (2008) to 57.1% for the BA group in Jacobson et al., (1996). The overall rates for ADM and all types of psychotherapy across studies were 44.1% and 49.4% respectively.

In terms of completer samples, ADM recovery on the BDI ranged from 42.1% in Salminen et al. (2008) to 63% in Jarrett et al. (1999). CBT recovery ranged from 25.0% in Constantino et al. (2008) to 63.8% in Jacobson et al. (1996). The CBT completer rate in Constantino et al. (2008) was lower than the ITT rate because 2 recovered CBT patients failed to complete treatment. Recovery for non-CBT psychotherapies ranged from 27.3% for ICT in Constantino et al. (2008) to 63.2% for AT in Jacobson et al. (1996). The overall rates for ADM and all types of psychotherapy across studies were 52.6% and 53.1% respectively.

HRSD recovery

In the ITT samples, Table 18 shows that ADM recovery on the HRSD ranged from 44.2% in DeRubeis et al. (2005) to 52% in Salminen et al. (2008). CBT recovery ranged from 38.3% in DeRubeis et al. (2005) to 66% in Jacobson et al. (1996). Recovery for non-CBT psychotherapies ranged from 46.2% for STPP in Salminen et al. (2008) to 62.8% for AT in Jacobson et al, (1996). The overall rates for ADM and all types of psychotherapy across studies were 46.4% and 53.5% respectively.

In the completer samples, ADM recovery ranged from 52.5% in DeRubeis et al. (2005) to 68.4% in Salminen et al. (2008). CBT recovery ranged from 44.0% in DeRubeis et al. (2005) to 70.2% in Jacobson et al. (1996). Recovery in non-CBT psychotherapies ranged from 57.1% for STPP in Salminen et al., (2008) to 71.1% for AT in Jacobson et al. (1996). The overall rates for ADM and all types of psychotherapy across studies were 57.1% and 58.9% respectively.

Summary

The results for both measures and sample type indicate that the range of post-treatment recovery rates for psychotherapy was typically greater than for ADM. In addition, overall ITT psychotherapy recovery rates were 5.3% higher than ADM rates on the BDI and 7.1% higher on the HRSD. However, not all studies contributing to these overall figures had directly compared ADM with psychotherapy (Constantino et al., 2008; Jacobson et al.; 1996). For the pooled ITT sample of studies that directly compared ADM with psychotherapy, the overall BDI recovery rates for ADM and psychotherapy were 44.1% and 44% (David et al., 2008; Jarrett et al., 1999; Salminen et al., 2008) respectively. The corresponding HRSD rates were 46.4% and 42.6% respectively (DeRubeis et al., 2005; Jarrett et al., 1999; Salminen et al., 2008). For the pooled completer sample, the overall BDI recovery rates for ADM and psychotherapy were 52.6% and 48.1% respectively (David et al., 2008; Jarrett et al., 1999; Salminen et al., 2008). The corresponding HRSD rates were 57.1% and 49% respectively (DeRubeis et al., 2005; Jarrett et al., 1999; Salminen et al., 2008). Thus, with the exception of the ITT sample results for the BDI, ADM appeared to produce higher overall recovery rates than psychotherapy in direct comparisons. However, goodness of fit testing for the results of individual studies failed to show that any active treatment was superior to another in any sample and on either measure. Nevertheless, goodness of fit tests showed that both CBT and ADM in Jarrett et al. (1999) were superior to placebo on both measures in ITT samples ($X^2_{(2)} = 7.03$, $p = .03$ for both measures). Active

treatments in Jarrett et al. (1999) were no different to placebo in completer samples on both measures.

Patient attrition In clinical significance categories

The overall attrition rate for the 747 patients in active treatments across studies was 13.7%. In studies that compared psychotherapy directly with ADM the attrition was 15.6% for ADM and 11.6% for psychotherapy ($p > .1$). An analysis of pooled LOCF [13] data and IPD completer status for the BDI showed that 54.9% of dropouts in active treatments were unchanged, 25.5% improved and 19.6% recovered at their last observation (n = 51). The same analysis for the HRSD showed that 55.4% of dropouts were unchanged, 7.7% were worse, 27.7% improved and 9.2% recovered at their last observation (n = 65).

6.3.3 Comparing Published Clinical Significance Rates with Jacobson Recovery

Table 19 presents the post-treatment clinical significance rates reported in studies along with the criteria on which they were based. The corresponding Jacobson method rates are presented for comparison. These were based on the same outcome measure as used for published results with the exception of David et al. (2008). An analysis of IPD showed that unreliable change had not contributed to published results in 5 studies (Constantino et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). The minimum published HRSD entry criterion of 14 points in David et al. (2008) indicated that patients scoring 6 or less had demonstrated reliable change. A comparison of published and Jacobson results revealed that (i) published and Jacobson rates were identical in only one study, (ii) published rates were higher where published criterion scores on measures were less stringent than Jacobson score criteria, (iii) the use of additional outcome criteria in studies that used identical score criteria to those of the Jacobson method produced rates that differed substantially from Jacobson rates.

Table 19 shows that the published clinical significance criteria closely approximated those of the Jacobson method in two studies where the HRSD score criterion was identical to that of the Jacobson method (DeRubeis et al., 2005; Salminen et al., 2008). Consequently, the published completer sample remission rates in Salminen et al. (2008) were identical to Jacobson recovery rates for both ADM and STPP at 68% and 57% respectively. However, DeRubeis et al. (2005) included a small number of patients with an HRSD score greater than 7 as remitted in their results. Consequently, the published remission rates of 46% for ADM and 40% for CBT were both 2% higher than the corresponding Jacobson recovery rates.

[13] Salminen et al. (2008) was excluded as no LOCF data was available

Table 19. Post-treatment Clinical Significance Rates: A Comparison of Results based on Published & Jacobson Method Criteria.

Study	Treatment	n	Published criterion	Sample	Published (%)	Jacobson (%)
Constantino et al. (2008)	CBT	11	BDI ≤ 15	ITT	55	36
	ICT	11			82	27
David et al. (2008)	ADM	57	HRSD ≤ 6 & no MDD	ITT	50	46[α]
	CBT	56			50	46[α]
	REBT	57			45	42[α]
DeRubeis et al. (2005)	ADM	120	HRSD ≤ 7	ITT	46	44
	CBT	60			40	38
Jacobson et al. (1996)	AT	43	BDI ≤ 8 & no MDD	ITT	51	56
	BA	56			46	57
	CBT	50			56	62
Jarrett et al. (1999)	ADM	36	HRSD ≤ 9	ITT	58	50
	CBT	36			58	47
	Placebo	36			28	22
Salminen et al. (2008)	ADM	19	HRSD ≤ 7	Completer	68	68
	STPP	21			57	57

Key: AT = Coping with automatic thoughts; BA = Behavioural activation; BDI = Beck Depression Inventory; CBT = Cognitive Behavioural Therapy; HRSD = Hamilton Rating Scale for Depression; ICT = Integrative cognitive therapy; REBT = Rational Emotive Behavioural Therapy; STPP = Short Term Psychodynamic Psychotherapy.

α: BDI data only

The use of idiosyncratic criteria in the remaining studies could lead to published clinical significance rates that (i) differed markedly from Jacobson method rates (ii) changed the observed relative efficacy of treatments. For example, the published response criterion (BDI ≤ 15) in Constantino et al. (2008) was far less stringent than the Jacobson method (BDI ≤ 8). This produced published clinical significance rates of 55% for CBT and 82% for ICT which contrasted greatly with the corresponding Jacobson rates of 36% and 27% respectively. Consequently, the published advantage for ICT was reversed in favour of CBT according to Jacobson rates. This was the most extreme example of the difference between published and Jacobson rates. However, Table 19 shows that substantial differences also occurred in Jacobson et al. (1996) and Jarrett et al. (1999).

The BDI score criterion in Jacobson et al. (1996) was identical to that of the Jacobson method (BDI $\leq$ 8). Despite this, the requirement that patients were no longer depressed resulted in Jacobson recovery rates that were at least 5% higher than published. In addition Table 19 reveals that the differences between published and Jacobson rates were not consistent between treatments. For example, the Jacobson recovery rates for the AT and CBT groups were 5% and 6% higher than published rates respectively. This contrasted with the BA group where Jacobson rates were 11% higher than published. Consequently, the Jacobson method identified that BA was more efficacious than AT in contrast to published results.

In Jarrett et al. (1999), the criterion for response only required that patients score less than 9 on the HRSD. However, the use of a less stringent criterion than that of the Jacobson method produced published rates that were between 6% and 11% higher than the corresponding Jacobson rates. For example, the published rates for placebo and CBT were 28% and 58% respectively whilst the corresponding Jacobson rates were 22% and 47%. In addition, the Jacobson method revealed that more patients recovered in the ADM group (50%) compared to CBT (47%). This contrasted with published results where response rates were identical. Finally, a comparison of results for David et al. (2008) shows that published remission rates were consistently higher than Jacobson recovery rates. However, it was not possible to determine the reason for this as Jacobson method and published rates were based on different measures.

6.3.4 Jacobson Recovery: Agreement Between Measures in the Same Sample

An examination of the results in Table 17 and Table 18 for the 3 studies that used both the BDI and HRSD showed that (i) only the HRSD categorised any patient as worse, (ii) HRSD recovery rates were always higher or equal to BDI recovery rates in ITT samples. However, because BDI and HRSD data for drop-outs may have been obtained at different points during treatment, the ITT results in Table 17 and Table 18 were unsuitable for comparing the level of agreement between measures. Consequently, agreement between the BDI and HRSD for recovery was examined using the completer[14] samples of Jacobson et al. (1996), Jarrett et al. (1999) and Salminen et al. (2008). Table 20 presents a breakdown of the overall BDI and HRSD recovery rates for groups in studies in terms of the percentage who recovered according to (i) both the BDI and HRSD, (ii) the BDI only, (iii) the HRSD only. In Table 20 the 'BDI & HRSD' recovery rate (the 'dual rate') for each group denotes the

[14] All patients were assessed on both measures at pre- and post-treatment.

percentage of patients who recovered on both the BDI and HRSD. The 'BDI only' and 'HRSD only' rates denote the percentage of patients who recovered only on the BDI or HRSD respectively.

Table 20. Overall Percentage Recovery Rates on the BDI or HRSD: Comparisons with Recovery on Both Measures, the BDI Alone and the HRSD Alone.

Study	Treatment		Recovered on:			Overall recovery:	
		n	BDI & HRSD*	BDI only	HRSD only	BDI	HRSD
Jacobson et al. (1996)	AT	37	59.5	2.7	13.5	62.2	73.0
	BA	48	47.9	12.5	12.5	60.4	60.4
	CBT	47	55.3	8.5	14.9	63.8	70.2
Jarrett et al. (1999)	ADM	27	51.9	11.1	14.8	63.0	66.7
	CBT	31	41.9	9.7	9.7	51.6	51.6
	placebo	13	53.8	7.7	7.7	61.5	61.5
Salminen et al. (2008)	ADM	19	42.1	0	26.3	42.1	68.4
	STPP	21	47.6	0	9.5	47.6	57.1
Pooled total		243	50.6	7.4	13.6	58.0	64.2

* 'dual recovery'

The pooled results across all treatments show that the overall recovery rate for the HRSD was 6.2% higher than the overall rate for the BDI (64.2% versus 58% respectively). However, the finding that only 50.6% recovered according to both measures, that 7.4% recovered only on the BDI and that 13.6% recovered only on the HRSD suggested that agreement between measures was low. In addition, within some individual studies the agreement between overall BDI and HRSD rates was poor for some treatments. For example, in Jacobson et al. (1996) overall CBT recovery according to the BDI was 63.8%, whereas on the HRSD it was 70.2%. The poor agreement between overall BDI and HRSD rates meant that the rank order of treatments in Jacobson et al. (1996) and Salminen et al. (2008) differed according to the BDI and HRSD.

An inspection of the results in Table 20 for each group shows that the overall recovery rate for a specific measure equalled the dual rate plus the proportion of patients who recovered only on that measure. The table reveals that the identical overall BDI and HRSD rates in 3

of the 8 groups were the same only because the same percentage of patients recovered only on the BDI as only recovered on the HRSD (BA in Jacobson et al., 1996; CBT and placebo in Jarrett et al., 1999). Thus, the apparently perfect agreement between overall BDI and HRSD rates for these groups obscured the fact that 25% of the BA group in Jacobson et al. (1996) and 19.4% of the CBT and 15.4% of the placebo groups in Jarrett et al. (1999) recovered on only a single measure.

In the remaining 5 groups the percentage recovered only on the HRSD was higher than the percentage recovered only on the BDI. Consequently, the overall HRSD rate for each group was higher than the overall BDI rate. The magnitude of the difference between overall HRSD and overall BDI rates was equal to the 'HRSD only' rate minus the 'BDI only' rate. An examination of the table reveals that the difference between the 'HRSD only' and 'BDI only' rates in each of the 5 groups ranged from 3.7% for ADM in Jarrett et al. (1999) to 26.3% for ADM in Salminen et al. (2008). This variability produced variable agreement between overall BDI and overall HRSD rates in the 5 groups and was partly responsible for the BDI and HRSD ranking treatments differently in Jacobson et al. (1996) and Salminen et al. (2008).

Table 21. Pooled Comparison of Recovery Status according to the BDI or HRSD for Completers Assessed on Both Measures (n).

		HRSD (n)		
		Recovered	Unrecovered	Total
BDI (n)	Recovered	123	18	141
	Unrecovered	33	69	102
	Total	156	87	243

In order to quantify agreement concerning Jacobson recovery between the BDI and HRSD, the number of patients recovering and not recovering on measures was tabulated. Table 21 shows the number of patients across studies who recovered according to both the BDI and HRSD (dual recovery), the BDI only and the HRSD only. The table shows that 123 (50.6%) were recovered and 69 (28.4%) were unrecovered according to both measures. Thus, agreement in simple percentage terms between measures was 79%. However, the chance adjusted agreement rate between the BDI and HRSD was only 56% (Kappa = .56, p <.001).

6.4 Discussion

6.4.1 Treatment Efficacy According to the Jacobson Method

The primary aim of this study was to use a standard recovery definition to quantify the efficacy of psychological treatments for depression. Of the seven studies that provided IPD for the present study, only one (Dekker et al., 2008) provided data that was unsuitable. By classifying patients as recovered, improved, unchanged or worse, the Jacobson method enabled a clearer understanding of the current level of treatment efficacy than previously existed. The overall ITT recovery rates across BDI and HRSD studies of 48.1% and 50.7% respectively revealed that approximately 50% of patients entering active treatments did not recover. However, ITT recovery rates in different studies were highly variable and ranged from 27.3% (ICT in Constantino et al., 2008) to 62% (CBT in Jacobson et al., 1996) on the BDI, and from 38.3% (CBT in DeRubeis et al. 2005) to 66% (CBT in Jacobson et al., 1996) on the HRSD.

An analysis of IPD showed that the overall attrition rate for active treatments across studies was 13.7% and that the majority of drop-outs were unchanged on measures. This suggested that patients who did not benefit from treatment were more likely to drop-out than those who experienced an improvement in symptoms. Whilst this could not be investigated using the IPD provided by authors, it was found that post-treatment recovery rates in completer samples were typically higher than in ITT samples. Overall completer recovery for active treatments was 53% in BDI studies and 58.2% in HRSD studies. These results show that the provision of a full course of treatment left a large proportion of patients unrecovered. As seen for ITT samples, completer recovery rates in individual studies were highly variable and ranged from 25% (CBT in Constantino et al., 2008) to 63.8% (CBT in Jacobson et al., 1996) on the BDI and from 44% (CBT in DeRubeis et al., 2005) to 71.1% (AT in Jacobson et al., 1996) on the HRSD. However, no significant differences were found on either measure between the active treatments within studies in both ITT and completer samples.

It is important to consider the efficacy rates reported here in light of research which demonstrates that that approximately 20% of mildly[15] depressed wait-list control patients will spontaneously recover over a 4 to 8 week period (Posternak and Miller, 2001). The placebo recovery rate in Jarrett et al.'s (1999) ITT sample did show that 22% of those receiving a sham treatment recovered over 10 weeks. This suggested that the active components of ADM and psychotherapy increased recovery over placebo in ITT samples by 28% and 25% respectively. Unfortunately, the lack of non-treatment control data for

[15] Pre-treatment BDI or HRSD score of 20 or less

remaining studies meant it was not possible to estimate the benefit of treatments over placebo nor to definitively confirm that treatments were more efficacious than placebo (Klein, 1996) .

6.4.2 Comparisons of Published & Jacobson Method Clinical Significance Rates

The published clinical significance rates of studies were compared with Jacobson method rates for the corresponding outcome measure. The one exception was David et al. (2008) where published rates based on the HRSD were compared with Jacobson rates based on the BDI. Comparisons showed that published rates were consistently higher than Jacobson rates in 4 studies (Constantino et al., 2008; David et al., 2008; DeRubeis et al., 2005; Jarrett et al., 1999) and consistently lower in one (Jacobson et al., 1996). Only Salminen et al. (2008) published clinical significance rates that were identical to Jacobson method rates. The rank order of treatment efficacy differed according to published and Jacobson criteria in 3 studies (Constantino et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999). These results revealed that the use of idiosyncratic clinical significance criteria typically overestimate recovery and prohibit a meaningful comparison of study outcomes.

An analysis of IPD showed that every patient who achieved a clinically significant outcome according to published criteria had changed reliably on measures. This meant that any disagreement between published and Jacobson method rates was due to differences in the definition of a clinically significant outcome. The published and Jacobson method recovery rates in Salminen et al. (2008) were identical because published and Jacobson criteria were effectively the same. However, two studies using the same score criterion as the Jacobson method published different results to those of the Jacobson method (DeRubeis et al., 2005; Jacobson et al., 1996). The disagreement between published and Jacobson method rates in DeRubeis et al. (2005) and Jacobson et al. (1996) occurred because additional published recovery criteria were employed. DeRubeis et al. (2005) published recovery rates that were 2% higher than the Jacobson rates for ADM and CBT because they included patients who had consistently met the HRSD recovery criterion prior to post-treatment assessment. Because DeRubeis et al.'s published rates were based on multiple assessments it may be argued that they were more accurate than Jacobson rates which were calculated using only pre- and post-treatment data. However, the consistent 2% difference between published and Jacobson rates suggested that unreliable symptomatic worsening is relatively rare and unlikely to bias conclusions based on the Jacobson method. Jacobson et al. (1996) used the same BDI score criterion as the Jacobson method but published lower recovery rates due to an additional requirement that patients were no longer depressed according to DSM III criteria. The published recovery rates for individual treatments in Jacobson et al. (1996)

were between 5% to 11% lower than Jacobson method rates. This meant that 12.6% of all patients that recovered according to the Jacobson method in this study were still depressed. This indicated that the BDI provided insufficient coverage of the symptom criteria required for a DSM III diagnosis of MDD.

The published score criteria used to define clinical significance in the three remaining studies differed from those of the Jacobson method (Constantino et al., 2008; David et al., 2008; Jarrett et al., 1999). The published HRSD recovery criteria in David et al. (2008) were more stringent than the Jacobson method HRSD criteria because patients were required to score 1 point lower on the HRSD and no longer be depressed according to DSM IV diagnostic criteria. The finding that David et al.'s published rates based on the HRSD were at least 3% higher than Jacobson rates based on the BDI suggested that Jacobson method BDI rates were more stringent than published criteria. However, the unavailability of HRSD data for patients in David et al. (2008) meant that the reason for this could not be explored further. In Constantino et al. (2008) and Jarrett et al. (1999) published comparisons were based on different definitions of treatment response. Constantino et al. (2008) defined response as a score of 15 or less on the BDI[16], whereas Jarrett et al. (1999) employed a score of 9 or less on the HRSD. Therefore, the published criteria employed in both of these studies were less stringent than the corresponding Jacobson method criteria. Consequently, published rates for the groups in Constantino et al. (2008) and Jarrett et al. (1999) could exceed Jacobson rates by as much as 55% and 11% respectively.

6.4.3 Agreement Between the BDI & HRSD

The BDI and HRSD are two of the most widely used outcome measures used in depression treatment studies. Consequently, the degree to which these measures agree concerning depressive symptomatology is of fundamental importance where study results are compared or combined in meta-analysis. A comparison of overall Jacobson recovery on the BDI with Jacobson recovery on the HRSD in the same sample revealed that the chance adjusted agreement between measures was 56% (kappa = 0.56) and that the rank order of treatments in studies could change between measures. The relatively low level of agreement between the BDI and HRSD was due to a substantial minority of patients recovering only on a single measure. However, twice as many patients recovered only on the HRSD as did only on the BDI. This finding accords with previous research showing greater symptomatic reduction on the HRSD compared to the BDI in the same sample (Lambert et al., 1986). However, the comparison of Jacobson recovery rates with the published results in Jacobson et al. (1996)

[16] Constantino et al. (2008) also required that symptomatic change was reliable.

indicated that the BDI provided insufficient coverage of depressive symptomatology. This, and the evidence that recovery on the HRSD may be a less stringent test of efficacy than recovery on the BDI suggests that Jacobson recovery on both measures is a better indicator of no longer meeting diagnostic status for an MDE. The symptomatic coverage offered by the use of both measures may thus accord with ACNP Task Force (Rush et al., 2006) recommendation that all of the DSM IV criteria (Table 1) assessed in diagnosis be assessed at post-treatment.

There are limitations which may affect the conclusions reached in this study. Firstly, IPD was not available for all eligible treatment studies which might have limited the representativeness of results. Also, studies varied in the exact version of the HRSD employed which was a potential confound. More importantly, the paucity of follow-up data meant it was not possible to assess whether the gains made during treatment persisted, or whether an increasing percentage of patients recovered during follow-up. A final limitation was that it was not possible to compare individuals' Jacobson recovery status with diagnostic status.

6.5 Summary & Concluding Remarks

The results of the primary analysis showed that, overall, between 50% to 60% of patients entering an active treatment failed to achieve recovery as assessed by the Jacobson method. This indicated that there is considerable room to improve the efficacy of treatments for major depression.

However, whilst there was a considerable range of recovery rates between studies, there was no evidence to suggest that the relative efficacies of specific types of active treatment in individual studies were significantly different. Thus, there was no evidence that the efficacy of specific psychotherapy types were different, or that medication was superior to psychotherapy. These latter conclusions were consistent with the published findings of individual studies which frequently employed different criteria to those of the Jacobson method. However, the use of idiosyncratic recovery criteria meant that published recovery rates were typically higher than those of the Jacobson method, and, considerably so where published results were based on less stringent score criteria on measures. Moreover, in addition to overestimating the efficacy of treatments, the use of idiosyncratic published recovery criteria could lead to changes in the rank ordering of treatments in individual studies in comparison to rates based on the Jacobson method. Consequently, the use of

idiosyncratic recovery criteria means that a balanced appraisal of outcomes for specific treatments across studies is problematic.

Finally, the results showed that agreement between measures was low and that Jacobson recovery rates on the HRSD were higher than on the BDI in samples assessed on both. A potential explanation for this was that neither of these measures adequately captures the full range of depressive symptomatology. This may suggest that Jacobson recovery on both measures is a better indicator of a return to the normative range of depressive symptomatology, or that a measure with better construct validity for major depressive disorder is required to accurately assess treatment outcomes in major depressive disorder.

Chapter Seven

Study 3

Does Severity of Depression at Pre-treatment Predict Recovery and Response Following Acute Treatment?

7.1 Introduction

In the previous chapter, the post-treatment efficacy of psychological and pharmacological treatments for major depressive disorder was evaluated across 6 studies (Constantino et al., 2008; David et al., 2008; DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008) according to the Jacobson method of clinical significance (Jacobson et al., 1999). The results showed that no active treatment intervention was superior to any other within individual studies. However, it was found that recovery rates following an active treatment intervention varied considerably between studies. Jacobson recovery in intention to treat samples (ITT) ranged from 27.3% (Constantino et al., 2008) to 62.0% (Jacobson et al., 1996) on the self-rated Beck Depression Inventory (BDI, Beck et al., 1961), and from 38.3% (DeRubeis et al., 2005) to 66.0% (Jacobson et al., 1996) on the clinician rated Hamilton Rating Scale for Depression (HRSD, Hamilton, 1960).

The patient characteristic with most empirical support concerning its influence on outcome is pre-treatment symptom severity (Hamilton and Dobson, 2002). Evidence indicates that higher levels of depressive symptoms at pre-treatment are predictive of poorer treatment outcomes and the need for longer lasting treatment (Shapiro et al., 1994). For example, in a sample of 37 CBT patients, Jarrett et al. (1991) found that increased BDI or HRSD pre-treatment severity predicted higher scores on the same measure following acute treatment. More recently, in a sample of 318 outpatients, Frank et al. (2011) found that increasing pre-treatment HRSD score was a non-specific predictor of increased time to remission (defined as an HRSD score ≤ 7) across both anti-depressant medication (ADM) and interpersonal therapy (IPT). Thus, Frank et al. (2011) found no evidence for an interaction between pre-treatment severity and the efficacy of specific treatments.

However, the results of the influential Treatment of Depression Collaborative Research Program (TDCRP, Elkin et al., 1989) revealed that, when compared to placebo, the relative efficacy of treatments did significantly change as a function of HRSD pre-treatment severity.

For example, Elkin et al. (1989) compared HRSD symptomatic improvement between ADM, IPT and CBT separately for two levels of pre-treatment HRSD severity. Less severe depression was defined as a pre-treatment HRSD score of 19 points or less, whilst more severe depression was defined as a score of 20 points and above (Elkin et al., 1989). Treatments were compared using analysis of covariance (ANCOVA) with marital status as covariate. The results showed that whilst ADM, IPT and CBT were no different to placebo in less severe samples, both ADM and IPT were superior to placebo in more severe samples. The same conclusions were reached when treatments for each sample were compared in terms of remission which was defined as a post-treatment HRSD score of 6 or less (Elkin et al., 1989). The finding that CBT was no different to placebo in either severity group contrasted with those for ADM and IPT. It also suggested that significant differences between the efficacy of active treatments may emerge at different pre-treatment severities. Indeed, in a more recent analysis of the TDCRP results, Elkin et al. (1995) did show that symptomatic improvement according to the BDI or HRSD in both ADM and IPT groups was superior to CBT in severe depression. Elkin et al. (1995) compared treatments using random regression models that provided greater statistical power than previously due to the inclusion of interim outcome data (Elkin et al., 1995). However, the validity of Elkin et al.'s (1995) findings has been challenged on the basis that slower response to treatment in the CBT group led to a biased estimate of its post-treatment efficacy within ITT analyses (Jacobson and Hollon, 1996).

The TDCRP results (Elkin et al., 1995; Elkin et al., 1989) have influenced the development of guidelines that recommended the use of medication over psychotherapy in the treatment of severe depression (DeRubeis et al., 1999; Driessen et al., 2010). However, conclusions concerning the efficacy of any treatment needs to be based on empirical evidence from more than one study (Chambless and Hollon, 1998). In order to determine whether the efficacy of ADM was superior to CBT in severe depression, DeRubeis et al. (1999) conducted a 'mega-analysis' of ITT post-treatment BDI and HRSD data from 4 methodologically similar studies (Rush et al., 1977; Murphy et al., 1984; Hollon et al., 1992; Elkin et al., 1989). DeRubeis et al. (1999) obtained individual patient data (IPD) from original study authors which enabled them to compare symptomatic improvement between ADM and CBT in patients who scored 20 or more on the HRSD. In deriving more severely depressed sub-samples from original study data, DeRubeis et al. (1999) found that the BDI mean pre-treatment severities for ADM and CBT in Murphy et al. (1984) were significantly different. Consequently, symptomatic improvement following ADM or CBT was compared using ANCOVAs that controlled for patients' pre-treatment severity (DeRubeis et al., 1999). Results for both the BDI (n = 132) and HRSD (n = 169) failed to demonstrate a significant difference between

ADM and CBT both within and across the four studies (DeRubeis et al., 1999). Thus, after controlling for individual differences in pre-treatment severity, DeRubeis et al. (1999) found no empirical evidence to support the superiority of ADM over CBT in severe depression.

It is apparent then, that pre-treatment severity is likely to be an important prognostic factor in the treatment of depression. The evidence indicates that higher severity patients demonstrate lower levels of symptomatic improvement (Jarrett et al., 1991) and that active treatments may be no more efficacious than placebo in less severe depression (Elkin et al., 1995; Elkin et al., 1989). In addition, the TDCRP results (Elkin et al., 1989) also highlighted the possibility that some types of treatment may be more effective than others in severe depression. Nevertheless, Elkin et al.'s (1995) controversial conclusion that ADM was superior to CBT has not been supported in analyses of evidence from multiple studies that controlled for pre-treatment severity (DeRubeis et al., 1999).

However, there are several methodological limitations that may have influenced the results of the studies described here. Firstly, no study controlled for the less than 100% reliability of the symptom measures employed in treatment comparisons. It is possible that the conclusions reached in individual studies were confounded by symptomatic change that could not be attributed to the effect of treatment. For example, it is impossible to know to what degree unreliable symptomatic change influenced the conclusions in Jarrett et al. (1991). It is also unclear to what degree unreliable symptomatic change affected estimates of mean symptomatic improvement that both Elkin et al. (1989) and DeRubeis et al. (1999) used to compare treatments. However, this method is itself unsatisfactory as it provides no information concerning the clinical significance of treatment differences which may themselves be statistically significant (Jacobson et al., 1999). A second limitation was that whilst Elkin et al. (1989) did compare clinically significant outcomes for ADM, IPT and CBT with placebo in more severe samples, they used an HRSD remission criterion of 6 or less. This was more stringent than the recovery criterion of 7 or less which was shown to best represent a return to normal functioning according to the Jacobson Method in chapter 6. Consequently, Elkin et al.'s (1995) finding that both ADM and IPT, but not CBT, were superior to placebo in severe samples was possibly based on an analysis that excluded some non-depressed patients. Finally, in the TDCRP Elkin et al. (1989) investigated the influence of pre-treatment severity on the relative efficacy of treatments compared to placebo using dichotomised samples. Consequently, their finding that CBT was no different to placebo in more severe depression may have capitalised on chance, as the HRSD cut-score used to define more severe depression was exploratory despite being selected *a priori* (Elkin et al., 1989).

The main purpose of this study was to determine whether pre-treatment severity on the BDI or HRSD predicted clinically significant recovery according to the methodology outlined by Jacobson and colleagues (Jacobson et al., 1984; Jacobson and Truax, 1991). Individual patient data derived from the studies described in chapter 6 were employed in hierarchical binary logistic regression analyses that used pre-treatment severity as the independent variable and Jacobson recovery status as the dependent variable. The use of Jacobson recovery status as a common outcome metric meant that limitations arising from idiosyncratic definitions of clinical significance and unreliable symptomatic change were avoided. It was hypothesised that increasing pre-treatment severity would significantly reduce the probability of Jacobson recovery on either measure. A secondary aim was to determine whether pre-treatment severity on the BDI or HRSD predicted Jacobson response, which was defined as a statistically reliable reduction in symptom score according to the Jacobson methodology (Jacobson and Truax, 1991). Hierarchical binary logistic regression analyses similar to those for Jacobson recovery status were undertaken but with Jacobson response status as the dependent variable. The exploratory nature of the Jacobson response analyses meant that no hypothesis was formed.

7.2 Method

The present study consisted of four one-stage fixed-effects IPD meta-analyses (Simmonds et al., 2005) employing hierarchical binary logistic regression. The four analyses examined whether pre-treatment severity predicted (i) Jacobson recovery and (ii) Jacobson response on the BDI and HRSD. Each analysis was stratified by study, treatment type and gender in order to reduce potential confounds due to baseline differences in recovery or response between the levels of each covariate (Simmonds et al., 2005; van Walraven, 2010). This method provided a more flexible and powerful approach than conventional meta-analysis by which to (i) examine the effect of pre-treatment severity on recovery or response and (ii) detect the existence of significant interactions between covariates (Higgins et al., 2001; Simmonds et al., 2005).

7.2.1 Individual Patient Data Used In Logistic Regression Analyses

The sample comprised selected individual patient data (IPD) from the BDI and HRSD studies described in study 2. Data for each patient consisted of pre-treatment severity and Jacobson recovery/response status on measures, study, treatment type (ADM or psychotherapy) and gender. To ensure that patient attrition did not confound results as a

result of swifter onset of ADM effects, only IPD for active treatment completers were used in analyses. IPD from 4 of 5 available studies were employed in BDI analyses (David et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). Constantino et al. (2008) was excluded from BDI analyses because the inclusion of IPD for only 22 patients meant that sparse data (Cohen et al., 2003) led to poor model fit in preliminary analyses. IPD from 4 of 5 available studies were employed in HRSD analyses (DeRubeis et al., 2005; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). Dekker et al. (2008) was excluded because IPD from this study did not correspond to a full course of treatment. Because a variety of pharmacological and psychological interventions were used across studies, treatment type was broadly categorised as either ADM or psychotherapy (see study 2 for details).

7.2.2 Data Analytic Strategy

The method below applies to all four regression analyses. For ease of presentation, the descriptions refer only to Jacobson recovery.

Included variables

Each analysis was a two-step, forced hierarchical binary logistic regression, with Jacobson recovery status as the dependent variable. Pre-treatment severity was entered as the predictor variable. The categorical variables study, treatment type and gender were entered as covariates. The inclusion of covariates meant that the existence of significant 1st order interactions between variables could be examined at the second step. Apart from study and treatment type, the only covariate provided for all patients across IPD studies was gender. This was included in regression models as some studies have reported significant interactions between gender and treatment type (Bigos et al., 2009; Parker et al., 2011).

Order of variables entered and coding scheme

Variables were entered into regression models according to the order shown in Table 22. The method of variable entry followed recommendations by Frazier et al. (2004). At step 1, the covariates study, treatment type (treatment), gender and pre-treatment severity (severity) were entered simultaneously. At step 2, all 1st order interaction terms for the variables entered at step 1 were included in the model. Table 22 also describes the meaning of a significant regression coefficient for each variable.

For categorical variables, effects coding was employed to ensure that the effect of each individual variable was calculated at the average effect of remaining variables (Cohen et al., 2003). For example, the two levels of the categorical variable ‘treatment’ were coded as -1

for ADM and +1 for psychotherapy, whilst 'gender' was coded as -1 for males and +1 for females. This meant that where $B_{treatment} = 0$, the B_{gender} coefficient referred to the log-odds of recovery in either males ($-B_{gender}$) or females ($+B_{gender}$) averaged across both ADM and psychotherapy. The difference between the log-odds of male and female recovery in this example is given by the difference between their respective coefficients ($2 \times B_{gender}$). The use of effects coding also meant that the constant (B_0) at each step in regression models represented the log-odds of recovery after controlling for the effects of all included variables. The continuous variable pre-treatment severity (BDI or HRSD) was centred at the sample mean in order to eliminate non-essential collinearity and produce interpretable regression coefficients (Cohen et al., 2003). To obtain a centred score for each patient, the sample mean pre-treatment severity was subtracted from their pre-treatment score. Centring meant that the regression coefficients of categorical variables referred to patients whose pre-treatment severity was equal to the sample mean severity.

The validity of results and their interpretation

Multicollinearity tests for the variables entered at step 1 were performed and tolerance values greater than 0.2 were taken to indicate that unstable regression coefficients did not threaten the validity of final models (Cohen et al., 2003). Models were also checked for bias arising from influential cases by examining the predicted probability, standardised residual, Cook's distance analogue, leverage and DFBeta values for each case in analyses (Cohen et al., 2003).

The results of each valid regression analysis were examined to determine the significance of each step, of the overall model, and of individual regression coefficients. Where a significant B coefficient was identified for a variable within a significant model, the predicted log-odds of recovery were calculated for each level of that variable. Subsequently, the predicted log-odds for each level were transformed into odds in order to calculate the more interpretable predicted probability of recovery (p) as given by the following relationship:

$$p = odds/(1 + odds).$$

The predicted probabilities of recovery for each level of significant variables were then plotted as a function of pre-treatment severity on the BDI or HRSD. The statistical package SPSS (version 12) was used to perform the regression analyses whose results are presented in tabular form. Microsoft Excel 2007 was used to calculate and plot the predicted probabilities across the range of BDI and HRSD pre-treatment severities.

Table 22. Variables Entered in Hierarchical Binary Logistic Regression Analyses Investigating Jacobson Method Recovery & Response.

Variable	Significant Result for Variable Indicates:
STEP 1	
Study†	Recovery* rates differed between studies
Treatment	Recovery rates differed between treatments
Gender	Recovery rates differed between genders
Severity	Recovery rates differed as a function of pre-treatment severity
STEP 2	
Study x Treatment	Recovery rates for ADM and psychotherapy differed in at least one study
Study x Gender	Recovery rates for males and females differed in at least one study.
Study x Severity	Recovery rates differed between studies as a function of pre-treatment severity
Treatment x Gender	Recovery rates differed between genders across treatments
Treatment x Severity	Recovery rates differed between treatments as a function of pre-treatment severity
Gender x Severity	Recovery rates differed between genders as a function of pre-treatment severity

† Refers to the overall significance of including Study in the model. To aid clarity, individual study coefficients are not described for Study nor its interaction terms.

* Response may be substituted for recovery throughout the table

7.3 Results

The first sections present the following descriptive statistics for each level of treatment type and gender across BDI and HRSD studies; (i) comparisons of the mean pre-treatment severity between recovered and unrecovered groups, (ii) comparisons of the mean pre-treatment severity between responders and non-responders ('no change'), (iii) Jacobson recovery and response rates. Following this, the results of the binary logistic regression analyses for Jacobson recovery and response in BDI and HRSD studies are presented.

7.3.1 Pre-treatment Severity, Recovery & Response across BDI & HRSD Studies

Table 23 presents the pre-treatment means of recovered versus unrecovered completers by treatment type and gender across BDI and HRSD studies. Table 24 presents the pre-treatment means of responders versus non-responders by treatment type and gender across BDI and HRSD studies. Significant differences between the pre-treatment means of the groups in Table 23 and Table 24 were examined using t-tests with Bonferroni corrections to adjust for multiple comparisons.

In terms of Jacobson recovery, Table 23 shows that there were no significant differences between the mean pre-treatment severity of recovered and unrecovered patients in any BDI group. For the HRSD, the overall mean pre-treatment severity was higher in unrecovered patients than in recovered ($t = 4.03$, $p < .01$). However, no differences were found between the pre-treatment means of the remaining HRSD groups in Table 23. A comparison of the post-treatment means between recovered and unrecovered patients within the groups in Table 23 was undertaken using t-tests adjusted for multiple comparisons. The results showed that for both the BDI and HRSD, the post-treatment mean of unrecovered patients within all groups was significantly higher than that of recovered patients (all $p < .01$).

In terms of Jacobson response, Table 24 shows that the overall mean pre-treatment severity of responders was significantly higher than that of non-responders in BDI studies ($t = 6.6$, $p < .01$). Furthermore, Table 24 reveals that the BDI pre-treatment mean in responders was significantly higher than in non-responders for the ADM ($t = 3.8$, $p < .01$), psychotherapy ($t = 5.5$, $p < .01$), male ($t = 6.6$, $p < .01$) and female groups ($t = 3.6$, $p < .05$). In contrast, there were no significant differences between the pre-treatment means of responders and non-responders within any of the HRSD groups in Table 24. A comparison of post-treatment means within the groups in Table 24 showed that for both measures, the post-treatment mean of non-responders was significantly higher than that of responders (all $p < .01$).

Table 23. BDI & HRSD Pre-treatment Mean Severity by Jacobson Recovery Status for Treatment Type & Gender.

		Recovered			Unrecovered				
Group	N	n	Mean	s.d.	n	Mean	s.d.	t	df
BDI									
Overall†	383	208	28.0	8.4	175	29.5	9.7	1.6	346.0
ADM	95	50	27.1	10.0	45	28.4	11.1	0.6	93
Psychotherapy	288	158	28.2	7.8	130	29.9	9.2	1.6	286
Males	109	56	29.4	10.1	53	27.1	10.5	1.1	107
Females	274	152	27.4	7.6	122	30.5	9.2	3.0	233.3
HRSD									
Overall†	383	223	19.6	4.3	160	21.4	4.3	4.0*	381
ADM	147	84	20.9	4.3	63	22.9	4.1	2.9	145
Psychotherapy	236	139	18.8	4.2	97	20.4	4.2	2.9	234
Males	123	63	19.8	4.5	60	22.0	4.2	2.9	121
Females	260	163	19.5	4.3	97	21.0	4.4	2.7	258

$p < .01$ adjusted for multiple comparisons (Bonferroni).

† Identical overall patient numbers in BDI and HRSD studies was coincidental.

Table 24. BDI & HRSD Pre-treatment Mean Severity by Jacobson Response Status for Treatment Type & Gender.

		Responded			No Change				
Group	N	n	Mean	s.d.	n	Mean	s.d.	t	df
BDI									
Overall†	383	305	29.9	9.2	78	23.9	6.4	6.6*	166.6
ADM	95	76	29.3	10.6	19	21.3	7.5	3.8*	38.2
Psychotherapy	288	229	30.1	8.7	59	24.8	5.9	5.5*	131.0
Males	109	81	30.8	10.4	28	20.9	5.1	6.6*	94.8
Females	274	224	29.5	8.7	50	25.6	6.5	3.6α	92.1
HRSD									
Overall†	383	329	20.6	4.4	54	18.8	3.9	2.8	381
ADM	147	129	22.0	4.3	18	20.1	4.0	1.8	145
Psychotherapy	236	200	19.7	4.3	36	18.2	3.8	2.0	234
Males	123	105	21.2	4.5	18	19.7	4.4	1.3	121
Females	260	224	20.3	4.4	36	18.4	3.6	2.5	258

* $p < .01$ adjusted for multiple comparisons (Bonferroni).

α $p < .05$ adjusted for multiple comparisons (Bonferroni).

† Identical overall patient numbers in BDI and HRSD studies was coincidental

7.3.2 Percentage Recovering & Responding across BDI & HRSD Studies

Table 25 presents the percentage of active treatment completers across BDI and HRSD studies who recovered or responded[17] at post-treatment according to Jacobson's criteria. The results presented in Table 25 should be considered as descriptive as they did not take the following potential confounds into account; (i) ADM and psychotherapy rates were not derived from studies that had all made direct comparisons between treatments, (ii) male and female rates did not control for treatment or study differences.

Table 25. Percentage Recovering & Responding by Treatment Type & Gender across BDI & HRSD Studies.

			Recovered		Responded	
		N	n	%	n	%
BDI	Overall†	383	208	54.3	305	79.6
	ADM	95	50	52.6	76	80.0
	Psychotherapy	288	158	54.9	229	79.5
	Males	109	56	51.4	81	74.3
	Females	274	152	55.5	224	81.8
HRSD	Overall†	383	223	58.2	329	85.9
	ADM	147	84	57.1	129	87.8
	Psychotherapy	236	139	58.9	200	84.7
	Males	123	63	48.8	105	85.4
	Females	260	163	62.7	224	86.2

† Identical patient numbers for BDI and HRSD was coincidental.

In terms of Jacobson recovery on the BDI, Table 25 shows that the overall rate was 54.3% in a sample of 383 completers. Thus, 45.7% of completers were unrecovered across BDI studies at post-treatment. Goodness of fit tests showed that BDI recovery rates did not significantly differ between ADM and psychotherapy (52.6% versus 54.9% respectively; $\chi^2 = 0.14$, $p = .70$), nor between males and females (51.4% versus 55.5% respectively; $\chi^2 = 0.53$, $p = .47$). In terms of Jacobson response on the BDI, Table 25 shows that the overall rate was 79.6%. Thus, 20.4% of completers across the overall BDI sample demonstrated no reliable change in symptomatology at post-treatment. Goodness of fit tests showed that BDI response rates did not significantly differ between ADM and psychotherapy (80.0% versus 79.5% respectively; $\chi^2 = 0.01$, $p = .92$), nor between males and females (74.3% versus 81.8% respectively; $\chi^2 = 2.66$, $p = .10$).

[17] N.B. response = statistically reliable reduction in symptom score

In terms of Jacobson recovery on the HRSD, Table 25 shows the overall rate was 58.2% in a sample of 383 patients. Thus, 41.8% of completers across the HRSD sample were unrecovered at post-treatment. Goodness of fit tests showed that recovery rates did not significantly differ between ADM and psychotherapy patients (57.1% versus 58.9% respectively; $\chi^2 = 0.12$, $p = .7$). However, goodness of fit tests revealed that the female recovery rate across HRSD studies was significantly higher than the male recovery rate (62.7% versus 48.8%, $\chi^2 = 6.64$, $p = .01$). In terms of Jacobson response on the HRSD, Table 25 shows that the overall rate was 85.9%. Thus, 14.1% of completers across the overall HRSD sample demonstrated no reliable change in symptomatology at post-treatment. Goodness of fit tests showed that response rates did not significantly differ between ADM and psychotherapy (87.8% versus 84.7% respectively; $\chi^2 = 0.68$, $p = .41$), nor between males and females (85.4% versus 86.2% respectively; $\chi^2 = 0.04$, $p = .84$).

7.3.3 Binary Logistic Regression Analyses for Jacobson Recovery

BDI results

Table 26 presents the results of the binary logistic regression analysis for Jacobson recovery across BDI studies. The independent variables study, treatment type, gender and pre-treatment severity were regressed on Jacobson recovery status. Pre-treatment BDI severity was centred using the overall mean of 28.7 points. Collinearity tests indicated that the regression analysis was appropriate as the lowest tolerance value for any of the 1st order terms in Table 26 was 0.92.

The results in Table 26 reveal that the variables entered at step1 did not produce a model that significantly predicted recovery (step $\chi^2 = 11.3$, $p = .08$). Also, step 2 did not improve prediction above that of step 1 (step $\chi^2 = 18.90$, $p = .06$). However, the inclusion of all interaction terms at step 2 did produce an overall model that significantly predicted recovery ($\chi^2 = 30.2$, $p = .025$). The results for step 2 show that the only the gender x severity interaction coefficient was significant (B = -0.03, $p = .02$). In order to better understand the interaction between gender and severity on BDI recovery, the 1st step analysis presented in Table 26 was undertaken separately for males and females (see Table 27). The pre-treatment BDI severity used in the male and female analyses were centred using a mean pre-treatment BDI score of 28.3 and 28.8 points respectively. Collinearity tests indicated that the regression analysis was appropriate as the lowest tolerance value for any of the 1st order terms in Table 27 was 0.87.

Table 26. Results of Logistic Regression Analysis for BDI Recovery.

	Step χ^2	$\Delta R^{2\alpha}$	Variable	B	*p*	S.E.	Odds
Step 0			Constant	0.17	.09	0.10	1.19
Step 1	11.30	.04			.08		
			Constant	0.10	.48	0.14	1.11
			Study$^{\beta}$		.05		
			Treatment	- 0.08	.53	0.13	0.92
			Gender	0.06	.58	0.12	1.06
			Severity	- 0.02	.08	0.01	0.98
Step 2†	18.90	.06			.06		
			Constant	0.04	.90	0.31	1.04
			Study$^{\beta}$		.16		
			Treatment	- 0.18	.54	0.30	0.83
			Gender	0.19	.24	0.16	1.21
			Severity	- 0.04	.09	0.02	0.96
			Study$^{\beta}$ x Treatment		.87		
			Study$^{\beta}$ x Gender		.81		
			Study$^{\beta}$ x Severity		.08		
			Treatment x Gender	- 0.18	.24	0.15	0.84
			Treatment x Severity	0.00	.80	0.01	1.00
			Gender x Severity	- 0.03	.02	0.01	0.97

† Model significance = .025; β: Results for individual studies not shown

Note: Effects coding employed for all categorical variables. Males, Salminen et al. (2008) and ADM served as baseline for Gender, Study and Treatment respectively.

Table 27. Results of Logistic Regression Analyses for BDI Recovery by Gender.

		Step χ^2	$\Delta R^{2\alpha}$	Variable	B	*p*	S.E.	Odds
Male:	Step 0			Constant	0.06	.77	0.19	1.06
	Step 1	5.32	.06			.38		
				Constant	- 0.04	.86	0.26	0.96
				Study$^{\beta}$		.47		
				Treatment	0.11	.66	0.24	1.11
				Severity	0.02	.34	0.02	1.02
Female:	Step 0			Constant	0.22	.07	0.12	1.25
	Step 1	15.96	.08			.01		
				Constant	0.21	.18	0.16	1.24
				Study$^{\beta}$		.08		
				Treatment	- 0.19	.24	0.17	0.83
				Severity	- 0.05	.01	0.02	0.95

α: Nagelkerke's pseudo R^2 ; β: Results for individual studies not shown

Note: Effects coding employed for all categorical variables. Salminen et al. (2008) and ADM served as baseline for Study and Treatment respectively.

The results for males in Table 27 show that the step 1 model was not significant (χ^2 = 5.32, *p* = .38) and that no covariate significantly predicted recovery. In contrast, the step 1 model for females was significant (χ^2 = 15.96, *p* = .01) and revealed that pre-treatment severity alone significantly predicted recovery (B = -0.05, *p* = .01). However, the results in Table 27 also suggest a trend for a significant difference between studies (*p* = .08) for female recovery which is not considered further. In order to clarify the meaning of the results in Table 26 and Table 27, Figure 3 presents the predicted probability of recovery for males and females as a function of BDI pre-treatment severity (across studies and treatments). The figure was based on the coefficients in the 2nd step model in Table 26 and used the following regression equation

$$\text{Log-odds (recovery)} = B_0 + B_{severity}.\ X_{severity} + B_{gender}.\ X_{gender} + B_{gender\ x\ severity}.\ X_{gender}.X_{severity}$$

where pre-treatment severity was centred at the mean for both genders of 28.7 points with X_{gender} coded as -1 and +1 for males and females respectively.

Figure 3. Predicted Probability of Male & Female Recovery as a Function of BDI Pre-treatment Severity.

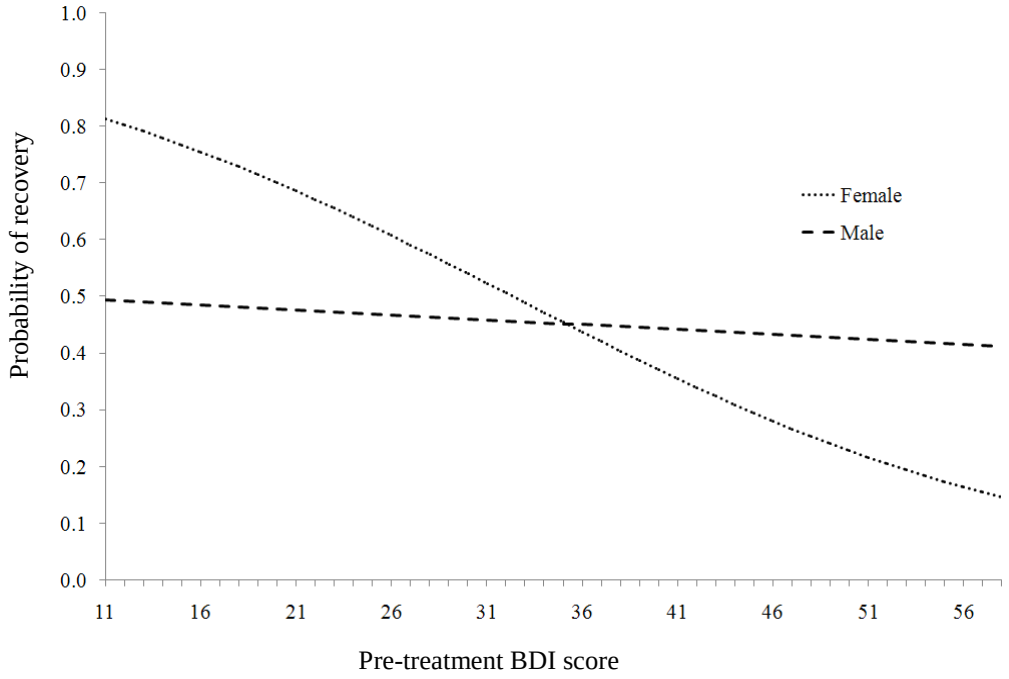


Figure 3 shows that the predicted probability of male recovery was effectively constant across the range of BDI pre-treatment severity. This finding agreed with the results for males in Table 27 where $B_{severity}$ was not significantly different from zero. In contrast, Figure 3 shows that the predicted probability of female recovery reduced with increasing BDI pre-treatment severity. This finding also agreed with the results for females in Table 27, where a significant $B_{severity}$ coefficient indicated that the predicted probability of recovery reduced with increasing BDI pre-treatment severity. However, whilst the predicted probability of recovery was the same for both sexes at a BDI pre-treatment score of approximately 35 points, Figure 3 indicates that significantly fewer males recovered at lower BDI scores than females and vice versa. Thus, despite an overall finding that gender did not significantly predict recovery (B_{gender} = 0.19, p = .24; Table 26), Figure 3 indicates that the significant gender x severity interaction term identified significant gender differences at both the lower and upper ranges of BDI pre-treatment severity. However, an analysis[18] using the 95% confidence intervals for $B_{gender\ x\ severity}$ suggested that the significant difference between predicted male and female recovery occurred for BDI scores of only 31 points and below.

Summary

The step 2 model BDI results showed that, overall, pre-treatment severity was not predictive of recovery ($B_{severity}$ = -0.04, p = .09). However, it was revealed that BDI pre-treatment severity was a significant predictor of recovery in females. Despite no difference in the overall recovery rate between genders (B_{gender} = 0.19, p = .24), the significant gender x severity interaction ($B_{gender\ x\ severity}$ = -0.03, p = .02) indicated that the probability of recovery differed between genders as a function of BDI pre-treatment severity. This interaction revealed that the probability of recovery in females significantly reduced as pre-treatment severity increased, whereas, in males the probability of recovery was effectively constant (Figure 3). Moreover, it was shown that for the same pre-treatment score, the predicted probability of recovery was significantly higher in females than males for BDI scores of 31 points and below. The non-significant results for remaining 1st order covariates study (p = .16) and treatment ($B_{treatment}$ = -0.18, p = .54) at step 2 indicated that the probability of recovery did not differ between studies or treatments. Also, the non-significant interaction between treatment and severity ($B_{treatment\ x\ severity}$ = 0.00, p = .80) provided no evidence that ADM and psychotherapy were differentially effective at differing pre-treatment severities on the BDI. Finally, the goodness-of-fit statistic, Nagelkerke's pseudo R^2 for the 2 step model in Table 26 was 0.1. This indicated that the overall model accounted for 10% of the null deviance (Cohen et al., 2003).

[18] This did not account for the error in $B_{severity}$ or B_{gender}

HRSD results

Table 28 shows the results of the binary logistic regression analysis for the HRSD. The independent variables study, treatment type, gender and pre-treatment severity were regressed on Jacobson recovery status. Pre-treatment HRSD severity was centred using the overall mean of 20.3 points. Collinearity tests indicated that the regression analysis was appropriate as the lowest tolerance value for any of the 1st order terms in Table 28 was 0.72.

Table 28. Results of Logistic Regression Analysis Investigating the effect of HRSD Pre-treatment Severity on the Probability of Recovery.

	Step χ^2	$\Delta R^{2\alpha}$	Variable	B	*p*	S.E.	Odds
Step 0			Constant	0.33	.001	0.10	1.39
Step 1	24.71	.08			.001		
			Constant	0.25	.07	0.14	1.29
			Study$^{\beta}$		.38		
			Treatment	- 0.20	.14	0.14	0.82
			Gender	0.26	.03	0.03	1.29
			Severity	- 0.10	.001	0.003	0.91

α: Nagelkerke's pseudo R^2

β: results for individual studies not shown.

Note: Effects coding employed for all categorical variables. Males, Salminen et al. (2008) and ADM served as baseline for Gender, Study and Treatment respectively.

The step 1 results in Table 28 show that that the inclusion of study, treatment, gender and pre-treatment severity resulted in a model that significantly predicted recovery on the HRSD (χ^2 = 24.71, *p* = .001). Step 2 results are not presented as (i) the step was not significant (step χ^2 = 4.81, *p* = .94), (ii) no significant interactions were identified despite a significant overall model (model χ^2 = 29.52, *p* = .03). The significant predictors of HRSD recovery in Table 28 were gender (B = 0.26, *p* = .03) and pre-treatment HRSD severity (B = -0.1, *p* = .001). Figure 4 presents the predicted probabilities of male and female recovery (across studies and treatments) as a function of HRSD pre-treatment severity. Figure 4 was based on the results in Table 28 and used the following regression equation:

$$\text{Log-odds (recovery)} = B_0 + B_{gender} . X_{gender} + B_{severity} . X_{severity}$$

Here, $B_{severity}$ was centred at the overall mean pre-treatment severity of 20.3 points and X_{gender} was -1 and +1 respectively for males and females.

Figure 4. Predicted Probability of Male & Female Recovery as a Function of HRSD Pre-treatment Severity.

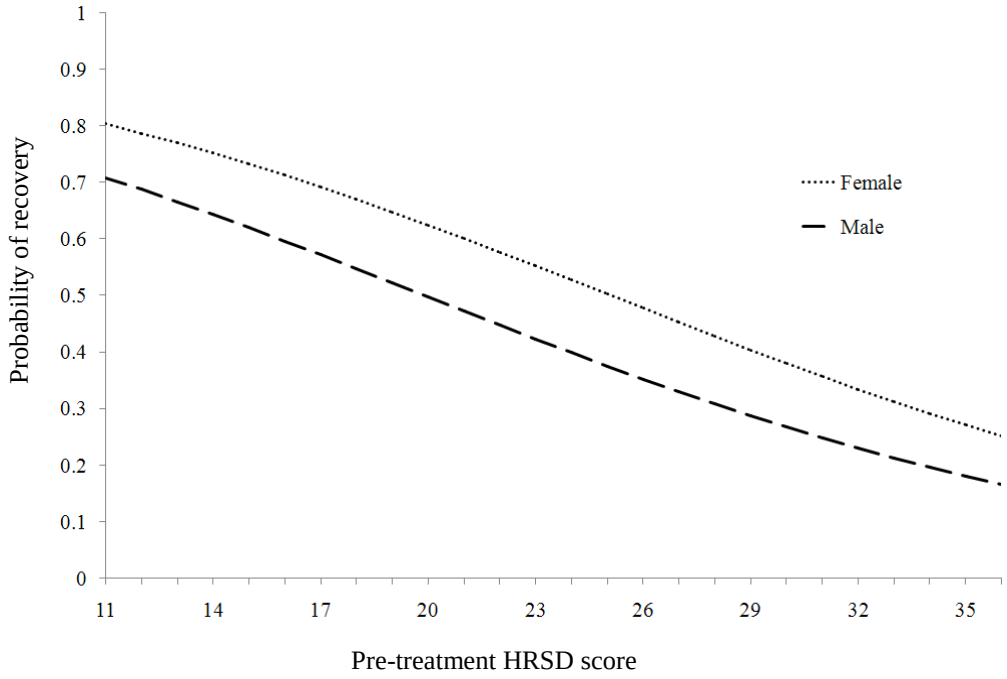


Figure 4 shows that (i) the predicted probability of recovery for both genders reduced with increasing pre-treatment severity and (ii) the magnitude of the probability of male recovery was significantly lower than female recovery at the mean HRSD score of 20.3 points. Whilst the figure also suggests that the difference between male and female recovery was significant across the entire range of HRSD pre-treatment severity, an examination of the results using the 95% confidence intervals[19] for $B_{severity}$ suggested that the difference was significant only for HRSD pre-treatment scores ranging from 16 to 24 points.

Summary

The step1 model HRSD results showed that both pre-treatment severity and gender significantly predicted recovery ($B_{severity}$ = -0.10, p = .001; B_{gender} = 0.26, p = .03). The result for severity indicated that the probability of recovery significantly reduced with increasing pre-treatment severity in both males and females. However, the significant result for gender also revealed that the probability of male recovery was significantly lower than that of females at the pre-treatment HRSD mean of 20.3 points. Moreover, for equivalent pre-treatment scores, it was shown that the predicted probability of recovery was significantly

[19] These are omitted for clarity.

lower in males than females for HRSD scores between approximately 16 and 24 points. The non-significant results for the remaining covariates study (p = .38) and treatment ($B_{treatment}$ = -0.20, p = .14) meant that the probability of recovery was not significantly different between studies or treatments. The use of a step 1 model for the HRSD meant that it was not possible to test for interactions between 1st order covariates. Finally, Nagelkerke's pseudo R^2 for the 1 step model in Table 28 was 0.08 indicating that the model accounted for 8% of the null deviance (Cohen et al., 2003).

7.3.4 Binary Logistic Regression Analyses for Jacobson Response

The role of pre-treatment severity on Jacobson response was investigated using the same method as for Jacobson recovery. Patients were classified as responding to treatment if they recovered or demonstrated a statistically reliable reduction in symptoms at post-treatment according to Jacobson criteria (Jacobson and Truax, 1991).

BDI results

Table 29 presents results of the binary logistic regression analysis for Jacobson response across BDI studies. The independent variables study, treatment type, gender and pre-treatment severity were regressed on Jacobson response status. Pre-treatment BDI severity was centred using the overall mean of 28.7 points. Collinearity tests indicated that the regression analysis was appropriate as the lowest tolerance value for any of the 1st order terms in Table 29 was 0.92.

The results in Table 29 show that the step 1 model was significant (step1 χ^2 = 39.92, p < .001) and that only pre-treatment severity (B = 0.09, p = .001) was predictive of Jacobson response on the BDI. Neither study, treatment or gender were significant predictors of response in the step 1 model. However, Table 29 shows that the step 2 model also significantly predicted Jacobson response on the BDI (step 2 χ^2 = 21.55, p = .03; overall model χ^2 = 61.47, p < .001). For the step 2 model the variables severity (B = 0.14, p = .001), study (p = .045), study x severity (p < .05) and gender x severity (B = -0.07, p = .03) were all predictive of response. Thus, the inclusion of interaction terms in the step 2 model led to a significant improvement in the prediction of response over that of the step 1 model.

Table 29. Results of Logistic Regression Analysis for BDI Response.

	Step χ^2	$\Delta R^{2\alpha}$	Variable	B	*p*	S.E.	Odds
Step 0			Constant	1.36	.001	0.13	3.91
Step 1					.001		
	39.92	.16	Constant	1.48	.001	0.20	4.38
			Study$^{\beta}$		.19		
			Treatment	- 0.23	.18	0.17	0.79
			Gender	0.19	.20	0.14	1.21
			Severity	0.09	.001	0.02	1.10
Step 2†	21.55	.08			.03		
			Constant	2.13	.001	0.5	8.44
			Study$^{\beta}$		.045		
			Treatment	- 0.90	.09	0.54	0.41
			Gender	- 0.02	.93	0.27	0.98
			Severity	0.14	.001	0.04	1.15
			Study$^{\beta}$ x Treatment		.70		
			Study$^{\beta}$ x Gender		.51		
			Study$^{\beta}$ x Severity		.03		
			Treatment x Gender	- 0.18	.36	0.20	0.83
			Treatment x Severity	- 0.05	.21	0.04	0.95
			Gender x Severity	- 0.07	.03	0.03	0.94

† Model significance < .001

S.E. = Standard error of B
α: Nagelkerke's pseudo R^2
β: results for individual studies not shown.

Note: Effects coding employed for all categorical variables. Males, Salminen et al. (2008) and ADM served as baseline for Gender, Study and Treatment respectively.

An interpretation of the step 2 results in Table 29 is presented separately for each significant predictor of response on the BDI.

Severity

The significant finding for severity ($B_{severity}$ = 0.14, p < .001) indicated that, overall, the probability of response increased with increasing BDI pre-treatment severity. This finding was similar to that for the step 1 model, where pre-treatment severity was the only significant predictor of response. However, the inclusion of 1st order interactions in the step 2 model meant that $B_{severity}$ increased from 0.09 to 0.14 at the second step (see Table 29).

Study

The significant finding for study (p = .045) indicated that there were significant differences in overall response rates between studies. An examination of the B_{study} coefficient for each study[20] showed that the B_{study} coefficient for Salminen et al. (2008) was the only one that significantly differed from zero (B = -1.64, p < .05). This revealed that the odds of response in Salminen et al. (2008) were 0.19 that of the overall rate based on all studies (David et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008) at a BDI pre-treatment severity of 28.7 points whilst controlling for treatment and gender.

Study x Severity

The significant study x severity interaction indicated that the predicted probability of response significantly differed between studies as a function of pre-treatment severity. An examination of individual study coefficients revealed that the $B_{study\ x\ severity}$ coefficient in two of the four studies was significantly different from zero (David et al., 2008; Salminen et al., 2008;). A significant $B_{study\ x\ severity}$ of +0.11 (p < .05) for David et al. (2008) indicated that the predicted increase in the log-odds of response for a 1 point increase on the BDI above 28.7 points in this study was significantly greater than the predicted increase based on all studies (David et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008). Conversely, a $B_{study\ x\ severity}$ of -0.14 (p < .05) for Salminen et al. (2008) indicated that the increase in the predicted log-odds of response for a 1 point increase on the BDI above 28.7 points in this study was significantly lower than the predicted increase based on all studies (David et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999; Salminen et al., 2008).

In order to better understand the effect of the significant study x severity interaction on overall response in David et al. (2008) and Salminen et al. (2008), the probability of response as a function of pre-treatment severity was derived for both the 1st and 2nd step regression models in Table 29. The regression equations employed for each model were:[21]

Step 1. $\text{Log-odds (response)} = B_0 + B_{study} + B_{severity}.X_{severity}$

Step 2. $\text{Log-odds (response)} = B_0 + B_{study} + B_{severity}.X_{severity} + B_{study\ x\ severity}.X_{severity}$

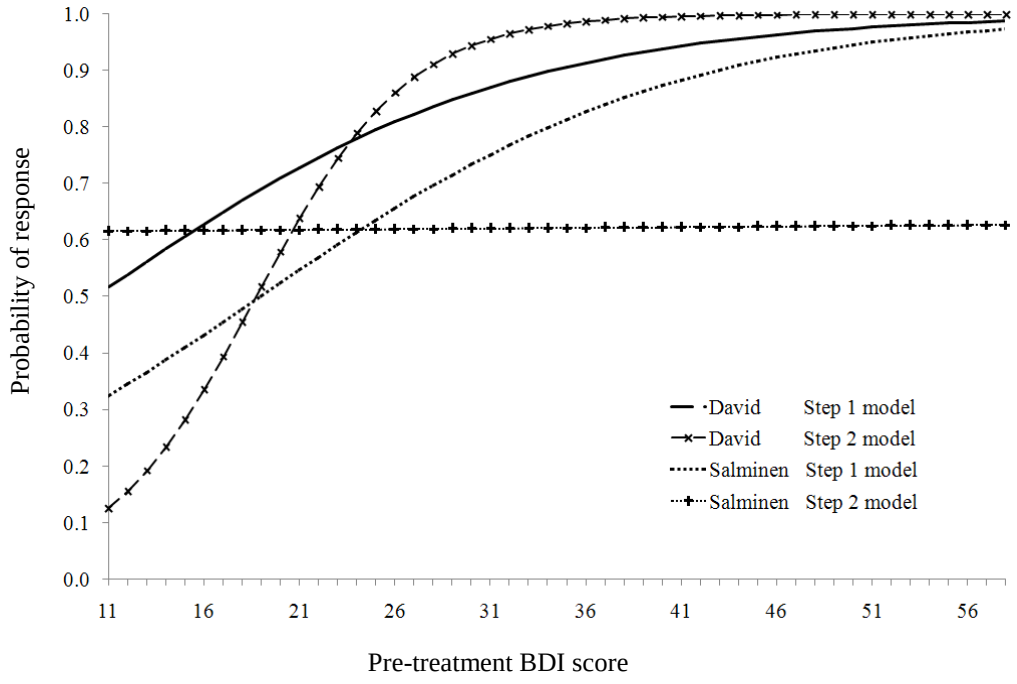
[20] Individual B_{study} coefficients are not presented in the table.

[21] Whilst study was not significant according to the step 1 results in Table 29, the step 1 B_{study} values for David et al. (2008) and Salminen et al. (2008) were included for the purpose of comparison.

Figure 5 presents a comparison of the predicted probability of response by BDI pre-treatment severity for David et al. (2008) and Salminen et al. (2008) according to both step 1 and step 2 regression results.

Figure 5 shows that for the step 1 model the probability of response in both David et al. (2008) and Salminen et al. (2008) increased with increasing pre-treatment severity. This corresponded to the significant result for $B_{severity}$ in the step 1 model in Table 29 where no other variable was predictive of response. In addition, Figure 5 shows that the probability of response in David et al. (2008) was higher than that in Salminen et al. (2008) for every pre-treatment BDI score. This was solely due to the (non-significant) difference between the B_{study} coefficients for David et al. (2008) and Salminen et al. (2008).

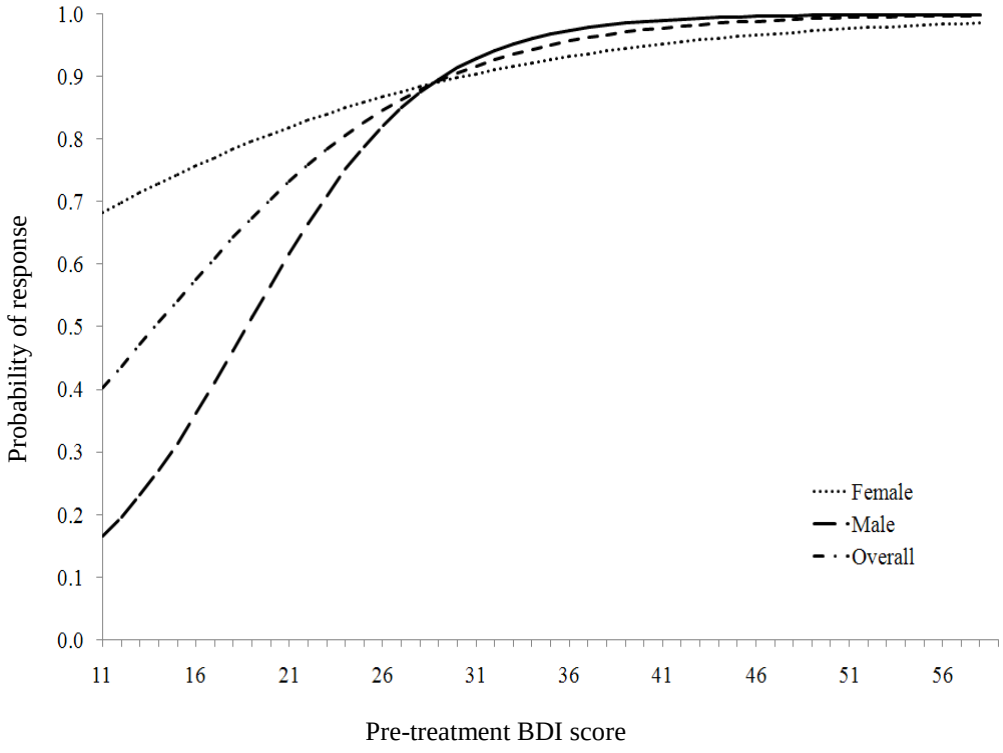
Figure 5. Predicted Probability of Response as a function of BDI Pre-treatment Severity: A Comparison of Steps 1 & 2 Regression Results for David et al. (2008) & Salminen et al. (2008).



In contrast to the step 1 model, the results for the 2nd step model in Figure 5 show that including the study x severity interaction led to dramatically different results. Here, the probability of response in David et al. (2008) was reduced at lower pre-treatment severities and increased at higher pre-treatment severities in comparison to step 1 results.

Nevertheless, the 1st and 2nd step model plots for David et al. (2008) broadly agree in that the probability of response significantly increased with increasing pre-treatment severity. However, Figure 5 reveals that the step 1 and 2 plots for Salminen et al. (2008) were very different. Here, the plot for the 2nd step model indicates that the probability of response was constant across all BDI pre-treatment severities. This finding cannot be attributed to the significantly lower probability of response in Salminen et al. (2008) (see page 132), as the B_{study} coefficient affects only the intercept and not the gradient of predicted probability plots. Thus, despite the 2 step model finding that severity was predictive of response overall, the significant study x severity interaction revealed that pre-treatment severity was not predictive of response in Salminen et al. (2008).

Figure 6. Predicted Probability of Male & Female Response as a function of BDI Pre-treatment Severity.



Gender x Severity

The significant gender x severity interaction for the 2nd step model in Table 29 revealed that the effect of pre-treatment severity on response was not consistent between genders. Figure 6 presents the predicted probability of response as a function of BDI pre-treatment severity

for males and females. Figure 6 was derived using the 2nd step coefficients in Table 29 in the following regression equation:

$$\text{Log-odds (response)} = B_0 + B_{gender}.X_{gender} + B_{severity}.X_{severity} + B_{gender\ x\ severity}.X_{gender}.\ X_{severity}$$

Here, pre-treatment severity was centred at the mean for both genders of 28.7 points with X_{gender} coded as -1 and +1 for males and females respectively.

Figure 6 shows that the probability of response increased more rapidly in males than females with increasing BDI pre-treatment score. Also, Figure 6 shows that the predicted probability of response was lower in males for BDI scores below 29 points. An analysis[22] using the 95% confidence intervals for $B_{gender\ x\ severity}$ suggested that the probability of male response was significantly lower than that of females for BDI scores below 24 points but significantly higher than for females above 31 points. However, Figure 6 shows that the difference between the probability of recovery for equivalent scoring males and females was very much smaller for BDI scores above 31 points than it was below 24 points. Consequently, the probability of BDI response was practically equivalent between genders for pre-treatment scores of 24 points and above. Thus, females were significantly more likely to demonstrate a reliable post-treatment reduction in BDI score than equivalent scoring males for BDI pre-treatment severities less than 24 points. Finally, Nagelkerke's pseudo R^2 for the 2 step model in Table 29 was 0.24 indicating that the model accounted for 24% of the null deviance (Cohen et al., 2003).

Summary

The results showed that, overall, increasing BDI pre-treatment severity was significantly associated with an increased probability of response. However, a significant finding for the covariate study revealed that the overall response rate in Salminen et al. (2008) was lower than seen in remaining studies (David et al., 2008; Jacobson et al., 1996; Jarrett et al., 1999). Furthermore, the significant study x severity interaction revealed that, in contrast to remaining studies, the probability of response in Salminen et al., 2008 did not change as a function of BDI pre-treatment severity. Whilst the 1st order covariate gender was not a significant predictor of response, the significant gender x severity interaction revealed a that the predicted probability of response in males was significantly lower than that of equivalent scoring females for BDI pre-treatment severities below approximately 24 points.

[22] This did not account for error in $B_{severity}$ or B_{gender}

HRSD results

Table 30 presents the results of the binary logistic regression analysis for response in HRSD studies. The independent variables study, treatment type, gender and pre-treatment severity were regressed on Jacobson response status. Pre-treatment HRSD severity was centred using the overall mean of 20.3 points. Collinearity tests indicated that the regression analysis was appropriate as the lowest tolerance value for any of the terms in Table 30 was 0.72. Only the results for the 1st step analysis are presented as sparse data at step 2 led to the model being highly influenced by individual cases.

Table 30. Results of Logistic Regression Analysis for HRSD Response.

	Step χ^2	$\Delta R^{2\alpha}$	Variable	B	*p*	S.E.	Odds
Step 0			Constant	1.81	.001	0.15	6.1
Step 1	10.04	.05			.12		
			Constant	1.79	.001	0.19	6.00
			Study$^{\beta}$		.58		
			Treatment	- 0.13	.50	0.19	0.88
			Gender	0.08	.62	0.16	1.08
			Severity	0.09	.049	0.05	1.10

α: Nagelkerke's pseudo R^2

β: results for individual studies not shown.

Note: Effects coding employed for all categorical variables. Males, Salminen et al. (2008) and ADM served as baseline for Gender, Study and Treatment respectively.

The results in Table 30 show that the step 1 model was not significant ($\chi^2 = 10.04$, $p = .12$) and that pre-treatment severity was only just predictive of HRSD response (B = 0.09, $p = .049$). If the significance of the overall model is ignored, the results for pre-treatment severity alone indicate that the probability of response significantly increased with increasing pre-treatment HRSD severity. Finally, Nagelkerke's pseudo R^2 for the non-significant step 1 model in Table 30 was 0.05 indicating that the model accounted for 5% of the null deviance (Cohen et al., 2003).

7.4 Discussion

The primary aim of this study was to determine whether pre-treatment severity was a significant predictor of Jacobson recovery on either a self- or clinician-rated measure (BDI or HRSD respectively) across the IPD studies obtained for study 2. A secondary aim was to determine whether pre-treatment severity on either measure was predictive of Jacobson response. The results showed that, with the exception of males on the BDI, the probability of recovery significantly decreased with increasing pre-treatment severity. The probability of response was found to increase with increasing pre-treatment severity on both measures. The inclusion of gender as a covariate in analyses revealed that the probability of female recovery was significantly higher than that of males at lower pre-treatment severities on both the BDI and HRSD. The results of all analyses failed to provide evidence that ADM was superior to psychotherapy on either measure across any range of pre-treatment severity.

The BDI recovery analysis showed that, overall, pre-treatment severity was not predictive of recovery (B = -0.04, p = .09; Table 26). However, the identification of a significant gender x severity interaction (B = - 0.03, p = .02) revealed that the null overall finding for BDI pre-treatment severity occurred because BDI pre-treatment severity was not predictive of male recovery. In contrast, the probability of female recovery significantly reduced with increasing BDI pre-treatment severity (Figure 3). The BDI gender x severity interaction also indicated that females were significantly and increasingly more likely to recover than equivalent scoring males as BDI pre-treatment scores fell below 31 points (see Figure 3). This gender difference could have occurred because females typically just met the BDI recovery criterion of 8 points or less, whereas males typically just missed it. If so, then the relatively small clinical difference between genders according to continuous outcome data might have become statistically significant when analysed in terms of categorical outcomes. However, the results of the BDI response analysis (Table 29) indicated that this was not the case. The BDI response results showed that, overall, increased BDI pre-treatment severity predicted an increased probability of response (B = 0.14, p = .001). However, the significant gender x severity interaction identified within the BDI response analysis (B = -0.07, p = .03) revealed that female response was significantly more likely than that of equivalent scoring males for pre-treatment severities below 24 points (Figure 6). Thus, the probability of reliable symptomatic improvement on the BDI was significantly lower in males than females at lower pre-treatment severities. This meant that the significantly higher probability of female recovery at pre-treatment BDI severities below 31 points was due to lower levels of reliable improvement in males and not an artefact of categorical data analysis.

The HRSD recovery analysis showed that, overall, HRSD pre-treatment severity was predictive of recovery (B = -0.10, p = .001; Table 28). The probability of recovery on the HRSD significantly reduced with increasing pre-treatment severity in both males and females. However, because gender was also a significant predictor of recovery on the HRSD (B_{gender} = 0.26, p = .03), the probability of female recovery was significantly higher than that of males with equivalent pre-treatment HRSD scores (Figure 4). An analysis of the 95% confidence intervals for $B_{severity}$ indicated that the probability of female recovery was significantly higher than that of equivalent scoring males for HRSD pre-treatment scores between 16 to 24 points. The results of the HRSD response analysis showed that increased pre-treatment severity significantly predicted an increased probability of response (B = 0.09, p = .049), albeit within a non-significant model (Table 30). The unavailability of a valid step 2 model within the HRSD response analysis meant it was not possible to determine whether there was a significant gender x severity interaction.

The results for both the BDI and HRSD revealed that, when study and treatment type were controlled for, there was a range of pre-treatment severity where the probability of female recovery was significantly higher than that of equivalent scoring males. In the BDI completer sample used in analyses, 66.8% (n = 256) of patients scored 31 points or below, whilst 66.6% (n = 255) of the HRSD sample scored from 16 to 24 points at pre-treatment. Thus, for the majority of patients, females were significantly more likely to recover than equivalent scoring males on both measures. The reason for this finding cannot be determined on the basis of the data used in analyses. According to the Centre for Cognitive Studies a BDI score greater than 30 denotes severe depression (Beck et al., 1988). In terms of the HRSD, the National Institute for Health and Clinical Excellence (NICE) has recommended that a score of 23 or denotes severe depression (Kriston and von Wolff, 2011). Consequently, the results of this study indicate that the probability of recovery in males and females with identical pre-treatment scores on the BDI or HRSD will only be equivalent in the most severely depressed samples. This significant interaction of pre-treatment severity with gender has the potential to confound the results of individual treatment studies that employ the BDI or HRSD to assess outcome. In order to minimise this risk of bias in individual studies it is necessary to satisfy two conditions. Firstly, the male to female ratio of treatment groups should be closely matched. Secondly, patients need to be assigned to treatment type using stratification on pre-treatment severity.

However, these findings need to be interpreted in terms of the explicatory power of the logistic regression analyses. The Nagelkerke's pseudo R^2 values indicated that the null deviance in the BDI recovery analysis was found to be 10%, whilst that for the HRSD

recovery analysis was 8%. These low null deviance values indicated that the regression models for recovery, and hence the variables entered in recovery analyses, accounted for a relatively small proportion of the variability in outcome. Consequently, it appears that the variables study, treatment, pre-treatment severity and gender were not major determinants of recovery in treatment completers. Clearly, there are important variables predictive of recovery that could not be included in regression models. In contrast, within the BDI response analysis Nagelkerke's pseudo R^2 for the 2 step model was 0.24. This indicated that the same variables taken together played a more important role in predicting BDI response as they accounted for 24% of the overall variability in outcome. However, it is likely that the higher pseudo R^2 value for the BDI response analysis was due to the significantly lower overall response in Salminen et al. (2008) plus significant between-study differences in the probability of response with increasing pre-treatment severity (B_{study} and $B_{study\ x\ severity}$ respectively). The significant finding for $B_{study\ x\ severity}$ in particular indicate that patients within studies did not respond equivalently after controlling for pre-treatment severity, treatment type and gender. Whilst the nature of the analyses in this study mean that it is impossible to know the reasons for these differences, it is possible that studywise differences in treatment integrity, or the intensity/duration of treatment were responsible. Whatever the reason, the results indicate that controlling only for pre-treatment severity in pooled analyses of primary studies is unlikely to eliminate potential sources of bias.

Finally, the finding that females were more likely to benefit from treatment than equivalent scoring males contrasts with the recent review by Parker et al. (2011). Parker et al. (2011) reviewed 15 studies that had investigated whether gender significantly influences psychotherapy outcome. Whilst several studies failed to identify any significant gender effects, Parker et al. (2011) reported that those finding superior outcomes in females were counterbalanced by those finding superior outcomes in males. On the basis of what was effectively a vote-counting analysis (Hedges and Olkin, 1980), Parker et al. (2011) concluded there was insufficient evidence to argue that gender significantly influences psychotherapy outcome. That the statistically more powerful and flexible method of IPD meta-analysis was able to detect significant gender differences using fewer studies supports its future use for the identification of predictors of treatment outcome.

There are several limitations that apply to this study. The broad division of treatments into ADM or psychotherapy meant that it was not possible to investigate the relative efficacy of specific types of ADM or psychotherapy in analyses. However, this rough division of treatments is not uncommon in meta-analytic comparisons of treatments for depression and the use of IPD overcame some of the limitations applying to conventional meta-analyses.

Perhaps the most important limitation was that because IPD studies differed in terms of duration and the variables investigated, it was not possible to include additional covariates currently believed to moderate treatment outcome. For example, it was not possible to include chronicity (MDD > 2 years), age, marital status and comorbidity in analyses, all of which have been reported to moderate treatment outcome (Fournier et al., 2009; Jarrett et al., 1991; Rush et al., 2005). Thus, it is possible that the significantly higher probability of female recovery at lower pre-treatment severities was due to confounds arising from one or more of these factors. Finally, the absence of controls meant that it was not possible to be certain that the results applied to treatments with demonstrable efficacy (Klein, 2000). This concern is particularly pertinent concerning treatment outcomes in less severely depressed patients, as both the TDCRP study (Elkin et al., 1989) and a recent patient-level meta-analysis (Fournier et al., 2010) showed that treatments were superior to placebo only in more severely depressed samples.

7.5 Summary & Concluding Remarks

Increasing pre-treatment severity predicted a lower recovery rate for both genders on the HRSD and for females on the BDI. Thus, pre-treatment severity moderated treatment outcome assessed by the BDI and HRSD. Males and females appear to have a different recovery pattern, with recovery in females significantly more likely than in males of equivalent pre-treatment score in all but the most severely depressed samples. Thus, gender is a significant moderator of outcome on these measures which suggests that to avoid confounded results, future treatment studies will need to balance comparison groups in terms of their male to female ratio and stratify groups by pre-treatment severity.

No evidence was found to suggest that ADM was superior to psychotherapy across the range of pre-treatment severity on either measure. However, the amount of outcome variation explained by the variables included in recovery analyses was relatively small. Nevertheless, the fact that IPD meta-analysis was capable of identifying significant gender differences in addition to theoretically important interactions supports their use in future investigations of factors that predict treatment outcome. A more complete understanding of such factors should enable the development of more efficacious treatments and assist clinicians in deciding which treatments work best for specific patient types.

Chapter Eight

General Discussion & Conclusions

This thesis has examined several methodological issues which contribute to the variability in outcome typically observed between psychological treatment studies for major depression. Because meta-analysis is now widely used to summarise the findings of such studies, the starting point was to survey the results of meta-analytic reviews that had included methodologically rigorous studies. In addition, it is recognised that effect sizes based on continuous data can be of limited clinical utility, therefore several meta-analytic reviews compared treatment outcomes in terms of the proportion of patients achieving an outcome criterion: (e.g. Casacalenda et al., 2002). The most reliable evidence obtained from the reviews in study 1 showed that at most 48% of patients starting psychotherapy[23] remitted by post-treatment (de Maat et al., 2006; de Maat et al., 2007). However, there was evidence that approximately half of patients receiving a placebo achieve remission (Casacalenda et al., 2002). There was robust evidence that the relative efficacy of psychotherapy and medication were no different at post-treatment (Casacalenda et al., 2002; de Maat et al., 2006; Parker et al., 2008) and limited evidence that acute phase psychotherapy conferred superior protection against recurrence in comparison to acute medication (Vittengl et al., 2007). Nevertheless, 73% of patients who achieve remission following a course of psychotherapy experienced a new major depressive episode within three years (Vittengl et al., 2007). These results suggest that psychotherapy is a relatively ineffective prophylactic intervention against further episodes of depression, although it may be the case that the intervention increases the duration between discrete episodes. Nonetheless, there remains a clear need to develop interventions that can improve the short and long term efficacy of psychological treatments for depression.

An examination of substantive review data in study 1 revealed that all reviews were at risk of producing biased results due to methodological factors that increase the variability in outcome between primary studies. Two factors were identified which could not be directly examined in this thesis due to the unavailability of appropriate information. First amongst these was the finding that some of the studies included in reviews provided insufficient detail concerning the establishment of therapist competence or adherence to treatment protocols. This meant that the integrity of psychological treatment could not be guaranteed across all studies and raised the possibility that some of the 'psychotherapy' offered to

[23] Aggregated across several psychotherapy models.

patients may have been no more effective than sham treatment. If true, this bias would have served to reduce overall estimates for the efficacy of psychotherapy in meta-analyses (Matt and Navarro, 1997) and could potentially obscure its superiority to medication when adequately provided. However, following an in-depth review of seven influential Cognitive Behavioural Therapy (CBT) studies for depression, Roth et al. (2010) found that investigators typically devote much effort to therapist selection, training and monitoring during clinical trials. This suggests that psychological treatments in many of the studies included in reviews were likely to have been provided to a high standard and that the risk of bias from this factor was low. Nevertheless, it is impossible to judge the validity of the results of individual studies, nor meta-analyses based on them, without clear evidence that treatments were provided as intended. Thus, it is important that psychotherapy researchers do not continue to neglect this important issue when reporting empirical research (Perepletchikova, 2009).

The second factor which could not be directly examined was that the primary studies in study 1 varied considerably in terms of the overall duration of psychological treatment. Because evidence suggests that medication may act more rapidly to reduce depressive symptoms than psychological treatments (Watkins et al., 1993; Elkin et al., 1989), it is possible that the inclusion of very brief studies comparing psychotherapy with medication biased the overall results of reviews in favour of medication. However, Dekker et al. (2008) has suggested that the more rapid symptomatic reduction afforded by medication over psychotherapy occurs only during the first month of treatment. Consequently, it is unlikely that treatment duration served as a source of bias in comparisons of psychotherapy with medication in the reviews in study 1, as the minimum duration of any study was 8 weeks. Nevertheless, as increasing symptomatic benefit is associated with increasing treatment duration (Howard et al., 1986; Shapiro et al., 1994), it is to be expected that a higher proportion of patients will remit in psychotherapy studies of longer duration. Consequently, the inclusion of such studies in meta-analysis will serve to increase the variability in remission estimates on which analyses are based. This in turn will reduce the power of meta-analysis to identify any significant overall differences between treatments that may exist. In addition, it will be difficult to determine how many patients should be expected to remit following treatment of a specific duration. A potential solution to this problem is that clinical significance rates are published for standard assessment intervals during treatment in future studies. Thus, reviewers could include only outcome data for patients treated over a specific time interval in meta-analysis. However, this approach is itself likely to be problematic, as Shapiro et al. (1994) have shown that therapists may adjust the pace of psychotherapy to correspond with the time available.

The first methodological factor that could be directly investigated in terms of its contribution to outcome variability and the risk of bias in meta-analysis was the use of idiosyncratic clinical significance criteria in primary studies. In study 2 post-treatment recovery rates calculated using Jacobson method criteria for the Beck Depression Inventory (BDI, Beck et al., 1961) and Hamilton Rating Scale for Depression (HRSD, Hamilton, 1960) were compared with published clinical significance rates based on the same measures. The primary studies providing individual patient data (IPD) for comparisons in study 2 met the same eligibility criteria as those in study 1. Study 2 showed that the published criteria for IPD studies typically led to published clinical significance rates that were higher than those of the Jacobson method. It was also found that all patients who met published clinical significance criteria demonstrated reliable change according to Jacobson method criteria. This meant that published and Jacobson method criteria were effectively the same where studies used a score of 8 or less on the BDI, or a score of 7 or less on the HRSD to define remission. However, where published criteria were much less stringent than Jacobson criteria, published rates could be up to 55% higher than Jacobson rates (e.g. Constantino et al., 2008). The evident outcome variability due to the idiosyncratic definitions in primary studies makes it difficult to know how closely published rates represent the optimum outcome of remission (APA 2000). This difficulty is compounded when the results of such studies are pooled in meta-analysis. Consequently, where the goal of meta-analysis is to compare treatments in terms of remission, it is necessary that included studies have employed appropriate definitions. However, such an approach is problematic, as study 1 showed that where primary studies publish clinical significance rates for the BDI or HRSD, they are frequently based on different criteria to those of the Jacobson method.

Study 2 also revealed that the rank order of treatment efficacy is not invariant to changing definitions of clinical significance. Therefore, it is possible that the use of idiosyncratic outcome criteria may inadvertently bias the results of individual studies in favour of one type of treatment. Whilst such bias ought to be random and thus cancel out in meta-analyses that include many studies, there is no guarantee that this will occur (Matt and Navarro, 1997). Moreover, the results of study 1 showed that there were relatively few high quality primary studies available in this area, as the maximum number included in reviews was 10. This indicates that statistically influential primary studies have the potential to bias the overall results of methodologically rigorous meta-analyses of psychological treatments for major depression. Indeed, that individual studies may influence review conclusions was illustrated by de Maat et al. (2007), who reported that their overall conclusions applied only to patients in Keller et al. (2000) and not the overall sample included in analysis.

When outcomes were compared between the BDI and HRSD in the same sample, the results of study 2 revealed poor agreement between measures in terms of Jacobson recovery (Kappa = 0.56). As fewer patients recovered on the BDI than on the HRSD it was shown that recovery on the BDI was a more stringent test of recovery. However, comparisons between IPD and published results for Jacobson et al. (1996) indicated that the BDI provided insufficient coverage of DSM III depressive symptomatology. Taken together, these findings indicated that recovery according to Jacobson criteria on either measure was a less stringent test of a return to normative symptomatic levels than remission identified via diagnostic criteria. Therefore, a better operationalisation of remission in future treatment studies would be to require that patients demonstrate recovery on both the BDI and HRSD according to Jacobson method criteria. Finally, study 2 showed that approximately 50% of patients starting psychotherapy in IPD studies recovered according to Jacobson method criteria on the BDI or HRSD by post-treatment.

The last study of this thesis examined whether pre-treatment severity on the BDI or HRSD was a significant predictor of recovery defined according to Jacobson criteria. This was the final methodological factor identified in study 1 with the potential to increase between-study outcome variability and bias the results of meta-analysis. The results of the IPD meta-analyses in study 3 showed that (i) increased pre-treatment severity on the BDI or HRSD generally predicted a reduced probability of recovery, (ii) there was no evidence that ADM was superior to psychotherapy at any pre-treatment severity. However, an interaction between pre-treatment severity and gender revealed that increased BDI pre-treatment severity was predictive of recovery only in females. It is impossible to rule out that this latter finding was due to unknown confounds. Nevertheless, the findings of the BDI response analysis, that males were less likely to demonstrate reliable change than equivalently depressed females, supports Sigmon et al.'s assertion that gender biases on self-reported depression measures warrant further research (Sigmon et al., 2005). The results for both measures revealed that females were significantly more likely to recover than equivalently depressed males at all but the most severe levels of pre-treatment severity. This suggests that a gender bias may also exist for clinician rated measures. Taken together, the findings of study 3 suggest that gender differences may serve to bias the results of individual studies unless, (i) gender ratios are matched between treatment groups, (ii) males and females are independently randomised to treatment group according to pre-treatment severity.

In summary, several methodological factors serve to increase outcome variation typically observed between primary studies of psychological treatments for major depression. Some

of this variation is random error which serves to reduce the ability of meta-analyses to detect significant treatment differences (Wilson and Lipsey, 2001) and risks that results are biased. The factor likely to contribute the most random error to meta-analyses of depression treatment studies is the method by which outcome is defined (Wilson and Lipsey,2001). This implies that future meta-analyses will be better placed to compare the relative efficacy of psychological treatments for depression only when a standard and appropriate definition of clinical significance is universally accepted. However, the moderating effects of treatment integrity, treatment duration, pre-treatment severity and other unknown factors will still lead to variability in outcome between individual studies. These moderators will still serve to reduce the power and precision of meta-analysis and render their results difficult to interpret.

Given the clinical heterogeneity associated with a single diagnosis of major depression, it is necessary to understand whether some treatments work better for particular patient types. For example, de Maat et al. (2007) fortuitously found that combined therapy is more efficacious only in moderately depressed patients with chronic depression. Unfortunately, it is very unlikely that the limitations of meta-regression (Higgins et al., 2001) will enable a clear understanding of the role that moderating variables such as chronicity play in affecting treatment outcome, even should a standard outcome definition be employed across studies. However, the finding that gender significantly moderated outcome in the IPD meta-analyses of study 3 indicate that this approach has greater power and flexibility to do so. Therefore, in the broadest terms, the results of this thesis indicate that IPD meta-analyses are to be preferred over those using summary data, as the latter can provide only very limited information concerning the efficacy of psychological treatments for major depression.

References

Akdemir, A., Turkcapar, M. G., Orsel, S. D., Demirergi, N., Dag, I. & Ozbay, M. H. 2001. Reliability and validity of the Turkish version of the Hamilton depression rating scale. *Comprehensive Psychiatry,* 42**,** 161-165.

Akiskal, H. S. & Akiskal, K. K. 2007. In search of Aristotle: temperament, human nature, melancholia, creativity and eminence. *Journal of Affective Disorders,* 100**,** 1-6.

Alonso, J., Angermeyer, M. C., Bernert, S., Bruffaerts, R., Brugha, T. S., Bryson, H., De Girolamo, G., De Graaf, R., Demyttenaere, K., Gasquet, I., Haro, J. M., Katz, S. J., Kessler, R. C., Kovess, V., Lepine, J. P., Ormel, J., Polidori, G., Russo, L. J., Vilagut, G., Almansa, J., Arbabzadeh-Bouchez, S., Autonell, J., Bernal, M., Buist-Bouwman, M. A., Codony, M., Domingo-Salvany, A., Ferrer, M., Joo, S. S., Martinez-Alonso, M., Matschinger, H., Mazzi, F., Morgan, Z., Morosini, P., Palacin, C., Romera, B., Taub, N. & Vollebergh, W. A. M. 2004. Prevalence of mental disorders in Europe: Results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatrica Scandinavica,* 109**,** 21-27.

Alonso, J., Ferrer, M., Romera, B., Vilagut, G., Angermeyer, M., Bernert, S., Brugha, T. S., Taub, N., McColgen, Z., De Girolamo, G., Polidori, G., Mazzi, F., De Graaf, R., Vollebergh, W. A. M., Buist-Bowman, M. A., Demyttenaere, K., Gasquet, I., Haro, J. M., Palaca, N, C., Autonell, J., Katz, S. J., Kessler, R. C., Kovess, V., LePine, J. P., Arbabzadeh-Bouchez, S., Ormel, J. & Bruffaerts, R. 2002. The European study of the epidemiology of mental disorders (ESEMeD/MHEDEA 2000) project: Rationale and methods. *International Journal of Methods in Psychiatric Research,* 11**,** 55-67.

Alonso, J. & Lepine, J. P. 2007. Overview of key data from the European Study of the Epidemiology of Mental Disorders (ESEMeD). *Journal of Clinical Psychiatry,* 68**,** 3-9.

Andrews, G. & Harvey, R. 1981. Does psychotherapy benefit neurotic patients. *Archives of General Psychiatry,* 38**,** 1203-8.

Angst, J., Gamma, A., Benazzi, F., Ajdacic, V. & Rössler, W. 2007. Melancholia and atypical depression in the Zurich study: Epidemiology, clinical characteristics, course, comorbidity and personality. *Acta Psychiatrica Scandinavica,* 115**,** 72-84.

Ankuta, G. Y. & Abeles, N. 1993. Client Satisfaction, Clinical Significance, and Meaningful Change in Psychotherapy. *Professional Psychology: Research and Practice,* 24**,** 70-74.

American Psychiatric Association 1980. *Diagnostic and statistical manual of mental disorders: DSM-III,* Washington, American Psychiatric Association.

American Psychiatric Association 2000. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR,* Washington, American Psychiatric Association.

Atkins, D. C., Bedics, J. D., McGlinchey, J. B. & Beauchaine, T. P. 2005. Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology,* 73**,** 982-989.

Baer, D. M., Wolf, M. M. & Risley, T. R. 1968. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis,* 1**,** 91-&.

Barlow, D. H. 1981. On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology,* 49**,** 147-155.

Beautrais, A. L., Joyce, P. R., Mulder, R. T., Fergusson, D. M., Deavoll, B. J. & Nightingale, S. K. 1996. Prevalence and comorbidity of mental disorders in persons making serious suicide attempts: A case-control study. *American Journal of Psychiatry,* 153**,** 1009-1014.

Bebbington, P., Dunn, G., Jenkins, R., Lewis, G., Brugha, T., Farrell, M. & Meltzer, H. 2003. The influence of age and sex on the prevalence of depressive conditions: Report from the National Survey of Psychiatric Morbidity. *International Review of Psychiatry,* 15**,** 74-83.

Beck, A. T., Erbaugh, J., Ward, C. H., Mock, J. & Mendelsohn, M. 1961. An Inventory for Measuring Depression. *Archives of General Psychiatry,* 4**,** 561-571.

Beck, A. T., Rush, A. J., Shaw, B. F. & Emery, G. 1979. *Cognitive Therapy of Depression,* New York, Guilford Press.

Beck, A. T., Steer, R. A. & Brown, G. K. 1996. *Manual for the Beck Depression Inventory - second edition,* San Antonio Texas, The Psychological Corporation.

Beck, A. T., Steer, R. A. & Garbin, M. G. 1988. Psychometric Properties of the Beck Depression Inventory - 25 Years of Evaluation. *Clinical Psychology Review,* 8**,** 77-100.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D. & Stroup, D. F. 1996. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association,* 276**,** 637-639.

Bergin, A. E. 1966. Some implications of psychotherapy research for therapeutic practice. *Journal of Abnormal Psychology,* 71**,** 235-246.

Berman, J. S., Miller, R. C. & Massman, P. J. 1985. Cognitive Therapy Versus Systematic-Desensitization - Is One Treatment Superior. *Psychological Bulletin,* 97**,** 451-461.

Bernal, M., Haro, J. M., Bernert, S., Brugha, T., De Graaf, R., Bruffaerts, R., LePine, J. P., De Girolamo, G., Vilagut, G., Gasquet, I., Torres, J. V., Kovess, V., Heider, D., Neeleman, J., Kessler, R. & Alonso, J. 2007. Risk factors for suicidality in Europe: Results from the ESEMED study. *Journal of Affective Disorders,* 101**,** 27-34.

Bernert, S., Matschinger, H., Alonso, J., Haro, J. M., Brugha, T. S. & Angermeyer, M. C. 2009. Is it always the same? Variability of depressive symptoms across six European countries. *Psychiatry Research,* 168**,** 137-144.

Bhar, S. S. & Beck, A. T. 2009. Treatment Integrity of Studies That Compare Short-Term Psychodynamic Psychotherapy with Cognitive-Behavior Therapy. *Clinical Psychology: Science and Practice,* 16**,** 370-378.

Bigos, K. L., Pollock, B. G., Stankevich, B. A. & Bies, R. R. 2009. Sex differences in the pharmacokinetics and pharmacodynamics of antidepressants: An updated review. *Gender Medicine,* 6**,** 522-543.

Birnbaum, H. G., Kessler, R. C., Kelley, D., Ben-Hamadi, R., Joish, V. N. & Greenberg, P. E. 2010. Employer burden of mild, moderate, and severe major depressive disorder: Mental health services utilization and costs, and work performance. *Depression and Anxiety,* 27**,** 78-89.

Blackburn, I. M., Bishop, S., Glen, A. I. M., Whalley, L. J. & Christie, J. E. 1981. The Efficacy of Cognitive Therapy in Depression - a Treatment Trial Using Cognitive Therapy and Pharmacotherapy, Each Alone and in Combination. *British Journal of Psychiatry,* 139**,** 181-189.

Boissel, J. P., Blanchard, J., Panak, E., Peyrieux, J. C. & Sacks, H. 1989. Considerations for the Metaanalysis of Randomized Clinical-Trials - Summary of a Panel Discussion. *Controlled Clinical Trials,* 10**,** 254-281.

Boisvert, C. M. & Faust, D. 2006. Practicing psychologists' knowledge of general psychotherapy research findings: Implications for science-practice relations. *Professional Psychology: Research and Practice,* 37**,** 708-716.

Boland, R. J. & Keller, M. B. 2008. Course and Outcome of Depression. *In:* Gotlib, I. H. & Hammen, C. L. (eds.) *Handbook of depression.* 2nd ed. New York: Guilford Press.

Boughton, S. & Street, H. 2007. Integrated review of the social and psychological gender differences in depression. *Australian Psychologist,* 42**,** 187-197.

Carter, G. C., Cantrell, R. A., Zarotsky, V., Haynes, V. S., Phillips, G., Alatorre, C. I., Goetz, I., Paczkowski, R. & Marangell, L. B. 2012. Comprehensive review of factors implicated in the heterogeneity of response in depression. *Depression and Anxiety,* 29**,** 340-354.

Casacalenda, N., Perry, J. C. & Looper, K. 2002. Remission in major depressive disorder: a comparison of pharmacotherapy, psychotherapy, and control conditions. *American Journal of Psychiatry***,** 1354-1360.

Centre for Reviews and Dissemination 2009. *Systematic Reviews: CRD's guidance for undertaking reviews in health care,* York, University of York.

Centre for Reviews and Dissemination 2009a. *University of York's Centre for Reviews and Dissemination's abstract reporting format,* Available: http://www.crd.york.ac.uk/crdweb/Home.aspx?DB=DARE [Accessed 25/11/2009].

Chambless, D. L. & Hollon, S. D. 1998. Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology,* 66**,** 7-18.

Christensen, L. & Mendoza, J. L. 1986. A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy,* 17**,** 305-308.

Churchill, R., Hunot, V., Corney, R., Knapp, M., McGuire, H., Tylee, A. & Wessely, S. 2001. A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression. *Health Technology Assessment,* 5**,** 1-173.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences,* New Jersey, Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* New Jersey, Lawrence Erlbaum Associates.

Constantino, M. J., Marnell, M. E., Haile, A. J., Kanther-Sista, S. N., Wolman, K., Zappert, L. & Arnow, B. A. 2008. Integrative cognitive therapy for depression: A randomized pilot comparison. *Psychotherapy,* 45**,** 122-134.

Coryell, W., Endicott, J., Andreasen, N. C., Keller, M. B., Clayton, P. J., Hirschfeld, R. M. A., Scheftner, W. A. & Winokur, G. 1988. Depression and panic attacks: The significance of overlap as reflected in follow-up and family study data. *American Journal of Psychiatry,* 145**,** 293-300.

Coyne, J. C. 1994. Self-reported distress - analog or ersatz depression. *Psychological Bulletin,* 116**,** 29-45.

Cristancho, M. A., O'reardon, J. P. & Thase, M. E. 2011. Atypical depression in the 21st century: Diagnostic and treatment issues. *Psychiatric Times,* 28**,** 42-46.

Cuijpers, P., Van Straten, A., Warmerdam, L., Andersson, G., Cuijpers, P., Van Straten, A., Warmerdam, L. & Andersson, G. 2008. Psychological treatment of depression: a meta-analytic database of randomized studies. *BMC Psychiatry,* 8**,** 36.

Daly, R. W. 2007. Before depression: the medieval vice of acedia. *Psychiatry,* 70**,** 30-51.

David, D., Szentagotai, A., Lupu, V. & Cosman, D. 2008. Rational emotive behavior therapy, cognitive therapy, and medication in the treatment of major depressive disorder: a randomized clinical trial, posttreatment outcomes, and six-month follow-up. *J Clin Psychol,* 64**,** 728-46.

Davila, J., Stroud, C. B. & Starr, L. R. 2008. Depression in Couples and Families. *In:* Gotlib, I. H. & Hammen, C. L. (eds.) *Handbook of depression.* 2nd ed. New York: Guilford Press.

De Maat, S. M., Dekker, J., Schoevers, R. & De Jonghe, F. 2006. Relative efficacy of psychotherapy and pharmacotherapy in the treatment of depression: A meta-analysis. *Psychotherapy Research,* 16**,** 562-572.

De Maat, S. M., Dekker, J., Schoevers, R. A. & De Jonghe, F. 2007. Relative efficacy of psychotherapy and combined therapy in the treatment of depression: a meta-analysis. *European Psychiatry: the Journal of the Association of European Psychiatrists,* 22**,** 1-8.

Decker, H. S. 2007. How Kraepelinian was Kraepelin? How Kraepelinian are the neo-Kraepelinians? - from Emil Kraepelin to DSM-III. *History of Psychiatry,* 18**,** 337-360.

Dekker, J. J. M., Koelen, J. A., Van, H. L., Schoevers, R. A., Peen, J., Hendriksen, M., Kool, S., Van Aalst, G. & De Jonghe, F. 2008. Speed of action: The relative efficacy of short psychodynamic supportive psychotherapy and pharmacotherapy in the first 8 weeks of a treatment algorithm for depression. *Journal of Affective Disorders,* 109**,** 183-188.

Derubeis, R. J., Gelfand, L. A., Tang, T. Z. & Simons, A. D. 1999. Medications versus cognitive behavior therapy for severely depressed outpatients: mega-analysis of four randomized comparisons. *American Journal of Psychiatry,* 156**,** 1007-13.

Derubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., O'reardon, J. P., Lovett, M. L., Gladis, M. M., Brown, L. L. & Gallop, R. 2005. Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry,* 62**,** 409-416.

Dobson, K. S. 1989. A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology,* 57**,** 414-419.

Driessen, E., Cuijpers, P., Hollon, S. D. & Dekker, J. J. M. 2010. Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. *Journal of Consulting and Clinical Psychology,* 78**,** 668-680.

Elkin, I., Gibbons, R. D., Shea, M. T., Sotsky, S. M., Watkins, J. T., Pilkonis, P. A. & Hedeker, D. 1995. Initial Severity and Differential Treatment Outcome in the National Institute of Mental-Health Treatment of Depression Collaborative Research-Program. *Journal of Consulting and Clinical Psychology,* 63**,** 841-847.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J. & Parloff, M. B. 1989. National-Institute-of-Mental-Health Treatment of Depression Collaborative Research-Program - General Effectiveness of Treatments. *Archives of General Psychiatry,* 46**,** 971-982.

Eysenck, H. J. 1952. The effects of psychotherapy: an evaluation. *Journal of Consulting Psychology,* 16**,** 319-24.

Field, A. P. 2003. The Problem in Using Fixed-Effects Models of Meta-Analysis on Real-World Data. *Understanding Statistics,* 2**,** 105.

Fink, M., Bolwig, T. G., Parker, G. & Shorter, E. 2007. Melancholia: Restoration in psychiatric classification recommended. *Acta Psychiatrica Scandinavica,* 115**,** 89-92.

Fleischmann, A., Bertolote, J. M., Belfer, M. & Beautrais, A. 2005. Completed suicide and psychiatric diagnoses in young people: A critical examination of the evidence. *American Journal of Orthopsychiatry,* 75**,** 676-683.

Follette, W. C. & Callaghan, G. M. 1996. The importance of the principle of clinical significance-defining significant to whom and for what purpose: A response to Tevgey, Lambert, Burlingame, Hansen. *Psychotherapy Research,* 6**,** 133-143.

Fournier, J. C., Derubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C. & Fawcett, J. 2010. Antidepressant Drug Effects and Depression Severity A Patient-Level Meta-analysis. *Jama-Journal of the American Medical Association,* 303**,** 47-53.

Fournier, J. C., Derubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D. & Gallop, R. 2009. Prediction of Response to Medication and Cognitive Therapy in the Treatment of Moderate to Severe Depression. *Journal of Consulting and Clinical Psychology,* 77**,** 775-787.

Frank, E., Cassano, G. B., Rucci, P., Thompson, W. K., Kraemer, H. C., Fagiolini, A., Maggi, L., Kupfer, D. J., Shear, M. K., Houck, P. R., Calugi, S., Grochocinski, V. J., Scocco, P., Buttenfield, J. & Forgione, R. N. 2011. Predictors and moderators of time to remission of major depression with interpersonal psychotherapy and SSRI pharmacotherapy. *Psychological Medicine,* 41**,** 151-162.

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., Rush, A. J. & Weissman, M. M. 1991. Conceptualization and Rationale For Consensus Definitions of Terms in Major Depressive Disorder - Remission, Recovery, Relapse and Recurrence. *Archives of General Psychiatry,* 48**,** 851-855.

Frasure-Smith, N. & Lesperance, F. 2005. Reflections on depression as a cardiac risk factor. *Psychosomatic Medicine,* 67**,** S19-S25.

Frazier, P. A., Tix, A. P. & Barron, K. E. 2004. Testing Moderator and Mediator Effects in Counseling Psychology Research. *Journal of Counseling Psychology,* 51**,** 115-134.

Friedman, M. A., Detweiler-Bedell, J. B., Leventhal, H. E., Horne, R., Keitner, G. I. & Miller, I. W. 2004. Combined psychotherapy and pharmacotherapy for the treatment of major depressive disorder. *Clinical Psychology: Science and Practice,* 11**,** 47-68.

Garfield, S. L. 1981. Psychotherapy: A 40-year appraisal. *American Psychologist,* 36**,** 174-183.

Gloaguen, V., Cottraux, J., Cucherat, M. & Blackburn, I. M. 1998. A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders,* 49**,** 59-72.

Goldberg, D. 2006. The "NICE Guideline" on the treatment of depression. *Epidemiologia e Psichiatria Sociale,* 15**,** 11-15.

Hamilton, K. E. & Dobson, K. S. 2002. Cognitive therapy of depression: Pretreatment patient predictors of outcome. *Clinical Psychology Review,* 22**,** 875-893.

Hamilton, M. 1960. A Rating Scale for Depression. *Journal of Neurology Neurosurgery and Psychiatry,* 23**,** 56-62.

Hardy, R. J. & Thompson, S. G. 1998. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine,* 17**,** 841-856.

Hatzenbuehler, L. C., Parpal, M. & Matthews, L. 1983. Classifying college students as depressed or nondepressed using the Beck depression inventory: An empirical analysis. *Journal of Consulting and Clinical Psychology,* 51**,** 360-366.

Hautzinger, M., De Jong-Meyer, R., Treiber, R., Rudolf, G. A. E. & Thien, U. 1996. Efficacy of cognitive behavior therapy, pharmacotherapy, and the combination of both in non-melancholic, unipolar depression. *Zeitschrift Fur Klinische Psychologie-Forschung Und Praxis,* 25**,** 130-145.

Hedges, L. V. & Olkin, I. 1980. Vote-Counting Methods in Research Synthesis. *Psychological Bulletin,* 88**,** 359-369.

Herceg-Baron, R. L., Prusoff, B. A., Weissman, M. M., Dimascio, A., Neu, C. & Klerman, G. L. 1979. Pharmacotherapy and Psychotherapy in Acutely Depressed-Patients - Study of Attrition Patterns in a Clinical-Trial. *Comprehensive Psychiatry,* 20**,** 315-325.

Herrmann, C., Brand-Driehorst, S., Kaminsky, B., Leibing, E., Staats, H. & Ruger, U. 1998. Diagnostic groups and depressed mood as predictors of 22-month mortality in medical inpatients. *Psychosomatic Medicine,* 60**,** 570-7.

Higgins, J. P. T. & Green, S. (eds.) 2006. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]*, Chichester UK, John Wiley & Sons, Ltd. .

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z. & Thompson, S. G. 2001. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine,* 20**,** 2219-2241.

Hiller, W., Schindler, A. C. & Lambert, M. J. 2012. Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. *Psychotherapy Research,* 22**,** 1-11.

Hollon, S. D., Derubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M. & Tuason, V. B. 1992. Cognitive Therapy and Pharmacotherapy for Depression - Singly and in Combination. *Archives of General Psychiatry,* 49**,** 774-781.

Hollon, S. D. & Flick, S. N. 1988. On the meaning and methods of clinical significance. *Behavioral Assessment,* 10**,** 197-206.

Howard, K. I., Kopta, S. M., Krause, M. S. & Orlinsky, D. E. 1986. The Dose-Effect Relationship in Psychotherapy. *American Psychologist,* 41**,** 159-164.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z. & Lutz, W. 1996. Evaluation of psychotherapy - Efficacy, effectiveness, and patient progress. *American Psychologist,* 51**,** 1059-1064.

Hsu, L. M. 1989. Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment,* 11**,** 459-467.

Hugdahl, K. & Ost, L. G. 1981. On the difference between statistical and clinical significance. *Behavioral Assessment,* 3**,** 289-295.

Jackson, S. W. 1981. Acedia the sin and its relationship to sorrow and melancholia in medieval times. *Bulletin of the history of medicine,* 55**,** 172-181.

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., Gortner, E. & Prince, S. E. 1996. A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting & Clinical Psychology,* 64**,** 295-304.

Jacobson, N. S., Follette, W. C. & Revenstorf, D. 1984. Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy,* 15**,** 336-352.

Jacobson, N. S. & Hollon, S. D. 1996. Prospects for future comparisons between drugs and psychotherapy: Lessons from the CBT-versus-pharmacotherapy exchange. *Journal of Consulting and Clinical Psychology,* 64**,** 104-108.

Jacobson, N. S. & Revenstorf, D. 1988. Statistics for Assessing the Clinical-Significance of Psychotherapy Techniques - Issues, Problems, and New Developments. *Behavioral Assessment,* 10**,** 133-145.

Jacobson, N. S., Roberts, L. J., Berns, S. B. & McGlinchey, J. B. 1999. Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology,* 67**,** 300-307.

Jacobson, N. S. & Truax, P. 1991. Clinical-significance - a statistical approach to defining meaningful change in psychotherapy-research. *Journal of Consulting and Clinical Psychology,* 59**,** 12-19.

Jacobson, N. S., Wilson, L. & Tupper, C. 1988. The Clinical-Significance of Treatment Gains Resulting from Exposure-Based Interventions for Agoraphobia - a Reanalysis of Outcome Data. *Behavior Therapy,* 19**,** 539-554.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J. & McQuay, H. J. 1996. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials,* 17**,** 1-12.

Jarrett, R. B., Eaves, G. G., Grannemann, B. D. & Rush, A. J. 1991. Clinical, cognitive, and demographic predictors of response to cognitive therapy for depression: A preliminary report. *Psychiatry Research,* 37**,** 245-260.

Jarrett, R. B., Schaffer, M., McIntire, D., Witt-Browder, A., Kraft, D. & Risser, R. C. 1999. Treatment of atypical depression with cognitive therapy or phenelzine - A double-blind, placebo-controlled trial. *Archives of General Psychiatry,* 56**,** 431-437.

Katz, M. M., Secunda, S. K., Hirschfeld, R. M. A. & Koslow, S. H. 1979. NIMH Clinical Research Branch Collaborative Program on the Psychobiology of Depression. *Archives of General Psychiatry,* 36**,** 765-771.

Kazdin, A. & Kazdin, A. E. 1977. Assessing the Clinical or Applied Importance of Behavior Change through Social Validation. *Behavior Modification,* 1**,** 427-452.

Keller, M. B. 2003. Past, Present, and Future Directions for Defining Optimal Treatment Outcome in Depression: Remission and Beyond. *JAMA,* 289**,** 3152-3160.

Keller, M. B., Lavori, P. W., Lewis, C. E. & Klerman, G. L. 1983. Predictors of relapse in major depressive disorder. *Journal of the American Medical Association,* 250**,** 3299-3304.

Keller, M. B., McCullough, J. P., Klein, D. N., Arnow, B., Dunner, D. L., Gelenberg, A. J., Markowitz, J. C., Nemeroff, C. B., Russell, J. M., Thase, M. E., Trivedi, M. H., Zajecka, J., Blalock, J. A., Borian, F. E., Jody, D. N., Debattista, C., Koran, L. M., Schatzberg, A. F., Fawcett, J., Hirschfeld, R. M. A., Keitner, G., Miller, I., Kocsis, J. H., Kornstein, S. G., Manber, R., Ninan, P. T., Rothbaum, B., Rush, A. J., Vivian, D. & Rothbaum, B. 2000. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *New England Journal of Medicine,* 342**,** 1462-1470.

Kessler, R., White, L. A., Birnbaum, H., Qiu, Y., Kidolezi, Y., Mallett, D. & Swindle, R. 2008. Comparative and interactive effects of depression relative to other health problems on work performance in the workforce of a large employer. *Journal of Occupational and Environmental Medicine,* 50**,** 809-816.

Kessler, R. C. 1999. The World Health Organization International Consortium in Psychiatric Epidemiology (ICPE): Initial work and future directions - The NAPE lecture 1998. *Acta Psychiatrica Scandinavica,* 99**,** 2-9.

Kessler, R. C. 2007. The global burden of anxiety and mood disorders: Putting the European Study of the Epidemiology of Mental Disorders (ESEMeD) findings into perspective. *Journal of Clinical Psychiatry,* 68**,** 10-19.

Kessler, R. C., Angermeyer, M., Anthony, J. C., De Graaf, R., Demyttenaere, K., Gasquet, I., De Girolamo, G., Guzman, S., Gureje, O., Haro, J. M., Kawakami, N., Karam, A., Levinson, D., Mora, M. E. M., Browne, M. a. O., Posada-Villa, J., Stein, D. J., Tsang, C. H. A., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Heeringa, S., Pennell, B. E., Berglund, P., Gruber, M. J., Petukhova, M., Chatterji, S. & Ustun, T. B. 2007. Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative. *World Psychiatry*, 6**,** 168-176.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., Wang, P. S. & National Comorbidity Survey, R. 2003. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA,* 289**,** 3095-105.

Kessler, R. C., Birnbaum, H. G., Shahly, V., Bromet, E., Hwang, I., McLaughlin, K. A., Sampson, N., Andrade, L. H., De Girolamo, G., Demyttenaere, K., Haro, J. M., Karam, A. N., Kostyuchenko, S., Kovess, V., Lara, C., Levinson, D., Matschinger, H., Nakane, Y., Browne, M. O., Ormel, J., Posada-Villa, J., Sagar, R. & Stein, D. J. 2010. Age differences in the prevalence and co-morbidity of DSM-IV major depressive episodes: Results from the WHO world mental health survey initiative. *Depression and Anxiety,* 27**,** 351-364.

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G. & Nelson, C. B. 1993. Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders,* 29**,** 85-96.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U. & Kendler, K. S. 1994. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry,* 51**,** 8-19.

Kessler, R. C. & Ustun, T. B. 2004. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International

Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research,* 13**,** 93-121.

Kessler, R. C. & Wang, P. S. 2008. Epidemiology of Depression. *In:* GOTLIB, I. H. & HAMMEN, C. L. (eds.) *Handbook of depression.* 2nd ed. New York: Guilford Press.

Kiesler, D. J. 1966. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin,* 65**,** 110-136.

Klein, D. F. 1996. Preventing hung juries about therapy studies. *Journal of Consulting and Clinical Psychology,* 64**,** 81-87.

Klein, D. F. 2000. Flawed meta-analyses comparing psychotherapy with pharmacotherapy. *American Journal of Psychiatry,* 157**,** 1204-1211.

Kriston, L. & Von Wolff, A. 2011. Not as golden as standards should be: Interpretation of the Hamilton Rating Scale for Depression. *Journal of Affective Disorders,* 128**,** 175-177.

Krueger, R. F., Watson, D. & Barlow, D. H. 2005. Introduction to the special section: Toward a dimensionally based taxonomy of psychopathology. *Journal of Abnormal Psychology,* 114**,** 491-493.

Lambert, M. J., Hatch, D. R., Kingston, M. D. & Edwards, B. C. 1986. Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: a meta-analytic comparison. *J Consult Clin Psychol,* 54**,** 54-9.

Lambert, M. J. & Ogles, B. M. 2009. Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research,* 19**,** 493-501.

LaMothe, R. 2007. An analysis of acedia. *Pastoral Psychology,* 56**,** 15-30.

LeCrubier, Y. 2007. Widespread underrecognition and undertreatment of anxiety and mood disorders: Results from 3 European studies. *Journal of Clinical Psychiatry,* 68**,** 36-41.

Leichsenring, F. 2001. Comparative effects of short-term psychodynamic psychotherapy and cognitive-behavioral therapy in depression: A meta-analytic approach. *Clinical Psychology Review,* 21**,** 401-419.

Lesperance, F., Frasure-Smith, N., Talajic, M. & Bourassa, M. G. 2002. Five-year risk of cardiac mortality in relation to initial severity and one-year changes in depression symptoms after myocardial infarction. *Circulation,* 105**,** 1049-53.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J. & Moher, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *British Medical Journal,* 339**,** 1-28.

Lovaas, O. I. 1993. The development of a treatment-research project for developmentally disabled and autistic children. *Journal of Applied Behavior Analysis,* 26**,** 617-630.

Lunnen, K. M. & Ogles, B. M. 1998. A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology,* 66**,** 400-410.

Lux, V., Aggen, S. H. & Kendler, K. S. 2010. The DSM-IV definition of severity of major depression: inter-relationship and validity. *Psychological Medicine,* 40**,** 1691-1701.

Lynch, D., Laws, K. R. & McKenna, P. J. 2010. Cognitive behavioural therapy for major psychiatric disorder: does it really work? A meta-analytical review of well-controlled trials. *Psychological Medicine,* 40**,** 9-24.

Martinovich, Z., Saunders, S. & Howard, K. I. 1996. Some comments on "assessing clinical significance". *Psychotherapy Research,* 6**,** 124-132.

Matt, G. E. & Navarro, A. M. 1997. What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. *Clinical Psychology Review,* 17**,** 1-32.

McLean, P. D. & Hakstian, A. R. 1979. Clinical depression: comparative efficacy of outpatient treatments. *J Consult Clin Psychol,* 47**,** 818-36.

Melartin, T., Leskelä, U., Rystälä, H., Sokero, P., Lestelä-Mielonen, P. & Isometsä, E. 2004. Co-morbidity and stability of melancholic features in DSM-IV major depressive disorder. *Psychological Medicine,* 34**,** 1443-1452.

Moncrieff, J., Churchill, R., Drummond, C. D. & McGuire, H. 2001. Development of a quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research,* 10**,** 126-133.

Morrow, J. & Nolen-Hoeksema, S. 1990. Effects of Responses to Depression on the Remediation of Depressive Affect. *Journal of Personality and Social Psychology,* 58**,** 519-527.

Mueller, T. I., Lavori, P. W., Keller, M. B., Swartz, A., Warshaw, M., Hasin, D., Coryell, W., Endicott, J., Rice, J. & Akiskal, H. 1994. Prognostic effect of the variable course of alcoholism on the 10-year course of depression. *American Journal of Psychiatry,* 151**,** 701-706.

Mullen, P. D. & Ramirez, G. 2006. The promise and pitfalls of systematic reviews. *Annual Review of Public Health,* 27, 81-102.

Murphy, G. E., Carney, R. M., Knesevich, M. A., Wetzel, R. D. & Whitworth, P. 1995. Cognitive-Behavior Therapy, Relaxation Training, and Tricyclic Antidepressant Medication in the Treatment of Depression. *Psychological Reports,* 77**,** 403-420.

Murphy, G. E., Simons, A. D., Wetzel, R. D. & Lustman, P. J. 1984. Cognitive Therapy and Pharmacotherapy - Singly and Together in the Treatment of Depression. *Archives of General Psychiatry,* 41**,** 33-41.

National Institute for Health & Clinical Excellence 2009. Depression: the treatment and management of depression in adults [CG 90].

Niessen, L. W., Grijseels, E. W. M. & Rutten, F. F. H. 2000. The evidence-based approach in health policy and health care delivery. *Social Science and Medicine,* 51**,** 859-869.

Nolen-Hoeksema, S. 1987. Sex Differences in Unipolar Depression: Evidence and Theory. *Psychological Bulletin,* 101**,** 259-282.

Nolen-Hoeksema, S. 2000. The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology,* 109**,** 504-511.

Nolen-Hoeksema, S. 2012. Emotion regulation and psychopathology: The role of gender.

Nolen-Hoeksema, S. & Aldao, A. 2011. Gender and age differences in emotion regulation strategies and their relationship to depressive symptoms. *Personality and Individual Differences,* 51**,** 704-708.

Nolen-Hoeksema, S., McBride, A. & Larson, J. 1997. Rumination and psychological distress among bereaved partners. *Journal of Personality and Social Psychology,* 72**,** 855-862.

Nolen-Hoeksema, S., Morrow, J. & Fredrickson, B. L. 1993. Response Styles and the Duration of Episodes of Depressed Mood. *Journal of Abnormal Psychology,* 102**,** 20-28.

Nugent, W. R. 2006. The Comparability of the Standardized Mean Difference Effect Size Across Different Measures of the Same Construct. *Educational and Psychological Measurement,* 66**,** 612-623.

Nugent, W. R. 2009. Construct validity invariance and discrepancies in meta-analytic effect sizes based on different measures: A simulation study. *Educational and Psychological Measurement,* 69**,** 62-78.

Ogles, B. M., Lambert, M. J. & Sawyer, J. D. 1995. Clinical significance of the National Institute of Mental Health treatment of depression collaborative research program data. *Journal of Consulting and Clinical Psychology,* 63**,** 321-326.

Ogles, B. M., Lunnen, K. M. & Bonesteel, K. 2001. Clinical significance: History, application, and current practice. *Clinical Psychology Review,* 21**,** 421-446.

Onder, G., Liperoti, R., Soldato, M., Cipriani, M. C., Bernabei, R. & Landi, F. 2007. Depression and risk of nursing home admission among older adults in home care in Europe: Results from the Aged in Home Care (AdHOC) study. *Journal of Clinical Psychiatry,* 68**,** 1392-1398.

Oxman, A. D. 2004. Grading quality of evidence and strength of recommendations. *British Medical Journal,* 328**,** 1490-1494.

Parker, G., Blanch, B. & Crawford, J. 2011. Does gender influence response to differing psychotherapies by those with unipolar depression? *Journal of Affective Disorders,* 130**,** 17-20.

Parker, G. B., Crawford, J. & Hadzi-Pavlovic, D. 2008. Quantified superiority of cognitive behaviour therapy to antidepressant drugs: a challenge to an earlier meta-analysis.[see comment]. *Acta Psychiatrica Scandinavica,* 118**,** 91-7.

Perepletchikova, F. 2009. Treatment Integrity and Differential Treatment Effects. *Clinical Psychology: Science and Practice,* 16**,** 379-382.

Perepletchikova, F. & Kazdin, A. E. 2005. Treatment Integrity and Therapeutic Change: Issues and Research Recommendations. *Clinical Psychology: Science and Practice,* 12**,** 365-383.

Perepletchikova, F., Treat, T. A. & Kazdin, A. E. 2007. Treatment Integrity in Psychotherapy Research: Analysis of the Studies and Examination of the Associated Factors. *Journal of Consulting and Clinical Psychology,* 75**,** 829-841.

Pilling, S. 2008. History, context, process, and rationale for the development of clinical guidelines. *Psychology and Psychotherapy: Theory, Research and Practice,* 81**,** 331-350.

Posternak, M. A. & Miller, I. 2001. Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders,* 66**,** 139-146.

Posternak, M. A., Solomon, D. A., Leon, A. C., Mueller, T. I., Shea, M. T., Endicott, J. & Keller, M. B. 2006. The naturalistic course of unipolar major depression in the absence of somatic therapy. *Journal of Nervous and Mental Disease,* 194**,** 324-329.

Rachman, S. 1971. Obsessional ruminations. *Behaviour Research and Therapy,* 9**,** 229-235.

Robins, L. N., Helzer, J. E., Croughan, J. & Ratcliff, K. S. 1981. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry,* 38**,** 381-389.

Robinson, L. A., Berman, J. S. & Neimeyer, R. A. 1990. Psychotherapy for the treatment of depression: a comprehensive review of controlled outcome research. *Psychological Bulletin,* 108**,** 30-49.

Romera, I., Pérez, V., Menchón, J. M., Polavieja, P. & Gilaberte, I. 2011. Optimal cutoff point of the Hamilton Rating Scale for Depression according to normal levels of social and occupational functioning. *Psychiatry Research,* 186**,** 133-137.

Roth, A. D., Pilling, S. & Turner, J. 2010. Therapist Training and Supervision in Clinical Trials: Implications for Clinical Practice. *Behavioural and cognitive psychotherapy,* 38**,** 291-302.

Rush, A. J. 2007. The varied clinical presentations of major depressive disorder. *Journal of Clinical Psychiatry,* 68**,** 4-10.

Rush, A. J., Beck, A. T., Kovacs, M. & Hollon, S. 1977. Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research,* 1**,** 17-37.

Rush, A. J., Kraemer, H. C., Sackeim, H. A., Fava, M., Trivedi, M. H., Frank, E., Ninan, P. T., Thase, M. E., Gelenberg, A. J., Kupfer, D. J., Regier, D. A., Rosenbaum, J. F., Ray, O. & Schatzberg, A. F. 2006. Report by the ACNP Task Force on Response and Remission in Major Depressive Disorder. *Neuropsychopharmacology,* 31**,** 1841-1853.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., Thase, M. E., Kocsis, J. H. & Keller, M. B. 2003. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry,* 54**,** 573-583.

Rush, A. J., Zimmerman, M., Wisniewski, S. R., Fava, M., Hollon, S. D., Warden, D., Biggs, M. M., Shores-Wilson, K., Shelton, R. C., Luther, J. F., Thomas, B. & Trivedi, M. H. 2005. Comorbid psychiatric disorders in depressed outpatients: Demographic and clinical features. *Journal of Affective Disorders,* 87**,** 43-55.

Salminen, J. K., Karlsson, H., Hietala, J., Kajander, J., Aalto, S., Markkula, J., Rasi-Hakala, H. & Toikka, T. 2008. Short-term psychodynamic psychotherapy and fluoxetine in major depressive disorder: A randomized comparative study. *Psychotherapy and Psychosomatics,* 77**,** 351-357.

Salokangas, R. K. R., Vaahtera, K., Pacriev, S., Sohlman, B. & Lehtinen, V. 2002. Gender differences in depressive symptoms: An artefact caused by measurement instruments? *Journal of Affective Disorders,* 68**,** 215-220.

Sartorius, N., Ustun, T. B., Costa, J. A. S., Goldberg, D., Lecrubier, Y., Ormel, J., Von Korff, M. & Wittchen, H. U. 1993. An international study of psychological problems in primary care: Preliminary report from the World Health Organization Collaborative project on 'psychological problems in general health care'. *Archives of General Psychiatry,* 50**,** 819-824.

Saunders, S. M., Howard, K. I. & Newman, F. L. 1988. Evaluating the clinical significance of treatment effects: Norms and normality. *Behavioral Assessment,* 10**,** 207-218.

Schatzberg, A. F. & Kraemer, H. C. 2000. Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. *Biological Psychiatry,* 47**,** 736-744.

Schmitz, N., Wang, J., Malla, A. & Lesage, A. 2007. Joint effect of depression and chronic conditions on disability: Results from a population-based study. *Psychosomatic Medicine,* 69**,** 332-338.

Schulberg, H. C., Block, M. R., Madonia, M. J., Scott, C. P., Rodriguez, E., Imber, S. D., Perel, J., Lave, J., Houck, P. R. & Coulehan, J. L. 1996. Treating major depression in primary care practice - Eight-month clinical outcomes. *Archives of General Psychiatry,* 53**,** 913-919.

Scott, A. I. F. & Freeman, C. P. L. 1992. Edinburgh Primary Care Depression Study - Treatment Outcome, Patient Satisfaction, and Cost after 16 Weeks. *British Medical Journal,* 304**,** 883-887.

Seggar, L. B., Lambert, M. J. & Hansen, N. B. 2002. Assessing clinical significance: Application to the Beck depression inventory. *Behavior Therapy,* 33**,** 253-269.

Shapiro, D. A., Barkham, M., Rees, A., Hardy, G. E., Reynolds, S. & Startup, M. 1994. Effects of Treatment Duration and Severity of Depression on the Effectiveness of Cognitive-Behavioral and Psychodynamic Interpersonal Psychotherapy. *Journal of Consulting and Clinical Psychology,* 62**,** 522-534.

Shapiro, D. A., Barkham, M., Stiles, W. B., Hardy, G. E., Rees, A., Reynolds, S. & Startup, M. 2003. Time is of the essence: A selective review of the fall and rise of brief therapy research. *Psychology and Psychotherapy: Theory, Research and Practice,* 76**,** 211-235.

Sigmon, S. T., Pells, J. J., Boulard, N. E., Whitcomb-Smith, S., Edenfield, T. M., Hermann, B. A., Lamattina, S. M., Schartel, J. G. & Kubik, E. 2005. Gender differences in self-reports of depression: The response bias hypothesis revisited. *Sex Roles,* 53**,** 401-411.

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J. & Thompson, S. G. 2005. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials,* 2**,** 209-217.

Simon, G. E. & Vonkorff, M. 1992. Reevaluation of Secular Trends in Depression Rates. *American Journal of Epidemiology,* 135**,** 1411-1422.

Smith, M. L. & Glass, G. V. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 32**,** 752-760.

Solomon, D. A., Keller, M. B., Leon, A. C., Mueller, T. I., Lavori, P. W., Shea, M. T., Coryell, W., Warshaw, M., Turvey, C., Maser, J. D. & Endicott, J. 2000. Multiple recurrences of major depressive disorder. *American Journal of Psychiatry,* 157**,** 229-233.

Solomon, D. A., Keller, M. B., Leon, A. C., Mueller, T. I., Shea, M. T., Warshaw, M., Maser, J. D., Coryell, W. & Endicott, J. 1997. Recovery from major depression: A 10-year prospective follow-up across multiple episodes. *Archives of General Psychiatry,* 54**,** 1001-1006.

Speer, D. C. 1992. Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology,* 60**,** 402-408.

Spijker, J., De Graaf, R., Bijl, R. V., Beekman, A. T. F., Ormel, J. & Nolen, W. A. 2002. Duration of major depressive episodes in the general population: Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *British Journal of Psychiatry,* 181**,** 208-213.

Spitzer, R. L., Endicott, J. & Robins, E. 1978. Research diagnostic criteria. Rationale and reliability. *Archives of General Psychiatry,* 35**,** 773-782.

Spitzer, R. L. & Wakefield, J. C. 1999. DSM-IV Diagnostic Criterion for Clinical Significance: Does It Help Solve the False Positives Problem? *Am J Psychiatry,* 156**,** 1856-1864.

Staines, G. L. 2007. Comparative outcome evaluations of psychotherapies: guidelines for addressing eight limitations of the gold standard of causal inference. *Psychotherapy,* 44**,** 161-174.

Steer, R. A., Beck, A. T., Riskind, J. H. & Brown, G. 1987. Relationships between the Beck Depression Inventory and the Hamilton Psychiatric Rating Scale for Depression in depressed outpatients. *Journal of Psychopathology and Behavioral Assessment,* 9**,** 327-339.

Stewart, J. W., McGrath, P. J., Quitkin, F. M. & Klein, D. F. 2009. DSM-IV depression with atypical features: Is it valid. *Neuropsychopharmacology,* 34**,** 2625-2632.

Swendsen, J. D. & Merikangas, K. R. 2000. The comorbidity of depression and substance use disorders. *Clinical Psychology Review,* 20**,** 173-189.

Tamres, L. K., Janicki, D. & Helgeson, V. S. 2002. Sex Differences in Coping Behavior: A Meta-Analytic Review and an Examination of Relative Coping. *Personality and Social Psychology Review,* 6**,** 2-30.

Taylor, M. A. & Fink, M. 2008. Restoring melancholia in the classification of mood disorders. *Journal of Affective Disorders,* 105**,** 1-14.

Thase, M. E. 2009. Atypical depression: Useful concept, but it's time to revise the DSM-IV criteria. *Neuropsychopharmacology,* 34**,** 2633-2641.

Tingey, R. C., Lambert, M. J., Burlingame, G. M. & Hansen, N. B. 1996a. Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research,* 6**,** 109-123.

Tingey, R. C., Lambert, M. J., Burlingame, G. M. & Hansen, N. B. 1996b. Clinically significant change: Practical indicators for evaluating psychotherapy outcome. *Psychotherapy Research,* 6**,** 144-153.

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., Mors, O., Elkin, A., Williamson, R. J., Schmael, C., Henigsberg, N., Perez, J., Mendlewicz, J., Janzing, J. G. E., Zobel, A., Skibinska, M., Kozel, D., Stamp, A. S., Bajs, M., Placentino, A., Barreto, M., McGuffin, P. & Aitchison, K. J. 2008. Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychological Medicine,* 38**,** 289-300.

University of Sheffield. 2009. *Critical Appraisal and Using the Literature: Appraisal of Reviews,* Available: http://www.shef.ac.uk/scharr/ir/units/critapp/apprev.htm [Accessed 25th November 2009].

Van Walraven, C. 2010. Individual patient meta-analysis-rewards and challenges. *Journal of Clinical Epidemiology,* 63**,** 235-237.

Vittengl, J. R., Clark, L. A., Dunn, T. W., Jarrett, R. B. & Vittengl, J. R. 2007. Reducing relapse and recurrence in unipolar depression: a comparative meta-analysis of cognitive-behavioral therapy's effects. *Journal of Consulting & Clinical Psychology,* 75**,** 475-88.

Wampold, B. E. & Jenson, W. R. 1986. Clinical significance revisited. *Behavior Therapy,* 17**,** 302-305.

Wampold, B. E., Minami, T., Baskin, T. W., Callen Tierney, S., Wampold, B. E., Minami, T., Baskin, T. W. & Callen Tierney, S. 2002. A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. *Journal of Affective Disorders,* 68**,** 159-65.

Wang, P. S., Berglund, P., Olfson, M., Pincus, H. A., Wells, K. B. & Kessler, R. C. 2005. Failure and delay in initial treatment contact after first onset of mental disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry,* 62**,** 603-613.

Watkins, J. T., Leber, W. R., Imber, S. D., Collins, J. F., Elkin, I., Pilkonis, P. A., Sotsky, S. M., Shea, M. T. & Glass, D. R. 1993. Temporal Course of Change of Depression. *Journal of Consulting and Clinical Psychology,* 61**,** 858-864.

Wells, J. E. & Horwood, L. J. 2004. How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports. *Psychological Medicine,* 34**,** 1001-1011.

Westen, D., Novotny, C. A. & Thompson-Brenner, H. 2004. The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin,* 130**,** 631-663.

World Health Organisation 1992. *The ICD - 10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines.,* Geneva, World Health Organisation.

World Health Organisation 1993. *ICD-10, the ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research,* Geneva, World Health Organization.

World Health Organisation 2001. *The World Health Report 2001. Mental Health: New Understanding, New Hope,* Geneva, World Health Organisation.

World Health Organisation 2008. Integrating Mental Health into Primary Care: a Global Perspective. Geneva: World Health Organisation.

Williams, J. B. W. 1988. A Structured Interview Guide for the Hamilton Depression Rating-Scale. *Archives of General Psychiatry,* 45**,** 742-747.

Wilson, D. B. & Lipsey, M. W. 2001. The Role of Method in Treatment Effectiveness Research: Evidence From Meta-Analysis. *Psychological Methods,* 6**,** 413-429.

Wilson, G. T. 1998. The clinical utility of randomized controlled trials. *International Journal of Eating Disorders,* 24**,** 13-29.

Wolf, M. M. 1978. Social Validity: The Case for Subjective Measurement or How Applied Behavior Analysis is Finding its Heart. *Journal of Applied Behavior Analysis,* 11**,** 203-214.

Yoshimasu, K., Kiyohara, C. & Miyashita, K. 2008. Suicidal risk factors and completed suicide: Meta-analyses based on psychological autopsy studies. *Environmental Health and Preventive Medicine,* 13**,** 243-256.

Zimmerman, M., Chelminski, I. & Posternak, M. 2004a. A review of studies of the Hamilton Depression Rating Scale in healthy controls: Implications for the definition of remission in treatment studies of depression. *Journal of Nervous and Mental Disease,* 192**,** 595-601.

Zimmerman, M., Posternak, M. A. & Chelminski, I. 2004b. Implications of using different cut-offs on symptom severity scales to define remission from depression. *International Clinical Psychopharmacology,* 19**,** 215-220.

Appendices

Appendix A

Search Filters used to Search Electronic Databases in Study One

Cochrane Library DARE **Searched 11/03/2009**

Search	Filter	Results
1	explode Psychotherapy/	569
2	psychotherap* or "behavior* NEAR/6 therap*" or behaviour* NEAR/6 therap* or biofeedback and psycho* or cognitive NEAR/6 therap* or desensiti* and psychol* or "implosive therap*" or relax* NEAR/6 therap* or relax* NEAR/6 techniq* or therap* NEAR/6 touch* (all text)	622
3	bibliotherapy or imagery and psychotherap* or counsel* or "milieu therap*" or psychosoc* or psycholog* or support* NEAR/6 group* or guide* NEAR/6 image* or "gestalt therap*" or "nondirective therap*" or "play therap*" or psychoanaly* NEAR/6 therap* or "psychotherap* process*" (all text)	1549
4	OR 1-3	1716
5	explode Depression/	103
6	depressi* NEAR/3 disorder* or depressi* NEAR/3 symptom* depression or depressive* or depressed or dysthymia* (all text)	350
7	explode Depressive Disorder/	232
8	OR 5-7	410
9	4 AND 8	252
10	meta-analy* or metaanal* (all text) or meta-analysis.pt,ab,ti.	4494
11	explode Meta-Analysis as Topic explode all trees	300
12	systematic* NEAR/4 review* or systematic* NEAR/4 overview* (all text)	7732
13	extraction.ab.	0
14	letter or comment or editorial pt.	0
15	10 or 11 or 12 or 13	8402
16	15AND NOT 14	8402
17	9 and 16	**222**

Cochrane Database of Controlled Trials **Searched 11/03/2009**

Search	Filter	Results
1	explode Psychotherapy/	8429
2	psychotherap* or "behavior* NEAR/6 therap*" or behaviour* NEAR/6 therap* or biofeedback and psycho* or cognitive NEAR/6 therap* or desensiti* and psychol* or "implosive therap*" or relax* NEAR/6 therap* or relax* NEAR/6 techniq* or therap* NEAR/6 touch* (all text)	10890
3	bibliotherapy or imagery and psychotherap* or counsel* or "milieu therap*" or psychosoc* or psycholog* or support* NEAR/6 group* or guide* NEAR/6 image* or "gestalt therap*" or "nondirective therap*" or "play therap*" or psychoanaly* NEAR/6 therap* or "psychotherap* process*" (all text)	45982
4	OR 1-3	50722
5	explode Depression/	3042
6	depressi* NEAR/3 disorder* or depressi* NEAR/3 symptom* depression or depressive* or depressed or dysthymia* (all text)	11955
7	explode Depressive Disorder/	4761
8	OR 5-7	13714
9	4 AND 8	6660
10	meta-analy* or metaanal* (all text) or meta-analysis.pt,ab,ti.	1420
11	explode Meta-Analysis as Topic explode all trees	172
12	systematic* NEAR/4 review* or systematic* NEAR/4 overview* (all text)	292
13	extraction.ab.	2589
14	letter or comment or editorial .pt.	0
15	10 or 11 or 12 or 13	4172
16	15AND NOT 14	4172
17	9 and 16	**31**

Cochrane Database of Economic Evaluations **Searched 11/03/2009**

Search	Filter	Results
1	explode Psychotherapy/	281
2	psychotherap* or "behavior* NEAR/6 therap*" or behaviour* NEAR/6 therap* or biofeedback and psycho* or cognitive NEAR/6 therap* or desensiti* and psychol* or "implosive therap*" or relax* NEAR/6 therap* or relax* NEAR/6 techniq* or therap* NEAR/6 touch* (all text)	250
3	bibliotherapy or imagery and psychotherap* or counsel* or "milieu therap*" or psychosoc* or psycholog* or support* NEAR/6 group* or guide* NEAR/6 image* or "gestalt therap*" or "nondirective therap*" or "play therap*" or psychoanaly* NEAR/6 therap* or "psychotherap* process*" (all text)	1758
4	OR 1-3	1924
5	explode Depression/	161
6	depressi* NEAR/3 disorder* or depressi* NEAR/3 symptom* depression or depressive* or depressed or dysthymia* (all text)	384
7	explode Depressive Disorder/	306
8	OR 5-7	5372
9	4 AND 8	196
10	meta-analy* or metaanal* (all text) or meta-analysis.pt,ab,ti.	757
11	explode Meta-Analysis as Topic explode all trees	33
12	systematic* NEAR/4 review* or systematic* NEAR/4 overview* (all text)	1943
13	extraction.ab.	0
14	letter or comment or editorial .pt.	0
15	10 or 11 or 12 or 13	2343
16	15AND NOT 14	2343
17	9 and 16	**25**

Cochrane Database of Systematic Reviews **Searched 10/03/2009**

Search	Filter	Results
1	explode Psychotherapy/	147
2	psychotherap* or "behavior* NEAR/6 therap*" or behaviour* NEAR/6 therap* or biofeedback and psycho* or cognitive NEAR/6 therap* or desensiti* and psychol* or "implosive therap*" or relax* NEAR/6 therap* or relax* NEAR/6 techniq* or therap* NEAR/6 touch* (all text)	611
3	bibliotherapy or imagery and psychotherap* or counsel* or "milieu therap*" or psychosoc* or psycholog* or support* NEAR/6 group* or guide* NEAR/6 image* or "gestalt therap*" or "nondirective therap*" or "play therap*" or psychoanaly* NEAR/6 therap* or "psychotherap* process*" (all text)	2071
4	OR 1-3	2162
5	explode Depression/	50
6	depressi* NEAR/3 disorder* or depressi* NEAR/3 symptom* depression or depressive* or depressed or dysthymia* (all text)	518
7	explode Depressive Disorder/	25
8	OR 5-7	526
9	4 AND 8	390
10	meta-analy* or metaanal* (all text) or meta-analysis.pt,ab,ti.	4191
11	explode Meta-Analysis as Topic explode all trees	14
12	systematic* NEAR/4 review* or systematic* NEAR/4 overview* (all text)	5676
13	extraction.ab.	528
14	letter or comment or editorial .pt.	0
15	10 or 11 or 12 or 13	5676
16	15AND NOT 14	5676
17	9 and 16	390
18	Restrict to reviews (not protocols)	**290**

Cochrane Database of Technology Assessments **Searched 11/03/2009**

Search	Filter	Results
1	explode Psychotherapy/	97
2	psychotherap* or "behavior* NEAR/6 therap*" or behaviour* NEAR/6 therap* or biofeedback and psycho* or cognitive NEAR/6 therap* or desensiti* and psychol* or "implosive therap*" or relax* NEAR/6 therap* or relax* NEAR/6 techniq* or therap* NEAR/6 touch* (all text)	119
3	bibliotherapy or imagery and psychotherap* or counsel* or "milieu therap*" or psychosoc* or psycholog* or support* NEAR/6 group* or guide* NEAR/6 image* or "gestalt therap*" or "nondirective therap*" or "play therap*" or psychoanaly* NEAR/6 therap* or "psychotherap* process*" (all text)	335
4	OR 1-3	413
5	explode Depression/	42
6	depressi* NEAR/3 disorder* or depressi* NEAR/3 symptom* depression or depressive* or depressed or dysthymia* (all text)	80
7	explode Depressive Disorder/	48
8	OR 5-7	102
9	4 AND 8	41
10	meta-analy* or metaanal* (all text) or meta-analysis.pt,ab,ti.	242
11	explode Meta-Analysis as Topic explode all trees	1
12	systematic* NEAR/4 review* or systematic* NEAR/4 overview* (all text)	2061
13	extraction.ab.	0
14	letter or comment or editorial .pt.	0
15	10 or 11 or 12 or 13	2101
16	15AND NOT 14	2101
17	9 and 16	**19**

PsycINFO **Search 1967 to March Week 2 2009**

Search	Filter	Results
1	exp Psychotherapy/	131801
2	(psychotherap$ or "behavior$ adj6 therap$" or (behaviour$ adj6 therap$) or (biofeedback and psycho$) or (cognitive adj6 therap$) or (desensiti$ and psychol$) or "implosive therap$" or (relax$ adj6 therap$) or (relax$ adj6 techniq$) or (therap$ adj6 touch$)).tw.	86063
3	(bibliotherapy or (imagery and psychotherap$) or counsel$ or "milieu therap$" or psychosoc$ or psycholog$ or (support$ adj6 group$) or (guide$ adj6 image$) or "gestalt therap$" or "nondirective therap$" or "play therap$" or (psychoanaly$ adj6 therap$) or "psychotherap$ process$").tw.	403913
4	exp Postpartum Depression/ or exp Recurrent Depression/ or exp Atypical Depression/ or exp Endogenous Depression/ or exp "Depression (Emotion)"/ or exp Reactive Depression/ or exp Treatment Resistant Depression/ or exp Major Depression/	81130
5	(depressi$ adj3 disorder$).tw.	21335
6	(depressi$ adj3 symptom$).tw.	23760
7	(depression or depressive$ or depressed or dysthymia$).tw.	145304
8	or/1-3	512793
9	or/4-7	147147
10	8 and 9	40511
11	(meta-analy$ or metaanal$).tw.	9933
12	meta-analysis.pt,ab,ti.	7042
13	exp Meta Analysis/	2865
14	(systematic$ adj4 (review$ or overview$)).tw.	3996
15	extraction.ab.	2129
16	(letter or comment or editorial).pt.	0
17	or/11-15	15238
18	17 not 16	15238
19	10 and 18	525
20	limit 19 to English language	**479**

EMBASE **Search 1980 to 2009 Week 10**

Search	Filter	Results
1	exp Psychotherapy/	78759
2	(psychotherap$ or "behavior$ adj6 therap$" or (behaviour$ adj6 therap$) or (biofeedback and psycho$) or (cognitive adj6 therap$) or (desensiti$ and psychol$) or "implosive therap$" or (relax$ adj6 therap$) or (relax$ adj6 techniq$) or (therap$ adj6 touch$)).tw.	34009
3	(bibliotherapy or (imagery and psychotherap$) or counsel$ or "milieu therap$" or psychosoc$ or psycholog$ or (support$ adj6 group$) or (guide$ adj6 image$) or "gestalt therap$" or "nondirective therap$" or "play therap$" or (psychoanaly$ adj6 therap$) or "psychotherap$ process$").tw.	168193
4	(depressi$ adj3 disorder$).tw.	18770
5	(depressi$ adj3 symptom$).tw.	19729
6	(depression or depressive$ or depressed or dysthymia$).tw.	181469
7	Reactive Depression/ or Bipolar Depression/ or Depression/ or Recurrent Brief Depression/ or Masked Depression/ or Long Term Depression/ or Atypical Depression/ or Agitated Depression/ or Puerperal Depression/ or Postoperative Depression/ or Major Depression/ or Endogenous Depression/	142332
8	or/1-3	235329
9	or/4-7	228187
10	8 and 9	36871
11	(meta-analy$ or metaanal$).tw.	22902
12	meta-analysis.pt,ab,ti.	17783
13	exp Meta Analysis/	34829
14	(systematic$ adj4 (review$ or overview$)).tw.	18115
15	extraction.ab.	82426
16	(letter or comment or editorial).pt.	666180
17	or/11-15	137560
18	17 not 16	133890
19	10 and 18	1217
20	limit 19 to English language	**1060**

MEDLINE In-Process & Other Non-Indexed Citations **Searched 6/03/2009**

Search	Filter	Results
1	psychotherapy.af.	947
2	(psychotherap$ or "behavior$ adj6 therap$" or (behaviour$ adj6 therap$) or (biofeedback and psycho$) or (cognitive adj6 therap$) or (desensiti$ and psychol$) or "implosive therap$" or (relax$ adj6 therap$) or (relax$ adj6 techniq$) or (therap$ adj6 touch$)).tw.	1375
3	(bibliotherapy or (imagery and psychotherap$) or counsel$ or "milieu therap$" or psychosoc$ or psycholog$ or (support$ adj6 group$) or (guide$ adj6 image$) or "gestalt therap$" or "nondirective therap$" or "play therap$" or (psychoanaly$ adj6 therap$) or "psychotherap$ process$").tw.	9235
4	depression.af.	6561
5	(depressi$ adj3 disorder$).af.	1076
6	(depressi$ adj3 symptom$).af.	1401
7	Depressive Disorder.af.	559
8	(depression or depressive$ or depressed or dysthymia$).af.	8207
9	1 or 3 or 2	10587
10	8 or 6 or 4 or 7 or 5	8209
11	10 and 9	1583
12	(meta-analy$ or metaanal$).tw.	2131
13	meta-analysis.ab,ti,pt.	1816
14	Meta-Analysis.af.	1816
15	(systematic$ adj4 (review$ or overview$)).tw.	2189
16	extraction.ab.	7672
17	(letter or comment or editorial).pt.	42128
18	16 or 13 or 12 or 15 or 14	11268
19	18 not 17	10981
20	11 and 19	58
21	limit 20 to English language	**53**

MEDLINE **Search 1950 to February Week 4 2009**

Search	Filter	Results
1	exp Psychotherapy/	119668
2	(psychotherap$ or "behavior$ adj6 therap$" or (behaviour$ adj6 therap$) or (biofeedback and psycho$) or (cognitive adj6 therap$) or (desensiti$ and psychol$) or "implosive therap$" or (relax$ adj6 therap$) or (relax$ adj6 techniq$) or (therap$ adj6 touch$)).tw.	35200
3	(bibliotherapy or (imagery and psychotherap$) or counsel$ or "milieu therap$" or psychosoc$ or psycholog$ or (support$ adj6 group$) or (guide$ adj6 image$) or "gestalt therap$" or "nondirective therap$" or "play therap$" or (psychoanaly$ adj6 therap$) or "psychotherap$ process$").tw.	201805
4	exp Depression/	50959
5	(depressi$ adj3 disorder$).tw.	17707
6	(depressi$ adj3 symptom$).tw.	19872
7	exp Depressive Disorder/	59678
8	(depression or depressive$ or depressed or dysthymia$).tw.	203764
9	or/1-3	311659
10	or/4-8	234895
11	9 and 10	33833
12	(meta-analy$ or metaanal$).tw.	23815
13	meta-analysis.pt,ab,ti.	27840
14	exp Meta-Analysis/	20228
15	(systematic$ adj4 (review$ or overview$)).tw.	19228
16	extraction.ab.	91998
17	(letter or comment or editorial).pt.	931461
18	or/12-16	134304
19	18 not 17	132489
20	11 and 19	644
21	limit 20 to English language	**602**

SCOPUS **Searched 12/03/2009**

Search	Filter	Results
1	KEY(Depressi* or depressive* or depressed or dysthymia*)	**262504**
2	KEY(Psychotherapy)	**76148**
3	KEY("bibliotherapy or "self-help" or "self help" or (comput* and therap*) or (online and therap*))	**12978**
4	2 or 3	**88255**
5	KEY((meta analysis) OR (metaanalysis) OR (systematic review) OR (systematic overview))	**63807**
6	1 and 4 and 5	**562**
7	Limit to English language	**505**

Web of Knowledge **Searched 11/03/2009**

Science Citation Index Expanded (SCI-EXPANDED) 1945to search date

Social Sciences Citation Index (SSCI) 1956 to search date

Search	Filter	Results
1	Topic = ((behavior* or behaviour* or cognitive or meta?cognit* or implosive or psycho* or interpersonal or gestalt or person?cent* or activation* or bibliotherapy* or counsel* or supportive or non?directive or guided or image* or computer* or cbt) and therap*) Document Type=(ARTICLE OR REVIEW OR CORRECTION)	88436
2	Topic = ((behavior* or behaviour* or cognitive or meta?cognit* or implosive or psycho* or interpersonal or gestalt or person?cent* or activation* or bibliotherapy* or counsel* or supportive or non?directive or guided or image* or computer* or cbt) and psychotherap*) Document Type=(ARTICLE OR REVIEW OR CORRECTION/ ADDITION)	26907
3	1 OR 2 Restricted to English Language	91978
4	Topic = Depress* Document Type=(ARTICLE OR CORRECTION OR REVIEW) AND Languages=(ENGLISH)	75773
5	Topic=(dysthymi*) Document Type=(ARTICLE OR CORRECTION/ ADDITION OR REVIEW) AND Languages = (ENGLISH)	1522
6	4 OR 5	76816
7	Topic=(met*analy*)	28241
8	Title=("systematic review" or "systematic overview")	15610
9	7 OR 8	42775
10	Topic=(extraction) Document Type=(ARTICLE OR REVIEW OR CORRECTION) AND Languages=(ENGLISH)	82658
11	9 OR 10	92698
12	3 AND 6	7894
13	12 AND 11	**505**

Appendix B

Screening Tool used to Assess the Eligibility of Reviews in Study One

Article Reference	Criterion Present?	Y/N	Include?	Notes
	Reviews RCTs - psychotherapy for Depression?			
	Adults?			
	Research Diagnosis of Depression?			
	Individual Psychotherapy for Depression?			
	Adequate Controls?			
	Synthesis of Psychotherapy efficacy?			
	Psychosis/PD/Medical/Substance Abuse?			

Appendix C

Composite Tool used to Extract Substantive & Quality Data* from Individual Reviews in Study One

BIBLIOGRAPHIC DETAILS

Descriptives* *(Adapted from CRDs DARE format)*	Quality*/ risk of bias *(Adapted from ScHARR format)*	
Summary of Review (Our abstract):		
Authors Objectives:	Did the review address a clearly focused question?	Y/N?
Search Methods:	Was the search strategy adequate (i.e. did the reviewers identify all relevant studies?)	Y/N?

STUDY SELECTION

What are the included designs in the review?	
Were the inclusion/exclusion criteria specified?	Y/N?
What types of therapy were included?	
What attempts were made to identify the 'purity' of therapy?	
Participants?	
Diagnostic techniques?	
Severity?	
Duration?	
Did the review include the right kinds of studies?	Y/N?

STUDY OUTCOMES

Assessment points in time/follow up?
Comparison groups?
Measures of severity?
Outcomes used in individual studies?

Continued from previous page....

VALIDITY ASSESSMENT

How were individual studies determined suitable for inclusion?

Did the reviewers assess the quality of the included studies?	Y/N?
Were appropriate outcome measures used?	Y/N?

SYNTHESIS

Synthesis: Assessment of differences between included studies?		Was the method of data extraction reported?	Y/N?
		Heterogeneity found?	Y/N?
		Heterogeneity accounted for	Y/N?
		Are appropriate sub-group analyses presented?	Y/N?
If the results of the studies have been combined, was it appropriate to do so?	Y/N?		

RESULTS OF THE REVIEW

Authors' Conclusions:		Are the main results of the review presented (e.g. numerical results included with CIs)	Y/N?
		Justified on included evidence?	Y/N?
Are issues of generalisability addressed?	Y/N?	Justified on included evidence?	Y/N?
Authors statements concerning implications for practice/research?		Justified on included evidence?	Y/N?

Our Comments on the review as a whole based on the qualitative/quantitative findings

* See main text for information about origins of table design.

Appendix D

Synthesis Methods of Reviews in Study One

Review	Method of Synthesis[†]
Casacalenda et al. (2002)	Post-treatment remission percentages for each study's treatment condition were averaged across studies to give an average remission rate. The average remission rates for each treatment condition were then assessed for significant differences using analysis of variance. No weighting of individual study results by sample size, nor testing for between study heterogeneity were described. The authors used SAS version 8.0 data analysis software.
de Maat et al. (2006)	The relative efficacy of treatments within studies were calculated using odds and relative risk ratios for remission at post treatment and relapse during follow up. These effect sizes were then weighted according to study size and combined to produce an overall estimate of the odds or risk ratio for remission or relapse. The authors employed a fixed effects model, tested for heterogeneity between individual study results and used Review Manager 4.2 software of the Cochrane Collaboration.
de Maat et al. (2007)	As for de Maat et al. (2006) for post treatment outcomes only.
Friedman et al. (2004)	The relative efficacy of treatments within studies was calculated using Cohen's d for both symptom reduction and recovery status. Effect sizes were weighted according to study size and combined to produce a Cohen's d for both symptom reduction and recovery. Between study heterogeneity was tested for using Chi-squared analyses. Results were presented for both post treatment and follow up outcomes. The authors did not report the statistical significance of their results.
Leichsenring (2001)	Success rate differences between treatments were assessed in individual studies by testing for significant differences in correlation coefficients (Cramer's Phi). Correlation coefficients from studies were transformed and compared to test for significant heterogeneity. A weighted mean Phi value was subsequently derived for both post treatment and follow up outcomes.
Parker et al. (2008)	The relative efficacy of treatments within studies were calculated using Cohen's d for symptom reduction on the BDI and odds and risk ratios for response. These effect sizes were then weighted according to study size and combined to produce an overall estimate of symptom reduction or response. The authors employed a random effects model, tested for heterogeneity between studies, and used the META statistical package version 8.01.
Vittengl et al. (2007)	The synthesis was based on the studywise logit transformed proportions of patients experiencing relapse or recurrence of depression by treatment type. These proportions were then weighted according to study size to provide an overall estimate of the numbers relapsing by treatment type. The authors employed a random effects model and tested for heterogeneity between individual study results.

Appendix E

Details of five Borderline Reviews Included in Study One

Review	Further Information
Casacalenda et al. (2002)	High levels of personality disorder (PD) comorbidity was reported for 3 of the 6 constituent studies which indicated the need for referral to a third reviewer (Elkin et al., 1989; Schulberg et al., 1996; Scott and Freeman, 1992). The review was included as these studies did not treat depression specifically in the context of PD. The finding of high PD comorbidity was possibly due to a level of detail in reporting not seen in other reviews. For example, other eligible reviews included Elkin et al. (1989) but did not provide information concerning the proportion of patients with PD.
Friedman et al. (2004)	Presented pooled results for group and individual psychotherapy. Consequently, only the results of comparisons that met our eligibility criteria were included.
Leichsenring (2001)	High levels of reported personality comorbidity in some included studies were accepted for same reasons as for Casacalenda et al. (2002). High levels of co-morbid Generalised Anxiety Disorder and/or Panic Disorder were identified in one study which used DSM III criteria (Shapiro et al., 1994). The review was not excluded on the basis that the primary diagnosis of patients in this study was depression.
Parker et al. (2008)	One of nine studies included patients with dysthymia (Hautzinger et al., 1996) whose BDI and HRSD scores were required to be greater than 20, indicating moderate depression (Gotlib and Hammen, 2002). These patients were considered as depressed according to their symptom severities.
Vittengl et al. (2007)	Results of comparisons that were based on studies not meeting our eligibility criteria were excluded.