SCREENING FOR DEPRESSION IN OLDER ADULTS

Volume 1

Dr Claire Pocklington MBChB MRCPsych

MSc by Thesis

The University of Hull and The University of York

Hull York Medical School

June 2016

ABSTRACT

Despite being the most common mental disorder in older adults, depression is under-recognised. It poses diagnostic difficulties in this population for several reasons; for example, symptomatic and phenomenological differences, age-related biological and psychological factors, and the presence of physical comorbidities. Depression in older adults is an important clinical topic because outcomes are worse in comparison to younger adults. It is also associated with higher rates of morbidity and mortality, increased healthcare utilisation and economic costs. It is likely to become a more pressing issue in the future due to the projected increase in the older adult population.

Screening for depression could be a solution to improve detection rates and avert the negative consequences of depression. This dissertation explores the topic of screening for depression in older adults. It uses systematic review methods to examine two questions. First, what is the diagnostic accuracy of the Geriatric Depression Scale? Secondly, what is the clinical effectiveness of screening for depression in older adults?

Findings of this dissertation show that the diagnostic performance of the Geriatric Depression Scale, at the recommended cut-off score of 5, is acceptable for screening purposes. However, results suggest the possibility of selective reporting of cut-off scores post-hoc and therefore findings should be approached cautiously. The dissertation found limited evidence regarding the clinical effectiveness of screening for depression in older adults and therefore cannot make any recommendations for policy or practice.

LIST OF CONTENTS

LIST OF TABLES, FIGURES AND APPENDICES

Tables

Figures

Pg.

<u>Appendices</u> Pg.

LIST OF PUBLICATIONS

Pocklington C, Gilbody S, Manea L, McMillan D, et al. The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis. Int J Geriatr Psychiatry 2016 Feb 18. doi: 10.1002/gps.4407.

D McMillan and S Gilbody were both supervisors for this dissertation. Any uncertainty regarding the screening of citations and data extraction, which were performed independently by the author Claire Pocklington, was discussed with them. Laura Manea gave advice on analysis but this was conducted by the author Claire Pocklington.

See Appendix 5.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr Dean McMillan, whose expertise, advice and patience has allowed me the opportunity to complete this dissertation. His time and effort is much appreciated.

I would also like to express my gratitude and appreciation to Prof Simon Gilbody and Dr Laura Manea.

Finally, I would like to thank Jamie for his support, understanding and encouragement.

AUTHOR DECLARATION

I confirm that this work is original and that if any passage(s) or diagram(s) have been copied from academic papers, books, the internet or any other sources these are clearly identified by the use of quotation marks and the reference(s) is fully cited. I certify that, other than where indicated, this is my own work and does not breach the regulations of HYMS, the University of Hull or the University of York regarding plagiarism or academic conduct in examinations. I have read the HYMS Code of Practice on Academic Misconduct, and state that this piece of work is my own and does not contain any unacknowledged work from any other sources.

CHAPTER 1

Introduction, background and existing evidence

INTRODUCTION

The diagnosis of depression in older adults is problematic. Despite being the most common mental health condition in those aged over 65 years of age it is often under-detected and under-recognised (Weeks et al., 2003, Smalbrugge et al., 2008, Licht-Strunk et al., 2009, Conradsson et al., 2013). The failure to detect and recognise a health condition equates to no treatment for it, which in turn leads to worse outcomes for the patient (Weeks et al., 2003).

Depression in older adults poses diagnostic difficulties for several reasons including, for example, differences in symptomatology, the presence of comorbid physical conditions, misunderstanding and misattribution of the ageing process and reluctance to report and seek help (Chapman and Perry, 2008).

Depression is not just an important issue due to it being the most common mental health condition in older adults or the associated diagnostic difficulties. Depression in older adults has worse outcomes, in terms of morbidity and mortality, when compared to younger adults (Jongenelis et al., 2002, Friedman et al., 2005a, Nyunt et al., 2009b, Hegeman et al., 2012, Conradsson et al., 2013). The presentation of depression in older adults differs from that in younger adults. Depression is the commonest cause of suicide in older adults (Friedman et al., 2005a). It is associated with increased healthcare utilisation and economic costs (Rinaldi et al., 2003, Friedman et al., 2005a, Nyunt et al., 2009b), largely through indirect ways secondary to comorbid physical illnesses. Depression in older adults is treatable and good outcomes can be achieved (Pomeroy et al., 2001).

Diagnostic rates of depression in older adults could be improved through the use of a depression screening tools. Such tools could be utilised in a depression screening programme for older adults.

At present, screening for depression in older adults is not recommended in the UK. In fact, screening for depression in older adults is not recommended in any country. Case-finding for depression is recommended routine practice in the UK for individuals with long-term chronic physical health conditions. Other countries also employ case-finding for depression in those deemed high risk. There is currently an evidence gap regarding screening for depression in the older adult population in the UK.

The primary aim of this chapter is to explore the theory and process of screening in the context of depression in older adults. The first part of this chapter will provide an overview of the clinical topic, which will include diagnostic difficulties, clinical presentation and consequences, highlighting why it is such an important and significant clinical topic. This will underline differences in the presentation of depression between younger and older adults. The second part of this chapter will discuss the rationale for the introduction of screening for depression in older adults. This will be achieved by describing the process of screening, the arguments for and against screening in general and for mental health problems in particular, and the current evidence base in relation to screening for depression in this population.

DEPRESSION IN OLDER ADULTS

Description depression

Depression is a clinical syndrome. Both the Diagnostic and Statistical Manual (DSM) of Mental Disorder and International Classification of Diseases (ICD) diagnostic classification systems describe three core symptoms of depression; low mood, anhedonia and reduced energy levels (American PsychiatricAssociation, 2013, World HealthOrganization, 2001b). Other symptoms include impaired concentration, loss of confidence, suicidal ideation, disturbances in sleep and changes in appetite. Symptoms must have been present for at least a period of two weeks for a diagnosis of depression to be made. Major depression refers to the presence of all three core symptoms and, in accordance with ICD criteria, at least the presence of a further five other symptoms (Organization, 2001b). See Table 1 for severity criteria of a depressive episode according to ICD.

<table>
<tr><th>Criteria A – General:</th><th>Criteria B – Presence of ≥2 of the following:</th><th>Criteria C – 'Other' symptoms:</th></tr>
<tr><td>• Symptoms for at least 2 weeks
• Symptoms not attributable to psychoactive substance use or organic mental disorder</td><td>• Low mood
• Anhedonia
• Reduced energy levels/ increased fatigability</td><td>• Loss of confidence and self-esteem
• Feelings of guilt
• Suicidal thoughts
• Impaired concentration/ability to think
• Changes in psychomotor activity
• Sleep disturbance
• Changes in appetite with weight changes</td></tr>
<tr><td colspan="3"><b><u>Criteria for severity of depressive episode:</u></b></td></tr>
<tr><td><b>Mild episode:</b>
2 symptoms of criteria B</td><td><b>Moderate episode:</b>
≥2 symptoms of criteria B
+ symptoms of criteria C until minimum of 6 symptoms in total</td><td><b>Major episode:</b>
all 3 symptoms of criteria B
+ symptoms of criteria C until a minimum of 8 symptoms in total</td></tr>
</table>

Table 1: Severity criteria of a depressive episode according to ICD-10 (World Health Organisation, 2001)

Depressive symptoms, which can be clinically significant, can be present in the absence of a major depressive episode. Depressive symptoms are those that do not fulfil diagnostic criteria for a diagnosis of depression to be made. Depressive symptoms are a risk factor for the development of a major depressive disorder. Depressive symptoms can be collectively referred to as sub-threshold depression, sub-syndromal depression or minor depression (Cherubini et al., 2012).

It has been proposed that there are two types of depression; early-onset and late-onset depression. Late-onset depression refers to a new diagnosis in individuals aged 65 years of age or older. Over half of all cases of depression in older adults are newly arising (i.e. the individual has never experienced depression before) and thus late-onset type depression. The onset of depression in adolescence and adulthood is referred to as early-onset depression (Fiske et al., 2009).

It is proposed that late- and early-onset depression are different entities as aetiology, course and prognosis vary (Rapp et al., 2005, Fiske et al., 2009). A family history of depression and a past history of personality difficulties are associated with early-onset type depression. On the other hand, late-onset type depression is associated more with structural brain changes, vascular risk factors, cognitive deficits and the later development of dementia. It has been suggested that late-onset depression could be prodromal to dementia (Fiske et al., 2009).

It should be borne in mind that a distinction between early- and late-onset forms of depression does not rule out an older adult with a history of early-onset depression developing depression of a late-onset form. Distinguishing the form of depression (i.e. late- or early-onset) would be difficult clinically. However, a study by Rapp et al. has established differences in patients with late- and early-onset depression; older adults with late-onset depression were found to have cognitive deficits in attention and executive function, which were associated with increased anhedonia and cardiovascular comorbidity (Rapp et al., 2005) . Whereas older adults with recurrent depression (i.e. early-onset type) were found to have cognitive deficits concerning episodic memory (Rapp et al., 2005).

Diagnosis of depression, regardless of age, is reliant on clinical assessment; there is no diagnostic laboratory test or investigation. A diagnosis is made through history-taking and mental state examination.

Presentation of depression in older adults

The presentation of depression in older adults is markedly different to that in younger adults. The most significant and fundamental difference in presentation in older adults is that depression can be present with the absence of an affect component, i.e. subjective feelings of low mood or sadness are not experienced (Evans, 1995, Evans and Mottram, 2000, Alexopoulos, 2005, Arean and Ayalon, 2005, Fiske et al., 2009). The absence of an affective component is referred to as 'depression without sadness' (Alexopoulos, 2005, Arean and Ayalon, 2005). It is common instead for older adults to report a lack of feeling or emotion when depressed (Alexopoulos, 2005, Arean and Ayalon, 2005).

Anhedonia is also less prevalent in this population. However, reduced energy levels and fatigue are frequently reported (Alexopoulos, 2005, Arean and Ayalon, 2005).

Compared to younger adults, psychological symptoms of depression occur more frequently and are more prevalent in older adults (Mitchell et al., 2011). Such psychological symptoms include feelings of guilt, poor motivation, low interest levels, anxiety related symptoms and suicidal ideation. The presence of irritability and agitation are key features as well (Evans and Mottram, 2000). Hallucinations and delusions are also more common in older adults, particularly nihilistic delusions (i.e. a person believing their body is dead or a part of their body is not working properly or rotting).

Cognitive deficits are characteristic of depression in older adults (Evans and Mottram, 2000, Butters et al., 2004a) and are described as 'substantial and disabling' (Butters et al., 2004b). Such cognitive deficits mainly concern executive function (Alexopoulos et al., 2000, Lockwood et al., 2000). Pseudodementia is a phenomenon seen in older adults (Chapman and Perry, 2008). The term refers to cognitive impairment secondary to a psychiatric condition, most commonly depression (Kang et al., 2014). Pseudodementia has become synonymous with depression. Pseudodementia can be mistaken for an organic dementia and so older adults who are depressed can present primarily to mental health services with memory problems. Pseudodementia is classically associated with 'don't know' answers, whereas older adults with a true dementia will often respond with incorrect answers (Bieliauskas, 2013). The cognitive deficits, namely those surrounding executive function and attention, seen in depression in older adults contribute significantly to functional impairment (Rapp et al., 2005, Kang et al., 2014).

'Depression-executive dysfunction syndrome' is a more specific and descriptive term to describe the cognitive deficits found in older adults with depression (Lockwood et al., 2000). 'Depression-executive dysfunction syndrome' is associated with psychomotor retardation, which can be a core feature of depression in this population (Evans and Mottram, 2000, Lockwood et al., 2000, Beheydt et al., 2014). Psychomotor retardation describes a slowing of movement and mental activity (Bennabi et al., 2013). Like pure cognitive deficits, psychomotor retardation contributes significantly to functional impairment (Bennabi et al., 2013). Both executive dysfunction and psychomotor retardation have been found to be related to underlying structural changes in the frontal lobes (Lockwood et al., 2000, Rapp et al., 2005, Walther et al., 2012). Psychomotor retardation is further related to white matter changes in the motor system, which leads to impaired motor planning (Walther et al., 2012). Is it thought that psychomotor retardation comes hand-in-hand with cognitive deficits in the older adult population (Bennabi et al., 2013, Beheydt et al., 2014). There is conflicting evidence of whether the presence of psychomotor retardation is related to depression severity (Bennabi et al., 2013, Beheydt et al., 2014).

Somatisation and hypochondriasis are associated with depression in older adults and increasing age in general (El-Gabalawy et al., 2013). Somatisation and hypochondriasis have a higher incidence in older adults who have depression compared to younger adults (Shahpesandy, 2005). Somatisation is often overlooked in older adults by healthcare professional who actively search to attribute such symptoms to a physical cause. Somatisation is more common in those who have physical comorbidities. Somatisation is seen as associated with depression in older adults (Sheehan and Banerjee, 1999). Somatisation in older adults is associated with structural brain changes and cognitive deficits (Inamura et al., 2015).

As touched upon, depression in older adults is associated with functional impairment cognitively, physically and socially (Evans and Mottram, 2000, Butters et al., 2004b, Polenick, 2013). Such functional impairment is linked to loss of independent function and increased rates of disability (Rinaldi et al., 2003). Withdrawal from normal social and leisure activities can be marked (Evans and Mottram, 2000, Polenick, 2013). Social avoidance reduces interaction with others and is often a maintaining factor for depression (Polenick, 2013). Treatment of depression is often associated with an increased functional level but no improvement in other depressive symptomatology.

Self-neglect is a classical feature of depression (Evans and Mottram, 2000), with the presence of depressive symptoms in older adults being predictive of it (Abrams et al., 2002). In the context of depression, self-neglect refers to an inability or refusal, of an individual, to attend to their own health, hygiene, nutrition or social needs (Abrams et al., 2002). Self-neglect occurs secondary to other symptoms of depression, i.e. loss of energy, loss of motivation, cognitive deficits, etc. Self-neglect has also been associated with executive dysfunction (Hildebrand et al., 2014). Behavioural disturbances can be a common mode of presentation, especially for older adults living in institutionalised care (Evans, 1995, Evans and Mottram, 2000). Behavioural disturbances also include incontinence, food refusal, screaming, falling and violence towards others (Evans and Mottram, 2000).

On the whole the presentation of depression in older adults can be viewed as vague and somewhat non-specific, because reports of fatigue, poor sleep and reduced appetite can be attributed to a host of causes other than depression and therefore it is no surprise that a diagnosis of depression is overlooked and goes undetected by healthcare professionals (Birrer and Vemuri, 2004). Older adults themselves often attribute low energy levels, insomnia, poor appetite and weight loss to physical illness (Evans and Mottram, 2000, Alexopoulos, 2005). Physical illnesses can mask and mimic depression (Alexopoulos, 2005). Healthcare professionals, therefore, should have a high level of clinical suspicion when older adults present with physical complaints that could be ascribed to biological symptoms and signs of depression. There may be a role and value in screening or case-finding for depression reducing a need for a high level of clinical suspicion.

Diagnostic difficulties in older adults

Depression in older adults has been a condition that has constantly been under-recognised. Several issues account for this. Firstly, phenomenological differences are present. Many have argued that phenomenological issues contribute heavily to diagnostic difficulties (Prakash et al., 2009); both the DSM and ICD classification systems do not have specific diagnostic criteria for depression in older adults. Potentially invalid diagnostic criteria for depression in older adults could result in fundamental difficulties in understanding, with consequent impact on both clinical practice and research. See Figure 1. Hegeman et al. commented that age-related biological and psychological factors may contribute to differences seen in the phenomenology of early-onset and late-onset depression (Hegeman et al., 2012).

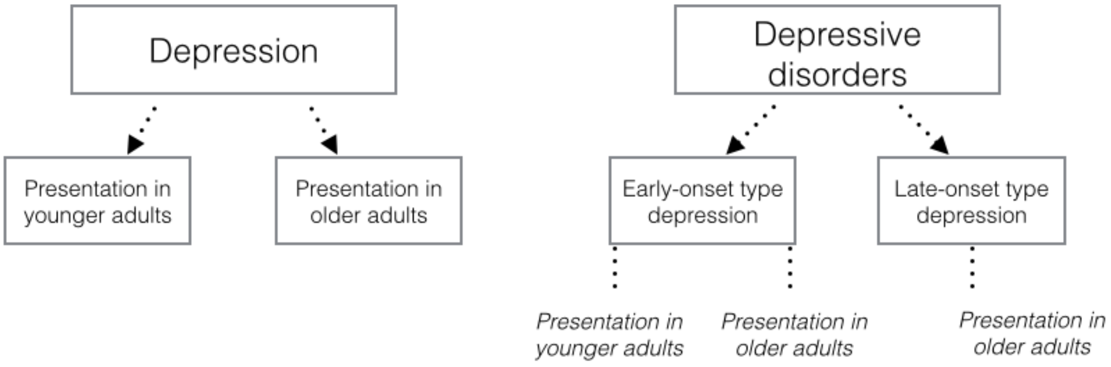
Depression
Presentation in younger adults
Presentation in older adults
Depressive disorders
Early-onset type depression
Late-onset type depression
Presentation in younger adults
Presentation in older adults
Presentation in older adults

Figure 1: The presentation of depression in younger and older adults

Diagnostic difficulties are also encountered because depression in older adults can present with vague symptoms, which do not correspond to the classical triad of low mood, low energy levels and anhedonia, which can all be cardinal symptoms in a younger population. The absence of an affective component (i.e. low mood) can lead to healthcare professionals disregarding the potential for the presence of depression and consequently not exploring for other depressive symptoms.

Furthermore, symptoms of depression, especially somatic ones, are often attributed to physical illnesses. Depressive somatic symptoms often lead to a diagnosis of depression being over looked; such symptoms 'mask' the clinical diagnosis of depression and hence the term 'masked depression' (Small, 1991). Depressive somatic symptoms are often attributed to physical illness and/or frailty by both the individual and healthcare professional (Friedman et al., 2005a). Healthcare professionals should have a low threshold for the presence of depression in older adults, especially those with physical comorbidities.

Further complicating diagnostic difficulties and under-recognition is the fact that older adults are less likely to report any symptoms associated with mental health problems and ask for help in the first place (Evans and Mottram, 2000, Crabb and Hunsley, 2006, Mitchell et al., 2010c); explanations for this include older adults being less emotionally open, having a sense of being a burden or nuisance, and believing symptoms are a normal part of ageing or secondary to physical illness (Evans and Mottram, 2000, Birrer and Vemuri, 2004, Alexopoulos, 2005, Mitchell et al., 2010c). Older adults also have a reluctance to report mental health problems due to their perception of associated stigma; many older adults hold the view the mental health problems are shameful, represents personal failure and leads to a loss of autonomy (Evans and Mottram, 2000).

There is an overlap between symptoms of depression and symptoms of dementia. The concept of pseudodementia has been discussed above. It is quite common for older adults with dementia to initially present with depressive symptoms. Depression has a high incidence in those with dementia, especially those with vascular dementia. Dementia is particularly difficult to diagnose in dementia due to communication difficulties; diagnosis is often based on observed behaviours (Alexopoulos et al., 2005, Alexopoulos, 2005).

Depression and comorbidity in older adults

Older adults who are depressed are more likely to have existing physical health conditions and more likely to develop physical health conditions (Chapman and Perry, 2008).

In those with pre-existing physical health problems, depression is associated with deterioration, impaired recovery and overall worse outcomes (Evans, 1995). For example, the relative risk of increased morbidity related to coronary heart disease is 3.3 in comparison to individuals without depression (Aromaa et al., 1994). Mykletun et al. established that a diagnosis of depression in older adults increased mortality by 70%

(Mykletun et al., 2009). Several causative routes account for poor physical illness outcomes. Older adults with depression are less likely to report worsening health. Depressive symptomatology indirectly affects physical illness through reduced motivation and engagement with management. Reduced motivation is often secondary to feelings of helplessness and hopelessness. Poor compliance with management advice, notably adherence to medications is observed (Evans et al., 1997). Lack of engagement with diet and exercise advice can also be noticeable. Feelings of hopelessness, helplessness and negativity will contribute to the failure to seek medical attention in the first place or report worsening health when seen by a healthcare professional. Such feelings may also contribute to poor compliance with management. Other symptoms of depression, such as reduced energy levels, impaired memory and impaired executive function can cause self-neglect of physical illness.

Depression affects biological pathways directly, which impairs physical recovery. Such biological effects include pro-inflammatory factors, metabolic factors, impact upon the hypothalamic-pituitary axis and autonomic nervous system changes (Katon, 2011).

Depression is particularly associated with specific physical illnesses; cardiovascular disease and diabetes mellitus. A study by Win *et al.* found that cardiovascular mortality is higher in older adults with depression because of physical inactivity; the study established that physical inactivity was accountable for a 25% increased risk in cardiovascular disease (Win et al., 2011). The relationships between depression and cardiovascular disease and depression and diabetes have been described as "bidirectional" (Katon, 2011).

Individuals who have depression are more likely to develop physical illnesses compared to those without depression. Higher incidents of cardiovascular disease and diabetes mellitus are seen in people - regardless of age - with depression. For example, a study by Brown et al. found that older adults with depression had a 1.46 relative risk increase for developing coronary heart disease compared to those without depression (Brown et al., 2011). The hypothalamic-pituitary axis dysfunction found in depression leads to increased levels of cortisol, which in turn, increases visceral fat. Increased visceral fat is associated with increased insulin resistance, promoting diabetes mellitus, and increased cardiovascular pathology (Katon, 2011).

Depression is a risk factor for the subsequent development of dementia; this is especially so if an older adult has no previous history of depression (i.e. depression is late-onset) (Alexopoulos et al., 2000). Treatment of depression will lead to an improvement in cognitive deficits, whether those cognitive deficits are secondary to pseudodementia or dementia.

Depression is associated with maladaptive health risk behaviours - e.g. substance misuse, a sedentary lifestyle - which, increase the risk of physical health problems. Depression can lead to physical health problems in the absence of maladaptive health risk behaviours; for example, a recent study by Rodic et al. found that individuals with depressive symptoms were at greater risk of developing any physical illness compared to individuals without depression symptoms; a 1.67 odds ratio ($p<0.001$) was calculated (Rodic et al., 2015). On further breakdown of results, a significant odds ratio (1.79) was established for the development of arthrosis and arthritis in those with depressive symptoms ($p<0.05$) (Rodic et al., 2015).

Healthcare utilisation and economic impacts

Older adults are less likely to report depressive symptoms to healthcare professionals explaining the under-utilisation of mental health services for depression (Speer and Schneider, 2003, Crabb and Hunsley, 2006). Despite older adults under-utilising mental health services they over utilise other healthcare services (Rinaldi et al., 2003, Speer and Schneider, 2003). For example, those presenting with non-specific medical complaints or somatisation have been found to have an increase use of healthcare services. Non-specific medical complaints and somatisation lead to an unnecessary use of resources, such as unnecessary consultations with healthcare professionals and investigations (Speer and Schneider, 2003). Increase in service utilisation means an increase in the associated economic cost of depression in older adults (Speer and Schneider, 2003, Weeks et al., 2003, Friedman et al., 2005a).

There is little difference in associated costs of depressive disorder and sub-threshold depression (Katon et al., 2003, Cuijpers et al., 2006, Cuijpers et al., 2007). One study found that compared to non-depressed older adults, depression was associated with a 43-51% increase in all healthcare costs (Katon et al., 2003). Healthcare costs of older adults with a comorbid physical illness and depression are far greater than those without depression – findings in diabetes mellitus are a good example (Finkelstein et al., 2003). A study by Unutzer et al. found that older adults with any type of chronic physical disease and depression had higher healthcare costs ($22,960) compared to those with just chronic physical disease ($11,956); the majority of the increased healthcare cost was associated with the chronic physical disease and not the care and treatment of the depression (Unutzer and Schoenbaum, 2009). Poor compliance with physical illness management is associated with missed appointments and a greater number of hospital admissions, which both have financial implications.

Aetiology and associations of depression in older adults

Depression in older adults is a more complex condition than that observed in younger adults, which contributes to low rates of detection. Depression in older adults is viewed by some as a different clinical entity than depression in younger adults. Depression in older adults is associated with structural changes to the brain. The consequences of failing to recognise depression in older adults can have more detrimental impact upon physical and mental health. These points stress the importance and need for improvements in recognizing depression in this population. This section will present further details regarding associated neurological structural changes and relations to physical health.

Late-onset type depression in older adults has been associated with the term 'vascular depression' (Baldwin, 2000, Sneed et al., 2008b, Sneed et al., 2011). Studies have found a significant higher rate and severity of white matter hyperintensities on MRI imaging in older adults with depression compared to those without depression (Hickie et al., 1997, Hickie et al., 2003, Sneed et al., 2011). White matter hyperintensities represent damage to the nerve cells; such damage is a result of hypo-perfusion of the cells secondary to small blood vessel damage (Debette and Markus, 2010). White hyperintensities are associated with vascular risk factors (e.g. age, hypertension, hypercholesterolaemia, obesity, diabetes mellitus, smoking) and are linked to cerebrovascular disease, such as stroke, vascular dementia. A relationship has been found between psychosocial stress and consequent development of vascular risk factors, which further supports the hypothesis of 'vascular depression' (Sneed et al., 2011). Clinically, 'depression-executive dysfunction syndrome' and psychomotor retardation are associated with vascular changes (Beheydt et al., 2014).

In older adults with depression, white matter hyperintensities are associated with structural changes to corticostriatal circuits and subsequent executive functional deficits. Loss of motivation or interest and cognitive impairment in depression are hallmark features of structural brain changes associated with the frontal lobes, which in turn are associated with a vascular pathology (Rapp et al., 2005). A study by Hickie et al. established that white matter hyperintensities in older adults with depression are associated with greater neurological impairment and poorer response to antidepressant treatment (Hickie et al., 1997). It is not fully understood why vascular depression responds less well to antidepressants; poor response has been linked directly to vascular factors but has also been associated with deficits in executive function (Sneed et al., 2008a, Sneed et al., 2011).

The relationship between cerebrovascular disease and depression is described as 'bi-directional' (Baldwin, 2000, Gothe et al., 2012); depression has been found to cause cardiovascular disease and vice versa (Gothe et al., 2012). Baldwin et al. direct the reader to the presence of post-stroke depression and the occurrence of depression in vascular dementia (Baldwin, 2000).

Younger and older adults share a number of fundamental risk factors for depression; such as female gender, personal history and family history (Evans and Mottram, 2000). Common risk factors that are found in the younger adults apply to an older population but, due to advancing age, older adults are more likely to encounter and experience them. Older adults have additional risk factors related to ageing, which are not just physiological in nature.

Age related changes:

Age related changes occurring in the endocrine, cardiovascular, neurological, inflammatory and immune systems have been directly linked to depression in older adults (Fiske et al., 2009).

The normal ageing process sees changes to sleep architecture and circadian rhythms with resultant changes to sleep patterns (Van Someren, 2000). Thus sleep disturbances are common in older adults and positively correlated to advancing age (Van Someren, 2000); over a quarter of adults over the age of 80 years report insomnia, and research has well-established that this is a risk factor for depression (Cole and Dendukuri, 2003, Pigeon et al., 2008). A meta-analysis by Cole *et al.* found sleep disturbances to be a significant risk factor for the development of depression in older adults (Cole and Dendukuri, 2003).

Sensory impairment:

Sensory impairments, whether secondary to the ageing or a disease process, are risk factors (Cole and Dendukuri, 2003, Huang et al., 2010). Research has found that hearing and vision impairments are linked to depression (Bernabei et al., 2011). A sensory impairment can lead to social isolation and withdrawal, which, in turn, are further risk factors for depression.

Physical illness:

Physical illness, regardless of age, is a risk factor for depression. Older adults are more likely to have physical illnesses and so in turn are more at risk of depression. See Table 2. Physical illness is associated with sensory impairments, reduced mobility, impairment in activities of daily living and impaired social function, all of which can lead to depression. Physical illnesses associated with chronicity, pain and disability pose the greatest risk for the subsequent development of depression (Evans and Mottram, 2000, Cole and Dendukuri, 2003, Huang et al., 2010). It is known that pain worsens depressive symptoms and vice versa (Alexopoulos, 2005). Physical illness affecting particular parts of the body, such as the cardiovascular, cerebrovascular and neurological systems, are more likely to cause depression (Fiske et al., 2009). Essentially, however, any serious or chronic illness can lead to the development of depression. It should be noted that a large proportion of older adults have physical illness but do not experience depression symptoms, therefore other factors must be at play (Harpole et al., 2005, Fiske et al., 2009).

Treatments of physical illness are directly linked to aetiology in depression, for example, certain medications are known to cause depression; cardiovascular drugs (e.g. Propranolol, thiazide diuretics), anti-Parkinson drugs (e.g. levodopa), anti-inflammatories (e.g. NSAIDs), antibiotics (e.g. Penicillin, Nitrofurantoin), stimulants (e.g. caffeine, cocaine, amphetamines), antipsychotics (e.g. Haloperidol), anti-anxiolytics (e.g. benzodiazepines), hormones (e.g. corticosteroids), and anticonvulsants (e.g. Phenytoin, Carbamazepine) (Evans and Mottram, 2000, Birrer and Vemuri, 2004). Polypharmacy is present in many older adults further increasing the risk of depression. Pharmacokinetic and pharmacodynamic age related changes also contribute to an increased risk of medication induced depression in older adults.

Cardiovascular	**Endocrine**	**Cerebrovascular/neurological**
Ischaemic heart disease Myocardial infarction	Addison's disease Cushing's disease Hypothyroidism Hyperthyroidism Diabetes mellitus Hypoglycaemia	Cerebral arteriosclerosis Cerebral infarction Intracranial tumour Parkinson's disease Multiple sclerosis Temporal lobe epilepsy Dementia
Metabolic	**Autoimmune disorders**	
Electrolyte abnormalities • Hypernatraemia • Hypercalacaemia • Hyperkalaemia • Hypokalaemia Folate deficiency Thiamine deficiency	Rheumatoid arthritis Systemic lupus erythematosus Pernious anaemia	

Table 2: Table of physical illnesses associated with depression
(Evans and Mottram, 2000, Fiske et al., 2009)

Dementia:

Dementia is common in old age and those with dementia are at higher risk of developing depression compared to those who do not have it (Conradsson et al., 2013). 20-30% of older adults with Alzheimer's disease have depression (Tsuno and Homma, 2009). Depression is a risk factor for the subsequent onset of dementia.

Psychosocial:

When compared to younger adults, older adults are at a greater risk of developing depression due to the increased likelihood of experiencing particular psychosocial stressors, in particular adverse life events. Stressors include lack of social support, social isolation, loneliness and financial hardship. Financial hardship and functional impairment often sees older adults downsizing in property. Deteriorating physical health often sees older adults no longer being able to manage living independently at home necessitating a move into institutional living. Bereavement, especially spousal, and the associated role change that follows this are risk factors for depression (Fiske et al., 2009).

Sub-threshold depression:

Sub-threshold depression is an established risk factor for major depression. If these symptoms were reported and recognised sooner the rate of conversion of sub-threshold depression to major depression could be reduced (Cherubini et al., 2012). However older adults are less likely to report such symptoms and ask for help in the first place (Evans and Mottram, 2000, Crabb and Hunsley, 2006).

Prevalence and epidemiology

The prevalence of depression in older adults in England and Wales was found to be 8.7% in 2007; however, if those with dementia are included this figure rises to 9.7% (McDougall et al., 2007). A meta-analysis by Luppa et al. established a 7.2% point prevalence of major depression and a 17.1% point prevalence of depressive disorder in older adults (Luppa et al., 2012). The projected lifetime risk of an older adult developing major depression by the age of 75 years old is 23% (Kessler et al., 2005). It has been found that one in four older adults experience depressive symptoms that require treatment (Royal College of General Practitioners, 2002).

Sub-threshold depression is 2-3 times more prevalent than major depression in older adults (Rinaldi et al., 2003, Cherubini et al., 2012). These depressive symptoms are often clinically relevant (Rinaldi et al., 2003, Birrer and Vemuri, 2004, National Ageing Research Institute, 2009). 8-10% of older adults per year with sub-threshold depressive symptoms go onto develop a major depressive episode (National Ageing Research Institute 2009, Meeks et al., 2011); sub-threshold depression is a known risk factor for the development of a major depressive episode (Cherubini et al., 2012).

There is conflicting evidence as to whether the incidence of depression increases with age. The World Health Organisation found that the incidence of major depression decreases with advancing age; whereas the incidence of clinically relevant depressive symptoms increases (World Health Organization, 2001a). However, a study by Conradsson et al. found that those over the age of 80 years are at the greatest risk of becoming depressed and a cohort study by Solhaug et al. found the highest incidence of depression (9.6%) in adults aged 86 – 90 years (Solhaug et al., 2012, Conradsson et al., 2013).

As with depression in younger adults incidence and prevalence are greater in women; 10.4% of women over the age of 65 years have depression compared to 6.5% of men (McDougall et al., 2007). Older women are more likely to experience recurrent episodes of depression compared to older men (Kessler et al., 1994). The gender gap in incidence and prevalence becomes narrower with increasing age (Fiske et al., 2009). It should be acknowledged however that women are more likely to present to healthcare services and seek help in comparison to men (Oliver et al., 2005, Mackenzie et al., 2006).

The prevalence of major depression in older adults varies by setting (Gellis, 2014). Highest rates are seen in long-term institutional care and inpatient hospital settings (Evans and Mottram, 2000, Djernes, 2006). For example, a cohort study established the prevalence of major depression to be 9.3% (95% CI 7.8 – 10.9) for older adults living at home and 27.1% (95% CI 17.9 – 36.3) for older adults living in institutional care (McDougall et al., 2007).

A recent meta-analysis established the prevalence of major depression and depressive symptoms in older adults in long-term care and found rates of 10% and 29% respectively (Seitz et al., 2010). One study of prevalence rates of major depression and depressive symptoms in primary care found rates of 6.5 – 9.0% and 10 – 25% respectively (Weyerer et al., 2008). Table 3 summaries prevalence rates of major depression by setting.

Setting	Prevalence rate (%)
Community	5 – 10
Primary care	10 – 30
Hospital inpatient	11 – 50
Long-term institutional care	10 – 43

Table 3: Prevalence rate of major depression by setting (Evans and Mottram, 2000, Djernes et al., 2006)

Depression is an important issue for older adults and is expected to become more important in the future as the population of people over the age of 65 years is expected to increase; based on UK population figures in 2012, The Kings Fund has estimated that by 2032 the proportion of older adults aged 65-84 years old will have increased by 39% whereas the proportion over the age of 85 years will have increased by 106% (Kings Fund, 2014). By 2020 it is estimated that depression will be the second leading cause of disability in the world regardless of age (Mathers and Loncar, 2006). This increase in population will consequently see the incidence and prevalence of depression rise.

Recognising, and so diagnosing, depression in older adults will become more important because of a greater demand on existing healthcare services and provisions, due to physical health consequences, impact upon healthcare utilisation and greater economic healthcare costs mentioned previously.

Prognosis of depression in older adults

Depression in older adults is associated with a slower rate of recovery (Arean and Ayalon, 2005) and worse clinical outcomes compared to younger adults (Fiske et al., 2009). Depression in older adults is associated with higher relapse rates (Mitchell and Subramaniam, 2005). Worse prognosis in older adults correlates with advancing age, physical comorbidities and functional impairment (Licht-Strunk et al., 2005). The structural brain changes associated with depression in older adults are linked, as discussed, to poorer treatment response.

Morbidity and mortality associated with depression can be described as primary or secondary; primary morbidity and mortality arises directly from the depressive illness; whereas secondary morbidity and mortality arises from physical health problems, which are secondary to depression. The greater morbidity and mortality of physical illnesses present in older adults with depression means greater associated financial costs compared to older adults with physical illnesses who do not have depression (Katon et al., 2005).

Outcomes from sub-threshold depression are on par with those of major depression; however sub-threshold depression which develops into major depression is associated with worse outcomes (Cherubini et al., 2012).

Proportionally more people over the age of 65 years commit suicide compared to younger people (Rodda et al., 2011). Depression is the leading cause of suicide in older adults (Birrer and Vemuri, 2004, Rodda et al., 2011); one study reports that 75% of older adults who killed themselves were depressed (Sawyer, 2012).

The vast majority of older adults who commit suicide have had contact with a health professional within the preceding month (Arean and Ayalon, 2005); this figure has been quoted as high as 70% (Fiske et al., 2009). This further supports and suggests the fact the depression is under-detected. Unlike younger adults, older adults are less likely to report suicidal ideation and can experience suicidal ideation without feeling low in mood (Fiske et al., 2009, Evans and Mottram, 2000). Older adults have few suicide attempts, compared to younger adults, because their suicide methods are more lethal (Alexopoulos, 2005).

Screening for depression could have substantial effects on morbidity and mortality rates for older adults. Screening for depression could improve prognosis because individuals will be diagnosed earlier and so treatment commenced. In turn this may reduce suicide rates considerably. In some cases, screening may lead to individuals being diagnosed where depression may have gone undetected otherwise.

SCREENING

The purpose of screening is not to diagnose people with a disease. The purpose of screening is to identify individuals who require diagnostic investigation. Having a positive screen result does not mean that a disease is present; a positive screen result means that there is more likelihood the disease is present than if a negative result had been found. A screening test is not a substitute for clinical assessment.

Screening for depression may be of value in the older adult population. Screening could improve detection rates for depression in older adults; without a screening test being administered depression may not be identified.

The value of selective screening (i.e. case-finding) for depression in individuals with long-term physical health problems and those with dementia has been recognized and is now recommended for use in clinical practice (National Institute of Care Excellence, 2009). Many older adults would fall within the remit of this guidance. Despite the consequences of depression in older adults the value of a depression screening programme is unknown.

Principles of screening

The Oxford dictionary defines the term screen as 'a system of checking a person or thing for the presence or absence of something' (Oxford Dictionary, 2015). The UK Screening Portal defines medical screening as '*a process of identifying apparently healthy people who may be at increased risk of a disease or condition'*. It goes on to say that such identified people can be offered information, further tests and interventions to reduce their risk. Risk can refer to the development of a particular disease in the first place, the development of complications from a disease, side-effects experienced from a treatment, etc. Screening enables at-risk people to be identified early allowing intervention earlier to prevent or reduce negative outcomes that have yet to occur. In essence early identification means early intervention. Screening can only reduce risk and cannot eradicate it (Public Health England, 2013).

In 1968, at the request of the World Health Organisation (WHO), Wilson and Jungner described the criteria for screening (Wilson and Jungner, 1968), which is still held today;

1. The condition should be an important health problem
2. There should be an accepted treatment for patients with recognised disease
3. Facilities for diagnosis and treatment should be available
4. There should be a recognisable latent or early symptomatic stage
5. There should be a suitable test or examination
6. The test should be acceptable to the population
7. The natural history, including development from latent to declared disease, should be adequately understood
8. There should be an agreed policy on whom to treat as patients
9. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole
10. Case-finding should be a continuing process

The National Screening Committee (Public Health England) has developed the WHO screening criteria further and published their own guidance criteria regarding the viability, effectiveness and appropriateness of a screening programme (Public Health England, 2013, updated 2015). The National Screening Committee guidance states that screening should be both clinically and cost effective. The guidance is comprised of domains that refer to the clinical condition of interest, the screening test, the treatment intervention for a positive screen result, the effectiveness of screening and implementation. There is a total of 20 criteria, which cover these domains. See Appendix 1 for the full guidelines.

Screening vs. case-finding:

Screening involves applying a screening test to everyone in a particular population. Whereas case finding, also known as 'selective screening', is a more targeted approach; only people who are known to be at greater risk of a condition undergo a screening test. In terms of older adults, cardiovascular and cerebrovascular disease are associated with an increased risk of depression (Drayer et al., 2005, Alexopoulos, 2005, Fiske et al., 2009) and so case-finding for depression could involve only applying screening tests for older adults who have cardiovascular and cerebrovascular disease.

In the UK, NICE recommends that case-finding for depression is undertaken in those with chronic physical health problems (National Institute of Care Excellence, 2009), which, as discussed, would involve a large majority of older adults. A similar case-finding approach occurs in New Zealand; the New Zealand Guidelines Group recommends the use of brief screening tools for people with chronic illness, previous history of mental illness or suicide attempt, or recent significant loss. The guidance specifically states the implementation of case-finding in high risk groups, such as older adults in residential care (New Zealand Guideline Group, 2008).

Sensitivity and specificity:

The ability of a screening test to identify people who are more likely to have a disease is crucially important to its functional purpose. Results of a screen can be positive or negative; positive means increased chance that the individual has the disease; whereas negative means less chance. A positive screen result should precipitate a diagnostic investigation/test to confirm the presence of disease.

Prior to a screening process it is not known if a person has a disease or not. As explained above it is hoped that screening detects such people, but not everyone with the disease will be detected via a positive screen result because some people who have a negative screen result will in fact have the disease. This is referred to as false negative results. Ideally for screening to be justifiable the number of false negatives should be low and the number of true positives should be high. True positive means that a person has screened positive and investigation has confirmed the presence of disease and hence a diagnosis is made. Sensitivity is the measure that refers to the ability of screening to detect true positives (calculated as true positives N / (true positive N + false negative N)).

Specificity on the other hand is the measure that refers to the ability of screening to detect true negatives i.e. people who screen negative and do not have the disease (calculated as true negative N / (true negative N + false positive N)). For a good screening test, the number of true negatives will be high and the number of false positives will be low. A false positive result on screening means that a person has undergone an unnecessary diagnostic investigation (indicated by screening result) as they do not have the disease. An unnecessary diagnostic investigation is an unnecessary expense. See Table 4.

	Disease present	**Disease not present**
Positive screening result	True positive (a)	False positive (b)
Negative screening result	False negative (c)	True negative (d)
	Sensitivity = a/(a+c)	**Specificity** = d/(b+d)

Table 4: Sensitivity and specificity calculations

The sensitivity and specificity measures of a screening test demonstrate how good it is. Such measures are calculated for all screening tests to evaluate their diagnostic accuracy. As outlined by Public Health England guidance criteria, the distribution of test values should be identified and a suitable cut-off level defined (Public Health England, 2015).

Benefits and risks of screening for depression in older adults

The National Screening Committee states that the introduction of a screening programme should 'do more good than harm' (Public Health England, 2013, updated 2015). Thus for a decision to be made about introducing screening for a given clinical condition there has to be balanced consultation and consideration for the associated pros, positives and benefits against the cons, negatives and limitations. The acceptability of the screening test also must be considered. As discussed, no screening programme for depression in older adults currently exists. This section aims to present a balanced argument for the topic.

Arguments against screening for depression in older adults

Screening can reduce an individual's risk of developing a particular condition but it does not provide complete protection against future development, for example a person may develop depression following screening. A negative screening result does not equate to a condition being prevented. The outcome of a screening test applies only to the 'here and now'. Healthcare professionals, and even the individual undergoing screening, can be too reassured and comforted by a negative result that they pay no attention to ongoing preventative measures.

A major harmful outcome of screening is that the screening test may not be accurate, resulting in a false positive or false negative result. A false negative result leads to unwarranted inaction and inappropriate follow-up actions. Those with a false negative result should be undergoing confirmatory diagnostic testing. A false negative result can delay diagnosis, which could worsen prognosis because treatment is also delayed. A false negative result means that healthcare professionals, and the person undergoing screening, are given incorrect reassurances that they do not have a condition. This may lead to future non-attendance for further screening. As discussed in this section failure to diagnose and delayed diagnosis in depression in older adults is associated with worse outcomes.

A false positive result causes unnecessary distress and worry to an individual (and relatives) as they will believe they have the condition. There is often a delay between the false positive result of screening and the negative result of confirmatory tests, which further adds to distress and worry. Confirmatory tests are unnecessary because the screening result is incorrect. As well as unwarranted distress and worry to individuals, false positive screening tests incur unwarranted and unnecessary financial costs and healthcare time expenditure. A false positive screening test can also mean that an individual is given an inappropriate label. Confirmatory diagnostic tests for depression involve clinical interviews and instruments, which cause no harm to the individual with a false positive test result on screening however.

Another issue with screening is that of ‘over diagnosis’. The term refers to individuals diagnosed that may never need treatment. Like false positive screening results, it causes unnecessary distress and worry. In reference to depression in older adults, ‘over diagnosis’ could be the identification of individuals with sub-threshold depression that would have improved with time and never progressed to a depressive disorder, or the identification of individuals with depressive disorder that would have improved without treatment. However, the natural history of sub-threshold depression is not fully understood and it is not possible to predict which individuals will procede to develop a depressive disorder. It is also not possible to predict whether individuals with depressive disorder will improve without treatment. The impact of 'over diagnosis' in depression in older adults is not fully understood.

A specific negative issue of screening for depression is that it can impair the development of rapport between the healthcare professional and the patient. Asking a list of screening questions can be seen as impersonal and not person-centred care, which has a detrimental effect to the doctor-patient relationship. Another issue is that when screening takes the form of asking a set of specific questions, as in the case of screening for depression, people may give responses that they feel are wanted/socially desirable, which may not be truthful, which in turn introduces bias.

Finally, screening for depression may improve recognition, and so increase diagnostic rates, but this may not increase the knowledge of the condition by healthcare professionals. Depression in older adults is under-recognised because it is a condition that is poorly understood. A screening programme for depression will not primarily aim to educate healthcare professionals to fill this knowledge gap.

Arguments in favour of screening for depression in older adults

Research has found that early diagnosis and treatment of depression in older adults is vital (Chapman and Perry, 2008). Depression in older adults is associated, as discussed previously, with high morbidity and mortality, which includes disability, a decreased level of functioning and a reduction in quality of life (Weeks et al., 2003, Rinaldi et al., 2003, Birrer and Vemuri, 2004). The potential benefits of screening are due to the earlier detection of a condition. Clinical outcomes, health utilisation and economic costs associated with depression also highlight the importance of early detection, diagnosis and treatment.

Delayed diagnosis is associated with worse prognosis (Weeks et al., 2003). Earlier detection would mean earlier treatment and so a reduced period of untreated illness. Untreated illness is associated with distress, a reduced functional level and poor quality of life. Earlier detection may imply that symptom severity would not be as great, which may lead to lower economic treatment costs; this is based on the assumption that severity would be less severe at point of diagnosis, as diagnosis would have occurred earlier, compared to severity if no screening had taken place.

Earlier detection would reduce an older adult's risk of subsequently developing dementia or a worse state of physical health (due to poor management of existing physical health or the development of physical health conditions). Recognising and treating depression in older adults would improve physical health states, particularly in regards to co-morbid physical illness. Improved states of physical health may lead to overall reduced economic care costs.

Earlier detection of sub-threshold depression, which can be clinically significant, would reduce the rate of conversion to major depression and therefore this is also important. Earlier recognition would improve the quality of life of individuals with such symptoms (Korte et al., 2012).

Prior to any screening programme being introduced, in accordance with the National Screening Committee screening guidance, an effective treatment should be available. Effective treatment is available for depression in older adults; counselling, psychotherapy, psychotropic medications, in particular antidepressants, and elective-convulsive therapy (ECT) (National Institute of Care Excellence, 2016).

There is an evidence gap to whether a screening programme for depression in older adults would 'do more good than harm' because it is not known if the benefits of screening would outweigh the negatives.

Ideally a screening test will have an acceptable sensitivity and specificity, so that false negatives results will be minimal. A rate of false negative results will always exist though even in an effective screening programme where benefits outweigh risk (Petticrew et al., 2000). For a screening programme to be beneficial it has to be clinically effective; an effective screening programme is one where clinical outcomes for patients improve when compared against clinical outcomes in the absence of screening.

Screening for depression

Screening for depression takes the form of responses to a collective group of items, which are referred to as rating scales. Items can be direct open questions but are most commonly statements, which require a response in terms of agreement. Agreement can be measured by dichotomous categories (i.e. yes or no) or on a Likert scale.

Depression is a clinical disease that has attracted much attention in terms of evaluation of rating scales. As well as detecting depression, rating scales can provide a measure of severity. Most rating scales, and consequent research, have focused on the use of rating scales in the general population and so not in older adults. Examples of commonly used depression rating scales include the Beck depression inventory, the Hamilton depression rating scale (Hamilton, 1960), the Montgomery Asberg depression rating scale (Montgomery and Asberg, 1979) and the Patient health questionnaire (Spitzer et al., 1999).

Depression rating scales have been validated for use in a younger adults and so may not be appropriate for use in an older adult population because, as discussed, depression presents differently in older adults. This consequently led to the development of the Geriatric Depression Scale (GDS), which is the most well-known and commonly used depression rating scale in older adults (Yesavage et al., 1983). Brief versions of this could potentially be an acceptable test to use in a depression screening programme for older adults. A key criterion for a screening programme is a screening test that works (i.e. has acceptable diagnostic accuracy).

The Geriatric Depression Scale

The Geriatric depression scale (GDS) was developed in 1982 by Yesavage et al. (Yesavage et al., 1983). The original GDS consists of 30-items, which were selected from a pool of 100-items. These 100 items were generated by an expert panel of researchers and clinicians in old age psychiatry and geriatric medicine and were deemed to reflect depression in older adults. The final 30-items were selected for inclusion in the GDS because they showed the highest correlation with the total score of 100. See Table 5. Though the GDS does not include any items that reflect somatic symptoms of depression this was not planned; the 100 generated items did capture somatic symptoms but these items showed poor correlation and hence were not included (Yesavage et al., 1983).

Administration of the GDS can be performed by a healthcare professional reading out the items or self-administrated by an individual given a paper copy to complete. Response to the included items of the GDS is in a yes/no format and answers should be in reference to the past seven days (Montorio and Izal, 1996). A GDS score of 0-9 is deemed normal, a score of 10-19 is interpreted as mild depression and a score of 20-30 is interpreted as severe depression. In the original study, at a cut-off score of 11, a sensitivity of 84% and a specificity of 95% were found (Yesavage et al., 1983).

Excluding somatic symptoms of depression improves the sensitivity and specificity of the GDS because it is unlikely to misdiagnose somatic symptoms, secondary to physical illness, as being indicative of depression. Existing depression rating scales, which were not specifically developed for use in older adults, had a tendency to misdiagnose depression in older adults due to the inclusion of somatic symptoms.

The original study by Yesavage et al. was conducted in a sample of 100 older adults living in the community. The authors validated their new scale against two existing depression rating scales, which were not specific to depression in older adults; the Hospital Depression Rating Scale (HAMD) (Hamilton, 1960) and the Zung Self-Rating Scale (SRS) for Depression (Zung, 1965). A statistically significant positive correlation, indicating concurrent validity, was found between the three depression rating scales (p=<0.001); for severe depression the r value for the GDS, SRS and HAMD was 0.83 (Yesavage et al., 1983, Montorio and Izal, 1996).

Though the GDS is a popular rating scale is it not always practicable to use in clinical practice because it is time consuming to perform. In response to these problems the authors developed a 15-item version (GDS-15) in 1986; the authors simply selected the fifteen items from the GDS that correlated the highest with the total score from the generated 100 items. See Table 5. The GDS-15 takes 5-7 minutes to complete and was found to have a similar sensitivity and specificity to the original GDS (Yesavage and Sheikh, 1986). For the GDS-15, a score of 0-5 is deemed normal whereas a score >5 is suggestive of depression.

Several even briefer versions of the GDS all now available; 12-item, 10-item, 8-item, 5-item, 4-item and 1-item. However, briefer versions tend not to have standardised items like the GDS and GDS-15. The GDS was developed in America and so in the English language. All versions of the GDS have all been translated into other languages and are used throughout the world. The GDS reflects western values of society and may not be entirely appropriate to capture depression in non-Western countries however.

GDS items *indicates included in the GDS-15	**Items in bold = 1 point**	
1. Are you basically satisfied with your life?*	yes	**no**
2. Have you dropped many of your activities and interests?*	**yes**	no
3. Do you feel that your life is empty?*	**yes**	no
4. Do you often get bored?*	**yes**	no
5. Are you hopeful about the future?	yes	**no**
6. Are you bothered by thoughts you can't get out of your head?	**yes**	no
7. Are you in good spirits most of the time?*	yes	**no**
8. Are you afraid that something bad is going to happen to you?*	**yes**	no
9. Do you feel happy most of the time?*	yes	**no**
10. Do you often feel helpless?*	**yes**	no
11. Do you often get restless and fidgety?	**yes**	no
12. Do you prefer to stay at home rather than go out and do things?*	**yes**	no
13. Do you frequently worry about the future?	**yes**	no
14. Do you feel you have more problems with memory than most?*	**yes**	no
15. Do you think it is wonderful to be alive now?*	yes	**no**
16. Do you feel downhearted and blue?	**yes**	no
17. Do you feel pretty worthless the way you are now?*	**yes**	no
18. Do you worry a lot about the past?	**yes**	no
19. Do you find life very exciting?	yes	**no**
20. Is it hard for you to get started on new projects?	**yes**	no
21. Do you feel full of energy?*	yes	**no**
22. Do you feel that your situation is hopeless?*	**yes**	no
23. Do you think most people are better off than you are?*	**yes**	no
24. Do you frequently get upset over little things?	**yes**	no
25. Do you frequently feel like crying?	**yes**	no
26. Do you have trouble concentrating?	**yes**	no
27. Do you enjoy getting up in the morning?	yes	no
28. Do you prefer to avoid social occasions?	**yes**	no
29. Is it easy for you to make decisions?	yes	**no**
30. Is your mind as clear as it used to be?	yes	**no**

Table 5: The original GDS

Current evidence base of the Geriatric Depression Scale

To date there are five existing systematic reviews that explore the diagnostic accuracy of the GDS. There are several justifications for why a further systematic review should be performed. Firstly, all existing reviews have focused mainly on the original, 30-item GDS neglecting briefer versions, which are more practical for a clinical setting. Secondly, the most recent literature search was conducted in 2009 (Dennis et al., 2012) meaning it is six years out of date. Finally, all of the previous reviews have methodological limitations, as detailed below.

A major issue with data synthesis is that four of the existing reviews (Watson and Pignone, 2003, Wancata et al., 2006, Mitchell et al., 2010a, Mitchell et al., 2010b) have calculated pooled diagnostic data (i.e. sensitivity and specificity) for the GDS-15 regardless of cut-off score. This is not recommended and generates results that are imprecise and are difficult to interpret.

See Table 6 for 'A measurement tool to assess systematic reviews' (AMSTAR) ratings of methodological quality of the five existing systematic reviews (Shea et al., 2007). None of the five reviews had searched grey literature sources and therefore ignore unpublished data, creating a publication bias. Only one of the existing reviews used a standardised quality assessment of the primary studies (Mitchell et al., 2010a). The methodological limitations of these existing reviews contribute to the accuracy of reported diagnostic data.

Description of existing systematic reviews:

Watson and Pignone, 2003: This systematic review aimed to establish the diagnostic accuracy of all depression rating scales in primary care for older adults. Older adults were defined as being greater then 65 years of age. The search strategy included three electronic databases (MEDLINE, PsycINFO, Cochrane library), two clinical service guides and existing reviews. Grey literature was not searched. The search strategy was limited to English language papers only. Searches were performed from 1996 to January 2002. Two authors independently reviewed all abstract and full-papers. No tool to assess methodological quality of included papers was used. There was no assessment of publication bias. Meta-analysis was planned but was not possible due a limited number of studies being identified. 18 primary studies met inclusion criteria, which resulted in

1550 participants. The paper reports diagnostic data for the GDS and the GDS-15; for the GDS, sensitivity and specificity ranged from 79-100% and 67-80% respectively. For the GDS-15, sensitivity and specificity ranged from 82-100% and 72-82% respectively.

Wancata et al. 2006: The systematic review conducted by Wancata et al. focused on the GDS and the GDS-15. No definition regarding the age of an older adult is documented. Five electronic databases (MEDLINE, EMBASE, CINAHL, Psyndex, Cochrane library) were searched up until September 2014; it is unclear what year the searches started from. The reference lists of all included papers were reviewed to identify further studies for inclusion. Grey literature was not included in the search strategy. Language of publication was limited to English, French and German. The abstracts of identified papers were independently read by two authors. In total 173 full-papers were independently read by two authors and 42 of these fulfilled inclusion criteria. Primary studies based in psychiatric settings were excluded. Primary studies could report on all subtypes of depression; not just major depression. Where primary studies reported data for more than one cut-off score the authors used the cut-off score when sensitivity and specificity data was closest together. No tool of methodological quality was used to assess included studies. Study setting is the only subgroup analysis reported. Publication bias is not explored. In total, 33 primary studies were identified for the GDS and 15 were identified for the GDS-15. The total number of included participants was 6314. Mean sensitivity and specificity data are reported for the GDS and GDS-15 but the authors do not state how this was calculated. Mean sensitivity and specificity (regardless of cut-off score) of the GDS was 0.75 and 0.77 respectively. Mean sensitivity and specificity (regardless of cut-off score) of the GDS-15 was 0.80 and 0.75 respectively.

Mitchell et al. 2010a: This first systematic review by Mitchell et al. concerns the diagnostic accuracy of the GDS and GDS-15 in a primary care setting. Unlike the above systematic review older adult was defined as 55 years of age or older. The search strategy included three electronic databases (MEDLINE, EMBASE, Web of Knowledge) and four full-text collections (Science Direct, Ingenta Select, Ovid Full-text, Blackwell-Wiley Interscience). The data range of the search strategy was from inception to 2009. There is no reporting of limitations on the search strategy. A reverse citation search of key papers was also used but it is unclear what constituted key papers. Grey literature was not included. Papers were independently reviewed by two authors; however, data extraction was only performed by one author. There is no reporting of assessment methodological quality of included studies. In total 17 studies were identified; seven for the GDS and ten for the GDS-15, which resulted in 3012 and 1762 participants respectively. Meta-analysis was performed. The reported sensitivity and specificity of the GDS was 77.4% and 65.4% respectively. The reported sensitivity and specificity of the GDS-15 was 81.3% and 78.4% respectively.

Mitchell et al. 2010b: The second systematic review by Mitchell et al. also explores the diagnostic of the GDS and the GDS-15, but in addition it includes the GDS-5 and GDS-4, in a medical (both inpatient and outpatient) and nursing home settings. They defined older adults as 65 years of age or older. Type of depression was not limited to major depression, for example, primary studies which just identified the presence of some depressive symptoms were included. The same electronic databases and full-text collections were searched from inception to 2009. Grey literature was not included in the search strategy. Two authors independently reviewed identified papers; however data extraction and analysis were performed by only one author. Methodological quality of primary studies was assessed using the ‘Quality assessment of studies of diagnostic accuracy included in systematic reviews’ (QUADAS) tool (Whiting et al., 2003). In total 21 studies were identified for the GDS, 12 studies for the GDS-15 and three studies for the GDS-4 and GDS-5. The total number of included participants was unclear. Meta-analysis was performed for overall pooled diagnostic data, pooled diagnostic data specific to setting and also in accordance to cognitive function. For the GDS, overall pooled sensitivity and specificity were 81.9% and 77.7% respectively. For the GDS-15, pooled sensitivity and specificity were 84.3% and 73.8% respectively. Diagnostic data for the GDS-5 and GDS-4 were combined for the purpose of meta-analysis; pooled sensitivity and specificity were 92.5% and 77.2% respectively.

	Systematic review				
AMSTAR criteria	Watson and Pignone, 2003	Wancata et al. 2006	Mitchell et al. 2010a	Mitchell et al. 2010b	Dennis et al. 2012
Was an ‘a priori’ design provided?	🟩	🟩	🟩	🟩	🟩
Was there duplicate study selection and data extraction?	🟩	🟩	🟥	🟥	🟩
Was a comprehensive literature search performed?	🟥	🟥	🟥	🟥	🟥
Was the status of publication used as an inclusion criterion?	🟥	🟥	🟥	🟥	🟥
Was a list of studies (included and excluded) provided?	🟥	🟥	🟥	🟥	🟥
Were the characteristics of the included studies provided?	🟩	🟩	🟩	🟩	🟩
Was the scientific quality of the included studies assessed and documented?	🟥	🟥	🟥	🟩	🟥
Was the scientific quality of the included studies used appropriately in formulating conclusions?	🟥	🟥	🟥	🟥	🟥
Were the methods used to combine the findings of studies appropriate?	🟥	🟨	🟩	🟥	🟩
Was the likelihood of publication bias assessed?	🟥	🟥	🟥	🟥	🟥
Was the conflict of interest included?	🟥	🟥	🟥	🟥	🟥

YES NO UNCLEAR

Table 6: AMSTAR assessment of existing systematic reviews of the diagnostic accuracy of the GDS

Dennis et al. 2012: Like Wancata et al., the systematic review by Dennis et al. focused on several different depression rating scales and was not just specific to the GDS. Older adults were defined as being 60 years of age or greater. The study only explored diagnostic accuracy of rating scales in an inpatient setting. Three electronic databases, EMBASE, MEDLINE and PsycINFO, were searched from inception to 2009. The search strategy was limited to the English language only. Grey literature was not searched. Four 'key journals' were also searched; Age and Ageing, International Journal of Geriatric Psychiatry, Journal of the American Geriatric Society, American Journal of Geriatric Psychiatry and International Psychogeriatrics. Three of the study authors independently reviewed all identified papers. A total of 14 papers met inclusion and exclusion criteria. It is unclear how many people were involved in data extraction. The 14 papers resulted in a total of 1550 study participants. Data were synthesised by a pooled analysis; pooled sensitivity and specificity data is presented for different cut-off scores of the GDS and GDS-15. For the GDS-15, at the recommended cut-off score of 5, sensitivity was found to be 79% (95% CI 70-86%) and specificity was found to be 77% (95% CI 73-81%). For the GDS, at the recommended cut-off score of 10, sensitivity was found to be 85% (95% CI 78-91) and specificity was found to be 82% (95% CI 78-85%).

The methodological limitations of existing reviews concerning the GDS highlight the need, and justification, for another systematic review of diagnostic test accuracy of the GDS, in particular briefer versions.

The clinical effectiveness of screening for depression in older adults

There is an evidence gap as to whether screening for depression in older adults is clinically effective. This leads directly to the stance of the National Screening Committee that screening should 'do more good than harm'. For a depression screening programme in older adults to be introduced, it has to be established if screening improves clinical outcomes (and that this outweighs any associated harms).

Evidence of the clinical effectiveness of screening for depression in older adults is lacking as the topic has not been the focus of much research.

Current evidence base of the clinical effectiveness of screening for depression in older adults

Studies exploring the clinical effectiveness of screening for depression in older adults are limited in number. To date only one systematic review that explores the effectiveness of screening for depression in older adults has been performed. This review, by O'Connor et al, was undertaken in 2009 and only identified four primary studies (O'Connor et al., 2009). The population of interest for the review were adults >18years of age and older adults; the authors do not define the age of older adults though. O'Connor et al. limited the search strategy to a primary care setting and English language papers only. Grey literature was not included in the search strategy. It was not possible to perform a meta-analysis. O'Connor et al. concluded that screening for depression does not improve clinical outcome. See Table 7 for AMSTAR ratings of this review.

	Systematic review
AMSTAR criteria	O'Connor et al., 2009
Was an 'a priori' design provided?	🟩
Was there duplicate study selection and data extraction?	🟥
Was a comprehensive literature search performed?	🟥
Was the status of publication used as an inclusion criterion?	🟥
Was a list of studies (included and excluded) provided?	🟩
Were the characteristics of the included studies provided?	🟩
Was the scientific quality of the included studies assessed and documented?	🟩
Was the scientific quality of the included studies used appropriately in formulating conclusions?	🟥
Were the methods used to combine the findings of studies appropriate?	🟩
Was the likelihood of publication bias assessed?	🟥
Was the conflict of interest included?	🟥
	YES NO UNCLEAR

Table 7: AMSTAR assessment of the existing systematic reviews of the clinical effectiveness of screening for depression in older adults

CONCLUSION

The aim of this dissertation is to explore and investigate the use of screening for depression in older adults. As presented above, depression is an important clinic topic in older adults. In the future depression will become an even more pressing issue because the older adult population is ever increasing. Improvements in detection, thus diagnosis, could lead to better clinical outcomes, reduced healthcare utilisation and reduced direct and indirect economic costs associated with depression. Screening for depression in older adults could:

- detect the illness in those who do not present with depressive symptoms or seek help in the first place
- differentiate depression from other conditions (i.e. physical health conditions) when symptoms are overlooked
- reduce associated morbidity and mortality
    - of depression
    - ± of comorbid physical illness
- reduce associated economic cost
    - of depression
    - ± of comorbid physical illness

However, as mentioned, screening for depression does not come without drawbacks or ethical concerns.

The GDS is the most well-known and widely used depression rating scale for use in older adults. Despite this there is no up-to-date systematic review regarding the diagnostic accuracy and validity of the different brief versions available.

Screening for depression is not routine in clinical practice in the UK. There is little understanding about the benefit of screening for depression in an older adult population. Therefore, a systemic review will be conducted to investigate the clinical effectiveness of depression screening. Owing to the expected limited number of studies available to address this issue there will be no limit on the screening tool used i.e. there will be no limit to just include studies utilising the GDS.

AIMS AND OBJECTIVES

The aim of this dissertation is to explore the topic of screening for depression in older adults. The aim will be addressed through conducting systematic reviews, which will include meta-analyses where appropriate, in response to the following objectives:

1) Establish the diagnostic accuracy of brief versions of the GDS
2) Establish the clinical effectiveness of screening for depression in older adults

CHAPTER 2

The diagnostic accuracy of brief versions of the Geriatric Depression Scale (GDS) in older adults: a systematic review and meta-analysis

INTRODUCTION

For a screening programme to exist there has to a be a suitable test available to screen for the condition in question. This test also has to be acceptable to the population. Brief versions of the GDS could be suitable and acceptable tools to use for depression screening. They offer more clinical appeal as they take less time to administer. As discussed in Chapter 1, existing evidence regarding the accuracy of brief versions of the GDS is out of date and incomplete. Previous systematic reviews have several methodological limitations, which serve as further justification for the need to perform this review.

This chapter aims to establish the up-to-date diagnostic accuracy of the GDS-15 and briefer versions of the GDS. The data presented here is an expanded version of the published article in the International Journal of Geriatric Psychiatry. See Appendix 5.

Research question

What is the diagnostic accuracy of brief versions of the GDS in older adults?

METHOD

Protocol

In accordance with the Centre for Reviews and Dissemination (CRD) guidance, a protocol for the review was written (Centre for Reviews and Dissemination, 2008). See Appendix 2 for the protocol. Some of the searches were undertaken prior to the registration of the protocol and therefore, in line with CRD guidance, it was not registered.

Reporting

The Preferred Reporting of Items for Systematic Reviews and Meta-analyses (PRIMSA) guidelines were used as a basis for reporting (Moher et al., 2009).

Search strategy

A) Search terms

The search strategy comprised of three separate components, which were combined with the Boolean operator 'AND'. The three components of the search strategy were terms referring to older adults, depression and the GDS. These three components had to appear in the citation title or abstract. For terms referring to 'depression' a mixture of subject heading search terms (e.g. MeSH) and free-text terms were used because use of MeSH terms only would not have identified all relevant studies. When used, subject heading search terms were exploded.

Only free-text search terms were used to capture the concept of 'older adult'. Initially subject heading search terms for 'older adult' were going to be used; however, using such terms produced too many irrelevant results.

There are no subject heading search terms for the GDS and therefore just free-text search terms were used. Initially, broad subject heading search terms that refer to the concept of screening in general were going to be included in the search strategy but this produced too many results, which were unmanageable for the time scale of the review, and results that were not relevant. Preliminary, background literature searching revealed that there are numerous methods of referring to different versions of the GDS; for example, the 15-item version can be titled 'GDS-15', 'GDS 15', 'GDS15', 'GDS short' and the 'geriatric depression scale 15'. Terms using truncation were first piloted and the number of results between the truncation search and full-text terms were compared with no difference being found, therefore truncation symbols were used in the final search strategy to capture terms referring to the GDS.

The syntax of the search strategy was customised to the different electronic databases used. See Appendix 3 for search strategies.

B) Electronic databases

The databases MEDLINE, EMBASE, Cumulative Index to Nursing and Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts and Reviews of Effects (DARE) and the Health Technology Assessment (HTA) were searched from 1982 to April 2014. Searching seven different electronic databases meant that a wider range of coverage – thus a more comprehensive search - was provided meaning that all relevant studies were more likely to be found. The time frame of dates searched had a lower limit of 1982 because the GDS was developed in this year and therefore any papers found before 1982 would not be relevant to the review.

Apart from a limitation applied to the lower date range of 1982 no other limitations were applied to the search strategy (i.e. no limit to English-language only, 'only human', etc.). No filters were applied to identify studies of diagnostic accuracy as there is evidence that reliance on these in reviews of diagnostic accuracy studies misses relevant citations (Relevo, 2012, Beynon et al., 2013).

C) Unpublished and grey literature

To reduce publication bias in the review and to have utilised the most comprehensive search strategy possible unpublished and grey literature was also searched. An information technician was consulted in order to establish which resources were the most appropriate to use because the initial list was too extensive. Searching of unpublished and grey literature included the following resources; Conference proceedings via Web of Science, http://ethos.bl.uk, www.guideline.gov and www.opengrey.eu.

D) Additional search strategies

The clinical trials register was searched; www.clinicaltrials.gov.

The reference lists of previous systematic reviews (Watson and Pignone, 2003, Wancata et al., 2006, Mitchell et al., 2010a, Mitchell et al., 2010b, Dennis et al., 2012) were manually checked to identify further studies that may not have been identified through the search strategy. The reference lists of all included primary studies were also checked to identify further studies.

Prior to piloting the search strategy, inclusion of a reverse-citation search of the original, 1982 study article describing the development and validation of the GDS by Yesavage et al. was planned. However, this produced 5079 results in the database ScienceDirect and 4589 in Web of Science. Owing to time restraints and resources available the decision was made not to include this reverse-citation search.

Citation management

All citations that were identified through the search strategy were exported into the electronic reference and bibliography managerial software package Endnote (Thompson Reuters, 2016). Duplicate citations were removed using the automated command and manually so that only one copy of a study remained.

Study selection

A) Inclusion and exclusion criteria

'Population, intervention, comparator and outcome' (PICO) inclusion and exclusion criteria were developed and applied to each sift stage of citation screening (Richardson et al., 1995). First stage PICO criteria must have been met for the study to progress to a second sift stage. For a study to be included in the review all second sift criteria must have been met. These criteria were piloted prior to use.

If no abstract was available, the citation was judged on the basis of title alone.

Comparison of the GDS against a gold standard diagnostic instrument was a criterion because lack of a gold-standard would mean data regarding sensitivity and specificity would be inaccurate as bias would have been introduced. A gold-standard test is the best available evidence and accepted proof that a disease is present or absent. Gold-standard diagnostic instruments for depression were based on the International Classification of Diseases (ICD) or Diagnostic and Statistical Manual (DSM) classification of diseases. Such diagnostic instruments are structured and standardised, which improves diagnostic accuracy and reliability. In diagnostic accuracy studies results of the screening test are compared to known cases that have been diagnosed by a gold-standard test. Without the number of known cases of depression being established data regarding the screening test sensitivity and specificity of the GDS cannot be calculated. Diagnoses of disease not made by a gold-standard diagnostic test maybe inaccurate and unreliable, which can lead to over- or under-estimation of the diagnostic accuracy of the screening test.

Background reading identified that some studies classify older adults as being 55 years of age or older. In order to not exclude such studies, inclusion criteria regarding age of population was 55 years of age or older.

Inclusion and exclusion criteria were developed from PICO criteria and applied at each sift of screening. See Table 8.

PICO	Sift	Criteria
Population	First and second	• Sample referred to as 'older adults' • No restrictions in terms of ethnicity or country • No restrictions in terms of physical comorbidity and cognitive impairment
Instrument	First	• If citation referred to GDS then this criterion was met
	Second	• GDS must have been implemented • Version of GDS specified
Comparator/ reference standard	First	• Documented use of diagnostic interview or instrument in title or abstract • Reference to diagnostic accuracy implying use of gold standard instrument (i.e. diagnostic accuracy, sensitivity, specificity, likelihood ratios, diagnostic odds ratio, etc.) in title or abstract • Reference to Diagnostic and Statistical Manual (DSM) or International Classification of Diseases (ICD) diagnosis in title or abstract
	Second	• Gold standard diagnostic interview or instrument used and specified (e.g. structured clinical instrument for DSM disorders (SCID), composite international diagnostic interview (CIDI), diagnostic interview schedule (DIS)) • Unrecognised and unfamiliar diagnostic interviews or instruments were assessed on an individual basis to determine whether it was a 'gold standard'
Outcome	First	• No criteria to have been met
	Second	• Sufficient data to construct a 2x2 contingency table for the GDS vs. gold standard diagnostic instrument for the diagnosis of major depression
Study	First and second	• No restriction regarding type of study design

Table 8: First and second sift inclusion and exclusion criteria

B) Screening of citations

The author (Claire Pockilngton) screened all citations for inclusion eligibility based on title and abstract alone. For studies identified as eligible, the full-article was obtained to allow assessment against inclusion and exclusion criteria at the second sift stage of screening. Any uncertainty encountered by Claire Pocklington at the second sift stage was discussed with a supervisor (Dean McMillan). If any disagreement had been encountered this would have been resolved, in accordance with the protocol, by a consensus or failing this a second supervisor (Simon Gilbody) would have been involved.

If data were published more than once only one paper would be used where appropriate. However, if there were overlaps in study samples both papers would be included but would be cited as one study.

Authors of studies were contacted if additional information was required or if data were missing.

C) *Study selection process*

Citation articles were accessed through the University of York, University of Leeds and NHS electronic library systems. In total, 193 studies were selected at the second sift change. Access to electronic library systems only permitted access to 151 of these, leaving a total of 42 articles to find elsewhere. Difficulties were encountered with obtaining full article versions of these 42 studies because the electronic library systems that were accessible did not subscribe to certain journals.

Seven full article versions of studies were requested from The University of Leeds who had copies of the journal in physical format in storage.

Authors of the remaining studies were emailed where full-article versions were not obtainable from library sources. The email explained the purpose of wanting a full-copy and the difficulty in accessing it. If an email address was not included in the abstract the search engine Google was used to find an email address of one of the authors. If there was no response within seven days a different author was then contacted by email. Of the 35 missing full-version articles, nine authors responded and were able to provide an electronic copy of their study.

It was clear in some instances from the title, abstract and citation of the missing studies that they were not research papers but conference abstracts only. Conference abstracts would not provide sufficient information for inclusion; however, authors were emailed and asked if there was a corresponding research paper or if further details were available elsewhere. Four studies were conference abstracts and therefore did not provide sufficient information or data. The authors were contacted by email for additional information but did not reply and therefore the studies were excluded.

Full article versions were not found for 22 studies despite the approaches taken above. 19 of these articles were accessible through the British Library and so requested. The British Library was visited in person so that the journal articles could be obtained. Requests for the remaining three articles were made through the University of York library service at a cost.

Data extraction

The following data were extracted:

1) Author, date of publication
2) Descriptive characteristics of the setting (country, healthcare setting)
3) Descriptive characteristics of sample (age, ethnicity, proportion female, cognitive function)
4) Sample size and prevalence of major depression
5) Descriptive characteristics of the GDS (version, subset or non-subset, administration mode, administered by, language)
6) Descriptive characteristics of the gold-standard (diagnostic classification system used, name of instrument/test)
7) Data to construct a 2x2 contingency table (number of true positives, true negatives, false positives and false negatives in relation to diagnosis of major depression)

The author, Claire Pocklington, extracted all data. Any uncertainty encountered was discussed with the supervisor Dean McMillan; if uncertainty was not resolved a second supervisor, Simon Gilbody, would have been involved, as outlined in the protocol. Extracted data were directly recorded onto a spreadsheet.

Study quality assessment

The quality of all included studies was assessed using the QUADAS-II. The QUADAS-II is a quality assessment tool specifically designed for use in systematic reviews of diagnostic accuracy studies (Whiting et al., 2011). It assesses both risk of bias and applicability of studies. For this review the QUADAS-II was tailored for use for quality assessment of primary studies of the diagnostic accuracy of the GDS in keeping with the authors' recommendations. Each primary study was assessed against several criteria, which fall into four domains; patient selection, index test (i.e. GDS), gold-standard reference test and flowing and timings. Each criterion is rated in one of three ways; 'met', 'not met' and 'unclear'. Each domain is then given an overall rating for the presence of bias, which can take the possibilities of 'low, 'unclear' and 'high'. Applicability of patient selection, index test and reference rate are also rated as 'met', 'not met' and 'unclear'. See Table 9.

Domain	Criterion
Participant selection	a) Consecutive or random sample b) Avoids case-control/avoid artificially inflated base prevalence rates c) Avoid inappropriate exclusions
Index test	a) GDS interpreted blind to reference test b) Threshold pre-specified or multiple cut-offs reported c) If translated, appropriate translation d) If translated, psychometric properties reported
Reference test	a) Reference test correctly classifies target condition b) Reference test interpreted blind to GDS c) If translated, appropriate translation d) If translated, psychometric properties reported
Flow/timing	a) Interval of two weeks or less between GDS and reference test b) All participants receive same reference test c) All participants included in analysis
Applicability of patient selection	
Applicability of index test	
Applicability of reference test	

Table 9: QUADAS-II domains and criteria

The domain of participant selection explores whether bias was introduced into a primary study by the process of how participants were selected to take part in the study. 'Applicability of the patient selection' refers to how the participants in the primary studies should be similar to the target population of the systematic review. Both the GDS and gold-standard diagnostic test in the primary studies should have been applied and interpreted blind from one another so that the results of one did not influence the results of the other. Non-blinding of the GDS and gold-standard diagnostic test would have introduced bias. Ideally the cut-off score for the GDS (i.e. score at which the GDS is reported as positive) should be predetermined prior to the analysis to avoid overestimation of diagnostic accuracy. The domain of flow/timing refers to the time frame and analysis of results. Ideally the GS and gold-standard diagnostic test should both have been administered with an interval of two weeks or less in between; an interval of greater than two weeks reduces reliability of the initial measure because the clinical situation could have changed over time.

Data synthesis

For each primary study a two-by-two table was constructed for the different versions of the GDS and the different cut-off points used, which categorised study participants into true positives, false negatives, true negatives and false positives. Sensitivity is reflected by the number of true positives and false negatives (i.e. sensitivity = true positive N / (true positive N + false negative N)), whereas specificity is reflected by the number of true negatives and false positives (i.e. true negative N / (true negative N + false positive N)).

Heterogeneity

Heterogeneity was visually assessed by observing the overlap of confidence intervals in forest plots. However, it was also formally measured, for the GDS-15, using the I^2 statistic, which can be interpreted as the proportion of total variability explained by heterogeneity. For the GDS-15 and the different cut-off points used, diagnostic odds ratios were computed, which allowed exploration of between-study heterogeneity using the I^2 statistic. I^2 produces a measure of variability, which can range from 0 – 100%, with 0% meaning studies are completely homogeneous whilst 100% means there is complete heterogeneity between studies. Tentative thresholds are recommended for interpretation of the I^2 statistic (Higgins et al., 2011);

- 0 – 40% - heterogeneity may not be important
- 30 – 60% - may represent moderate heterogeneity
- 50 – 90% - may represent substantial heterogeneity
- 75 – 100% - may represent considerable heterogeneity.

Meta-analysis

Diagnostic meta-analyses for brief versions of the GDS were pre-planned and performed if there were a sufficient number of comparable studies. Owing to different studies using different cut-off points more than one diagnostic meta-analysis was performed. The statistical software programme Stata was used for data analysis. Stata requires a minimum of four studies to perform meta-analysis. Pooled estimates of sensitivity, specificity, positive likelihood ratio and negative likelihood ratio and diagnostic odds ratios were calculated (including confidence intervals) by bivariate meta-analysis. Summary Receiver Operating Characteristics (sROC) were calculated to produce 95% confidence interval ellipses within ROC space.

Funnel plots were constructed to examine the potential role of publication bias.

Subgroup analyses

Three subgroup analyses, which were pre-specified, were performed. These allowed comparison of the effects that participant age, study setting and country had on pooled results. Mean participant age for each study was used to classify primary studies into three subsets: young-old (65 – 74 years of age), middle-old (75 – 84 years) and old-old (≥ 85 years). Primary studies were divided into subsets depending on study setting: primary care, secondary care, community and residential/nursing home. Primary studies were also divided into two subsets depending on country: Western or non-Western country. For each subgroup pooled odds ratios were computed and meta-analysis re-run. This facilitated further exploration of heterogeneity.

Sensitivity analyses

Like subgroup analyses, sensitivity analyses were also pre-specified. Sensitivity analyses included examining the influence of the prevalence of major depression in primary studies and exploring the effect of primary studies extracting a brief version of the GDS version from a longer version (e.g. participants completed the GDS-30 but a score for the GDS-15 was calculated). For prevalence of major depression primary studies were divided into three subsets; prevalence of major depression <10%, 10 – 20% and >20%. For whether GDS versions were extracted from larger GDS versions primary studies were divided into two subsets; extracted GDS used or non-extracted GDS used. Again, pooled odds ratios were computed and meta-analyses re-run for the new groups. Sensitivity analysis also explored risk of bias for methodological domains of the prima studies in accordance with the QUADAS-II, such as participant selection, use and administration of the GDS, use and administration of the reference test and flow/timing of study design. For each QUADAS-II domain, meta-analysis was re-run excluding primary studies that were rated as having a 'high' or 'unclear' risk of bias.

Meta-regression

Meta-regression analysis of the logic diagnostic odds ratio was performed to explore the effects of more than one study characteristic on pooled summary estimates. Seven explanatory variables were explored. The number of explanatory variables was not dependent upon the number of studies included in meta-regression. Characteristics of the primary studies are viewed as explanatory variables in regards to the individual effect they have on the diagnostic odds ratio. Meta-regression permits statistical heterogeneity to be explored in further detail by developing a model to explore and explain how influential explanatory variables are.

RESULTS

The search strategy identified 11,418 records, which resulted in 6637 post-deduplication. 197 records met initial inclusion criteria on the basis of screening titles and abstracts alone. Full text copies of these were obtained and examined.

Of this 193, 166 studies were then excluded - see exclusion table in Appendix 4. See Figure 2 for a PRISMA diagram of study selection.

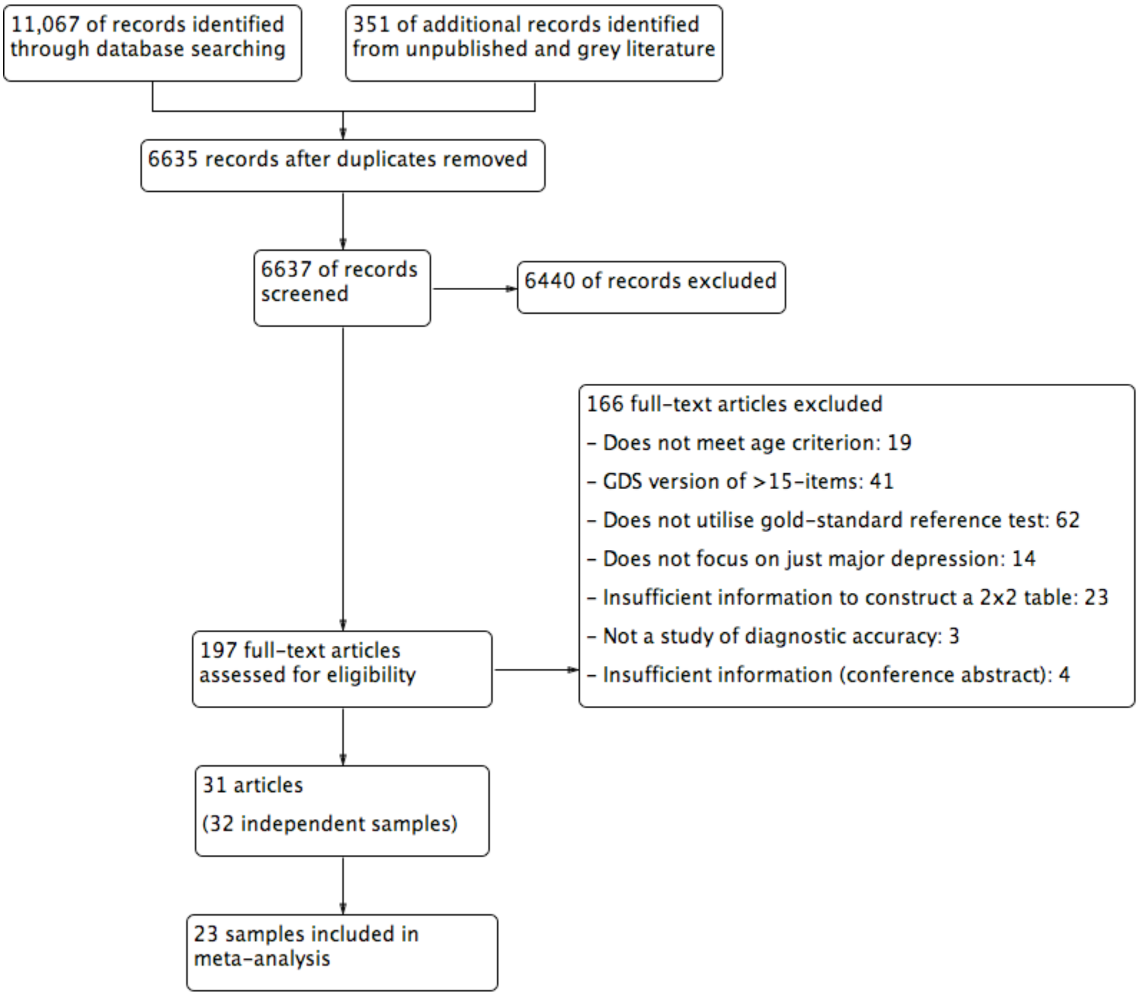
11,067 of records identified through database searching
351 of additional records identified from unpublished and grey literature
6635 records after duplicates removed
6637 of records screened
6440 of records excluded
166 full-text articles excluded
- Does not meet age criterion: 19
- GDS version of >15-items: 41
- Does not utilise gold-standard reference test: 62
- Does not focus on just major depression: 14
- Insufficient information to construct a 2x2 table: 23
- Not a study of diagnostic accuracy: 3
- Insufficient information (conference abstract): 4
197 full-text articles assessed for eligibility
31 articles
(32 independent samples)
23 samples included in meta-analysis

Figure 2: PRISMA diagram of study selection

There were six reasons for articles being excluded; does not meet age criterion (19 studies), GDS version >15 (41 studies), does not utilize gold-standard reference test (62 studies), does not focus on major depression (14 studies), insufficient information to construct 2x2 table (23 studies), not a study of diagnostic accuracy (3 studies) and insufficient information (conference abstract) (4 studies). See Appendix 4 for table of excluded studies.

The final 31 records resulted in 32 independent samples; two papers (Blank et al., 2004, Wongpakaran et al., 2013) separate sensitivity and specificity data for different study settings within the study and therefore each setting has been treated as a separate sample. Two samples (Allgaier et al., 2011, Broekman et al., 2011) both have two corresponding papers (Allgaier et al., 2013, Nyunt et al., 2009a) (respectively) that together provide complete information for the same sample. The sample by Allgaier et al. has two papers; one paper, published in English, provides diagnostic data regarding the GDS-15 and GDS-4, whereas the other, published in German, provides diagnostic data regarding the GDS-8 and GDS-4 (Allgaier et al., 2011, Allgaier et al., 2013).

The sample described in the studies by Broekman et al. and Nyunt et al. have overlapping samples but also different authorship; the paper by Broekman et al. has three authors, including Broekman, who are not authors on the Nyunt paper, which was published two years earlier. Nyunt, Niti and Pin are authors of both studies. The study by Nyunt et al. reports diagnostic data regarding the GDS-15. The study by Broekman et al. reports diagnostic data regarding the GDS-15 and the GDS-7. A discrepancy however was identified in the result sections and so the authors (Broekman and Nyunt) were contacted by email for clarification; Broekman et al. reported the sensitivity of the GDS-15, at a cut-off score of 5, as 97%, whereas Nyunt et al. reported the sensitivity as 96%.

The 32 samples amount to 13,141 participants. See Table 10 for study and sample characteristics.

Study	Sample characteristics	Sample size and % depressed	GDS characteristics	Diagnostic standard
Abas et al. (1998)	Country: UK Setting: primary care Age (yrs): Av. = 68.3 Ethnicity: African-Caribbean Female: 54.0% Cognition: 45% impaired (2% MMSE ≤9)	N = 164 Major depression: 20.0%	Version: 15 Administration mode: oral Administered by: interviewer Language: English	ICD GMS AGECAT
Allgaier et al. (2013)	Country: Germany Setting: community, nursing home Age (yrs): Av. = 84.5 (range 65 - 97) Ethnicity: not described Female: 73.9% Cognition: MMSE: ≥15 for inclusion. Mean MMSE 24.0	N = 92 Major depression: 14.1%	Versions: 15 with 8 and 4 subsets Administration mode: self-administration with assistance if required Administrated by: not stated Language: German	DSM-IV SCID
Almeida and Almeida (1999)	Country: Brazil Setting: secondary care, mental health outpatient clinic Age (yrs): Av. = 67.5 Ethnicity: not described Female: 84.4% Cognition: Mean MMSE 25.3	N = 64 Major depression: 64.1%	Version: 15 with 10, 4 and 1 subsets Administration mode: oral Administrated by: research team Language: Portuguese	ICD-10 ICD-10 Checklist of Symptoms
Arthur et al. (1999)	Country: UK Setting: primary care Age (yrs): Av. = 80.0 (range 77 - 83) Ethnicity: not described Female: 59% Cognition: median CAPE IO score 10	N = 201 Major depression: 6.0%	Version: 15 Administration mode: oral Administered by: practice nurse Language: English	ICD-10 SCAN
Bae and Cho (2004)	Country: South Korea Setting: secondary care, mental health outpatient clinic Age (yrs): Av. = 69.6 Ethnicity: not described Female: 65.0% Cognition: MMSE ≥15 for inclusion	N = 154 Major depression: 40.1%	Versions: 30 with 15 subset Administration mode: self-administration with assistance if required Administered by: research assistance if required Language: Korean	DSM-III-R DIS
Bijl et al. (2006)	Country: Netherlands Setting: primary care Age (yrs): Av. = 66.5 Ethnicity: not described Female: 64.2% Cognition: MMSE >18 for inclusion	N = 312 Major depression: 37.5%	Version: 15 Administration mode: oral Administered by: research assistant Language: Dutch	DSM-IV PRIME-MD

AMTS: abbreviated mental test score CAPE IO: Clifton assessment procedures for the elderly information/orientation
DIS: diagnostic interview schedule DSM: diagnostic and statistical manual of mental disorders
GMS AGECAT: geriatric mental state AGECAT ICD: international classification of diseases MMSE: mini-mental state examination
PRIME-MD: primary care evaluation of mental disorders SCID: structured clinical interview for DSM disorders

Table 10: Descriptive table of included study characteristics

Study	Sample characteristics	Sample size and % depressed	GDS characteristics	Diagnostic standard
Blank et al. (2004)[1]	Country: USA Setting: secondary care, outpatient clinic Age (yrs): Av. = 76.8 Ethnicity: 90.0% white Female: 76.0% Cognition: cognitive impairment excluded	N = 125 Major depression: 11.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Blank et al. (2004)[2]	Country: USA Setting: community, nursing home Age (yrs): Av. = 77.0 Ethnicity: 100% white Female: 67.0% Cognition: cognitive impairment excluded	N = 85 Major depression: 9.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Blank et al. (2004)[3]	Country: USA Setting: secondary care, inpatients Age (yrs): Av. = 80.0 Ethnicity: 93.0% white Female: 51.0% Cognition: cognitive impairment excluded	N = 150 Major depression: 8.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Broekman et al. (2011)	Country: Singapore Setting: community, social service users Age (yrs): Av. = 73.8 Ethnicity: 90.1% Chinese, 9.9% Malays and Indians Female: 59.0% Cognition: cognitive impairment excluded	N = 4253 Major depression: 3.4%	Versions: 15 with 7 subset Administration mode: oral Administrated by: nurses Language: English, Chinese, Malay	DSM-IV SCID
Castello et al. (2010)	Country: Brazil Setting: primary care Age (yrs): 59.5% 60-69, 40.5% 70-79 Ethnicity: not described Female: 72.7% Cognition: not assessed	N = 220 Major depression: 14.0%	Versions: 30 with 15, 10, 4 and 1 subsets Administration mode: oral Administered by: medical students Language: Spanish	DSM-IV SCID
Cullum et al. (2006)	Country: UK Setting: secondary care, inpatients Age (yrs): Av. = 80.2 Ethnicity: not described Female: 59% Cognition: AMTS ≥6 for inclusion	N = 221 Major depression 17.7%	Version: 15 Administration mode: oral Administered by: doctor Language: English	ICD-10 GMS

AMTS: abbreviated mental test score DIS: diagnostic interview schedule DSM: diagnostic and statistical manual of mental disorders
GMS AGECAT: geriatric mental state AGECAT MMSE: mini-mental state examination
PRIME-MD: primary care evaluation of mental disorders SCID: structured clinical interview for DSM disorders

Table 10: Descriptive table of included study characteristics cont.

Study	Sample characteristics	Sample size and % depressed	GDS characteristics	Diagnostic standard
D'ath et al (1994)	Country: UK Setting: primary care Age (yrs): Av. = 74.4 (range 65 - 92) Ethnicity: not described Female: 68.3% Cognition: not assessed	N = 120 Major depression: 34.0%	Versions: 15 with 10, 4 and 1 subsets Administration mode: oral Administered by: doctor Language: English	ICD-10 GMS
Davison et al. (2009)	Country: Australia Setting: community, residential home Age (yrs): Av. = 84.7 (range 67 - 97) Ethnicity: not described Female: 76.8% Cognition: cognitive impairment excluded	N = 168 Major depression: 16.1%	Version: 15 Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
De Craen et al. (2003)	Country: Netherlands Setting: community Age (yrs): Av. = 87.0 (range 86-88) Ethnicity: not described Female: 70.0% Cognition: 20% MMSE 0-18, 42% 19-27, 35% 28-30, 3% unknown	N = 79 Major depression: 10.0%	Version: 15 Administration mode: oral Administered by: interviewer Language: Dutch	ICD GMS AGECAT
Friedman et al. (2005)b	Country: USA Setting: primary care Age (yrs): Av. = 79.3 Ethnicity: 97% white Female: 58.2% Cognition: cognitive impairment excluded	N = 960 Major depression: 12.9%	Version: 15 Administration mode: oral Administered by: interviewer Language: English	DSM-IV MINI
Gerety et al. (1994)	Country: USA Setting: community, nursing home Age (yrs): Av. = 78.9 Ethnicity: 74% white Female: 56.0% Cognition: MMSE >15 for inclusion	N = 134 Major depression: 26.0%,	Versions: 30 with 15 subset Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
Izal et al. (2010)	Country: Spain Setting: mixed (community & day hospital) Age (yrs): Av. = 74.5 Ethnicity: not described Female: 69.0% Cognition: cognitive impairment excluded	N = 233 Major depression: 11.6%	Versions: 30 with 15, 10 and 5 subsets Administration mode: oral Administered by: psychologist Language: Spanish	DSM-IV SCID
Julian et al. (2009)	Country: USA Setting: community COPD patients Age (yrs): Av. = 66.4 Ethnicity: 91.5% white Female: 60.1% Cognition: not assessed	N = 188 Major depression: 11.2%	Version: 15 Administration mode: unclear Administrated by: unclear Language: English	DSM-IV MINI

AMTS: abbreviated mental test score DSM: diagnostic and statistical manual of mental disorders GMS AGECAT: geriatric mental state
ICD: international classification of diseases MINI: mini-international neuropsychiatric interview
MMSE: mini-mental state examination SCID: structured clinical interview for DSM disorders

Table 10: Descriptive table of included study characteristics cont.

Study	Sample characteristics	Sample size and % depressed	GDS characteristics	Diagnostic standard
Lee et al. (2013)	Country: Korea Setting: community Age (yrs): Av. = 72.1 Ethnicity: not described Female: 58.3% Cognition: not assessed	N = 1941 Major depression: 3.2%	Version: 15 Administration mode: oral Administered by: nurses, social workers and medical students Language: Korean	ICD-10 K-CIDI
Licht-Strunk et al. (2005)	Country: Netherlands Setting: primary care Age (yrs): 43.2% 55-64, 30.7% 65-74, 26.1%≥75 Ethnicity: not described Female: 64.5% Cognition: not assessed	N = 948 Major depression: 13.7%	Version: 15 Administration mode: self-administration Administered by: n/a Language: Dutch	DSM PRIME-MD
Lyness et al. (1997)	Country: USA Setting: primary care Age (yrs): Av. = 71.0 Ethnicity: 97.7% white, 2.3% black Female: 58.5% Cognition: not assessed	N = 130 Major depression: 9.2%	Version: 30 with 15 subset Administration mode: self-administration with assistance if required Administered by: n/a Language: English	DSM-III SCID
Malakouti et al. (2006)	Country: Iran Setting: community Age (yrs): 62.7% 59-74, 33.3% 75-83, 3.9% >85 Ethnicity: not described Female: 53.4% Cognition: not assessed	N = 204 Major depression: 10.7%	Version: 15 Administration mode: oral Administered by: psychologist and psychiatrist Language: Farsi	ICD-10 CIDI
Marc et al. (2008)	Country: USA Setting: community, nursing home Age (yrs): Av. = 78.3 Ethnicity: white 85%, black 11%, Hispanic 4% Female: 65.1% Cognition: MMSE ≥18 for inclusion	N = 526 Major depression: 15.4%	Version: 15 Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
McCabe et al. (2006)	Country: Australia Setting: community, nursing home Age (yrs): Av. = 86.6 (range 65-99), 89.4% ≥80 Ethnicity: not described Female: 74.0% Cognition: 54% mildly impaired, 46% moderately impaired	N = 113 Major depression: 17.7%	Version: 15 Administration mode: unclear Administered by: research assistant Language: English	DSM-IV SCID

CIDI: composite international diagnostic interview DSM: diagnostic and statistical manual of mental disorders
ICD: international classification of diseases K-CIDI: Korean composite international diagnostic interview
MINI: mini-international neuropsychiatric interview MMSE: mini-mental state examination
PRIME-MD: primary care evaluation of mental disorders SCID: structured clinical interview for DSM disorders

Table 10: Descriptive table of included study characteristics cont.

Study	Sample characteristics	Sample size and % depressed	GDS characteristics	Diagnostic standard
Neal and Baldwin (1994)	Country: UK Setting: secondary care, outpatient clinic Age (yrs): Av. = 77.2 (range 65-90) Ethnicity: not described Female: 62.0% Cognition: not assessed	N = 45 Major depression: 17.8%	Versions: 30 and 15 subset Administration mode: self-administrated Administrated by: n/a Language: English	ICD GMS AGECAT
Phelan et al. (2010)	Country: USA Setting: primary care Age (yrs): Av. = 78.0 Ethnicity: 32% non-White Female: 62.0% Cognition: not assessed	N = 69 Major depression: 11.5%	Version: 15 Administration mode: self-administration with assistance if required Administrated by: Research assistant if required Language: English	DSM-IV SCID
Rait et al. (1999)	Country: UK Setting: community Age (yrs): Av. = 69.1 Ethnicity: African-Caribbean Female: 50% Cognition: not assessed	N = 130 Major depression: 10.0%	Version: 15 Administration mode: oral Administered by: research interviewers Language: English	ICD GMS AGECAT
Van Marwijk et al. (1995)	Country: Netherlands Setting: primary care Age (yrs): 59.9% 65-74, 40.1% 75-94 Ethnicity: not described Female: 59.5% Cognition: not assessed	N = 586 Major depression: 5.6%	Versions: 30 with 15, 10, 4 and 1 subsets Administration mode: self-administrated Administrated by: n/a Language: Dutch	DSM-IV DIS
Watson et al. (2004)	Country: USA Setting: community, residential home Age (yrs): Av. = 83.0 (range 65-100) Ethnicity: not described Female: 72.0% Cognition: not assessed	N = 112 Major depression: 14.0%	Version: 15 Administration mode: oral Administered by: unclear? Language: English	DSM-IV SCID
Wongpakaran et al. (2013)[1]	Country: Thailand Setting: secondary care, outpatient clinic Age (yrs): Av. = 68.8 Ethnicity: not described Female: 67.9% Cognition: not assessed	N = 156 Major depression: 43.6%	Version: 15 Administration mode: self-administrated Administered by: n/a Language: Thai	DSM-IV MINI
Wongpakaran et al. (2013)[2]	Country: Thailand Setting: community, nursing home Age (yrs): Av. = 76.5 Ethnicity: not described Female: 55.6% Cognition: not assessed	N = 81 Major depression: 28.4%	Version: 15 Administration mode: self-administrated Administered by: n/a Language: Thai	DSM-IV MINI

CIDI: composite international diagnostic interview DIS: diagnostic interview schedule
DSM: diagnostic and statistical manual of mental disorders GMS AGECAT: geriatric mental state AGECAT
ICD: international classification of diseases MINI: mini-international neuropsychiatric interview
SCID: structured clinical interview for DSM disorders

Table 10: Descriptive table of included study characteristics cont.

Overview of studies

Publication year

The range of year of publication for the 32 samples identified ranged from 1994 to 2013.

Language

All of the 31 records were published in English with the exception of two; Allgaier et al. 2011, which was in German, and Lee et al. 2013, which was in Korean.

Country

Twelve of the 32 samples were set in European countries; six were from the UK (Abas et al., 1998, Arthur et al., 1999, Cullum et al., 2006, Dath et al., 1994, Neal and Baldwin, 1994, Rait et al., 1999), four from the Netherlands (Bijl et al., 2006, de Craen et al., 2003, Licht-Strunk et al., 2005, Van Marwijk et al., 1995), one from Germany (Allgaier et al., 2011, Allgaier et al., 2013) and one from Spain (Izal et al., 2010). Ten samples were from the USA (Blank et al., 2004 $^{1-3}$, Friedman et al., 2005b, Gerety et al., 1994, Julian et al., 2009, Lyness et al., 1997, Marc et al., 2008, Phelan et al., 2010, Watson et al., 2004). Two samples were from Australia (Davison et al., 2009, McCabe et al., 2006). Two samples were from Brazil (Almeida and Almeida, 1999, Castello et al., 2010). In total, 26 studies were based in Western countries. Six samples were based in non-Western countries: Iran (Malakouti et al., 2006), Singapore (Broekman et al., 2011, Nyunt et al., 2009a), South Korea (Bae and Cho, 2004, Lee, 2013) and Thailand (Wongpakaran et al., 2013 $^{1-2}$).

Setting

Ten of the studies were based in a primary care setting, seven in secondary care, fourteen in the community (eight of which in either a nursing or residential home) and one in a mixed setting (i.e. a combination of community and day hospital).

Age

The mean age of the samples ranged from 66.4 to 87.0 years. However, four studies did not report a mean age and instead reported sample age by proportion; for the study by Castello et al. 59.5% of the sample were 60-69 years of age and 40.5% were 70-79 years of age. Licht-strunk et al. reported 43.2% of the sample being 55-64 years of age, 30.7% aged 65-74 years of age and 26.1% ≥75 years of age. Malakouti et al. reported 62.7% of the sample being 59-74 years of age, 33.3% aged 75-84 years of age and 3.9% aged >85 years. The study by Van Marwijk et al. reported 59.9% of the sample being 65-74 years of age and 40.1% being 75-94 years of age. The studies with the highest mean age were based in a community setting of either a residential or nursing home: Watson et al. 83.0 years of age, Allgaier et al. 84.5 years of age, Davison et al. 84.7 years of age and McCabe et al. 86.6 years of age.

Ethnicity

20 studies did not report the ethnicity of the sample. In 9 studies, all based in the USA, the majority of the sample were white Caucasians; the proportion of sample that were white ranged from 68% to 100%. Three studies had samples where none of the participants was of white Caucasian ethnicity; two of the studies (Abas et al., 1998, Rait et al., 1999), both based in the UK, had a sample comprised of African-Caribbean participants only. In one study (Broekman et al., 2011) the study sample was comprised of 90.1% Chinese and 9.9% Malay and Indian.

Female

The proportion of the sample female in all studies ranged from 50.0% to 84.4%. In thirteen studies the proportion of the sample female is <60%. In twelve studies the proportion of the sample female is 60-70%. In seven studies the proportion of sample female is >70%.

Cognitive status

12 studies did not assess cognitive function. Nine studies excluded anyone with cognitive impairment. Six studies specified inclusion criteria as scoring above a certain cut-off score on a cognitive test; >15 on the mini-mental state examination (MMSE) (Folstein et al., 1975), (3 studies) >18 on the MMSE (2 studies), and >6 on the abbreviated mental test (AMT) (Hodkinson, 1972) (1 study). The five remaining studies measured cognitive function but did not specify this as an inclusion or exclusion criteria.

Sample size

Sample size in the 32 studies ranged from 45 to 4253 participants. Seven studies had a sample size <100, thirteen studies had a sample size of 100-200 and five studies had a sample size of 200-300. Six studies had sample sizes that ranged from 300 to 1000. The highest sample sizes were found in studies that had a community setting; Broekman et al. had a sample size of 4235 and Lee et al. had a sample size of 1941.

Prevalence of major depression

Prevalence of major depression ranged from 3.2% to 64.1%. Mean prevalence of depression was 17.7%. Nine studies reported prevalence <10%, 15 studies reported prevalence as 10-20% and seven studies reported prevalence as >20%. Prevalence of major depression was highest in the studies by Almeida and Almeida, Bae and Cho, and Wongpakaran et al. at 64.1%, 40.1% and 43.6% respectively; the studies by Almeida and Almeida, and Bae and Cho were based in mental health outpatient clinics in secondary care settings. The study by Wongpakaran et al. was also based in outpatient clinic in a secondary care setting but the type of clinic was not specified.

Briefer versions of the GDS examined

Seven briefer versions of the GDS were examined in the studies; GDS-1, GDS-4, GDS-5, GDS-7, GDS-8, GDS-10 and the GDS-15. See Table 11 below. Five studies reported data regarding the GDS-1. Five studies reported data regarding the GDS-4. The GDS-5, GDS-7 and GDS-8 only had one study providing data for each; none of these three studies was the same however. Five studies reported data for the GDS-10. 29 studies reported data regarding the GDS-15. With the exception of the GDS-1, different studies reported data at different cut-off scores.

The recommended cut-off score for the GDS-15 is 5 (Yesavage et al., 1983); 23 out of the 32 studies for the GDS-15 reported data at a cut-off score of 5.

GDS version		GDS-1	GDS-4		GDS-5	GDS-7	GDS-8	GDS-10				GDS-15												
Study	Cut-off score	n/a	1	2	2	2	5	2	3	4	5	1	2	3	4	5	6	7	8	9	10	11	12	13
Abas et al., 1998															x	x	x							
Allgaier et al., 2013		x	x	x			x										x	x	x					
Almeida and Almeida				x						x	x					x	x	x						
Arthur et al., 1999													x	x	x	x	x	x						
Bae and Cho., 2004																x	x	x	x	x	x	x		
Bijl et al., 2006																x								
Blank et al., 2004[1]																	x			x				
Blank et al., 2004[2]																	x	x						
Blank et al., 2004[3]																	x							
Broekman et al., 2011						x								x	x	x	x	x	x					
Castello et al., 2010		x	x	x						x	x				x	x	x							
Cullum et al., 2006																x		x	x	x	x			
D'ath et al., 1994		x	x	x					x	x						x	x							
Davison et al., 2009																x	x							
De Craen et al., 2003														x	x	x	x							
Friedman et al.,2005b																x	x	x						
Gerety et al., 1994																	x							
Izal et al., 2010					x				x							x								
Julian et al., 2009																x								
Lee et al., 2003															x	x	x	x	x	x	x			
Licht-Strunk etal.,2005																x								
Lyness et al., 1997																x								
Malakouti et al., 2006													x	x	x	x	x	x	x	x	x	x	x	x
Marc et al., 2008												x	x	x	x	x	x	x	x	x	x	x	x	x

Table 11: Table showing studies of brief versions of the GDS and cut-off scores reported

GDS version		GDS-1	GDS-4		GDS-5	GDS-7	GDS-8	GDS-10				GDS-15												
Study	Cut-off score	n/a	1	2	2	2	5	2	3	4	5	1	2	3	4	5	6	7	8	9	10	11	12	13
McCabe et al., 2006																	x							
Neal and Baldwin, 1994																x								
Phelan et al., 2010																x	x	x	x	x				
Rait et al., 1999															x									
Van Marwijk et al., 1995		x	x	x				x	x				x	x										
Watson et al., 2004																x								
Wongpakaran et al.,2013[1]															x	x	x							
Wongpakaran et al.,2013[2]																			x	x	x			

Table 11: Table showing studies of brief versions of the GDS and cut-off scores reported cont.

Administration mode of the GDS

The GDS was orally administered to participants in 21 studies; this involved either interviewers, members of the research team, doctors, nurses, psychologists or medical students reading out the question items of the GDS to the study participants. In nine studies, the GDS was self-administered; in four of these studies assistance was available if required. Administration mode was unclear in two studies (McCabe et al., 2006 and Julian et al., 2009).

Language of the GDS

Out of the 32 primary studies, 17 studies used the English version of the GDS. The remaining samples used translated versions, of which Dutch was the most common language. Four studies translated the GDS versions into Dutch (Bijl et al., 2006, de Craen et al., 2003, Licht-Strunk et al., 2005, Van Marwijk et al., 1995). Remaining languages of translated versions included Spanish (Castello et al., 2010 and Izal et al., 2010), Portuguese (Almeida and Almeida, 1999), German (Allgaier et al., 2011, Allgaier et al., 2013), Korean (Bae and Cho, 2004, Lee et al., 2013), Thai (Wongpakaran et al. 2013$^{1-2}$), and Farsi (Malakouti et al., 2006). One sample used a mixture of different language versions (Broekman et al., 2011, Nyunt et al., 2009).

Gold-standard diagnostic tests utilised

DSM and ICD diagnoses of major depression were included in all 32 studies. The majority of primary studies (22 studies) utilized a gold-standard reference based on DSM diagnostic criteria. In the remaining 10 studies, the gold-standard reference test was based on ICD diagnostic criteria.

Five different DSM gold-standard reference tests were identified; the most common was the Structured Clinical Interview for DSM disorders (SCID), with it being used in 11 studies. Five studies used the Diagnostic Interview Schedule (DIS), four studies used the Mini-International Neuropsychiatric Interview (MINI), and two studies used the Primary Care Evaluation of Mental Disorder (PRIME-MD).

Five different ICD gold-standard reference tests were identified: four studies used the Geriatric Mental State – Automated Geriatric Examination for Computer Assisted

Taxonomy (GMS-AGECAT), two studies used the Geriatric Mental State Schedule (GMS), two studies used the Composite International Diagnostic Interview (CIDI), one study used the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) and one study used the ICD-10 checklist of symptoms.

Quality assessment

Primary study quality was measured by the QUADAS-II as previously discussed. Table 12 and Table 13 shows quality assessment results. The overall rating of risk of bias for each study concerning each domain of the QUADAS-II varied.

Thirty-two independent samples have been identified from thirty-one citations. Two papers describe the methodology and results for the sample by Allgaier et al., therefore both records have been quality assessed as one and so only appear on the QUADAS-II results table once. The same applies to the papers by Broekman et al. and Nyunt et al., which describe the same sample. The QUADAS-II results appear under Broekman et al., 2011. As discussed, the paper by Blank et al. has been treated as three separate samples but for quality analysis it appears only once on the QUADAS-II. The same applies to the two samples by Wongpakaran et al., 2013.

For the domain of participant selection, out of the identified twenty-nine primary studies, twenty-five were rated as having an overall ‘low’ rating of bias. Two studies were rated as having an overall ‘high’ rating of bias; Almeida et al. and Bae et al. Both of these studies were based in secondary care mental health settings. The prevalence of depression in both studies was high; 64.1% and 40.1% respectively. The overall rating of bias was ‘unclear’ for two studies (Castello et al., 2010 and Lee et al., 2013).

For the domain of index test, overall rating of bias was more varied: twelve studies had a ‘low’ rating of bias, eight studies had a ‘high’ rating of bias and nine studies had an ‘unclear’ rating of bias. The study by Arthur et al. had an overall ‘high’ rating of bias because the brief version of the GDS was not interpreted blind to the gold-standard diagnostic reference test. The remaining seven studies (Castello et al., 2010, Friedman et al., 2005b, Gerety et al., 1994, Izal et al., 2010, Lyness et al., 1997, Neal and Baldwin, 1994, Rait et al., 1999) rated as having an overall ‘high’ rate of bias did not pre-specify cut-off score or report data for multiple cut-off scores. Lyness et al. did not pre-specify cut-off score and the brief version of the GDS was not interpreted blind. It was also unclear in the study by Friedman et al. if the brief version of the GDS was interpreted blind. For Castello et al., as well as cut-off score not being pre-specified, the psychometric properties were not reported despite the brief version of the GDS being translated into Spanish.

For studies rated as having an overall 'unclear' rate of bias for the domain of index test the reasons include it being unclear in five studies (Almeida and Almeida, 1999, Cullum et al., 2006, D'ath et al., 1994, Lee et al., 2013, Licht-Strunk et al., 2005) if the brief version of the GDS was interpreted blind to the gold-standard diagnostic reference test. In two studies (Van Marwijk et al., 1995, Watson et al., 2004) it is unclear if cut-off scores were pre-specified. In the remaining two samples (Allgaier et al., 2013, De Craen et al., 2003) it is unclear if translation was appropriate.

For the domain of reference test, only the study by Malakouti et al. was rated as having an overall 'high' rate of bias because it did not report the psychometric properties of the translated gold-standard, diagnostic reference test. Nineteen studies were rated as having an overall 'low' rate of bias. The overall rate of bias was 'unclear' in nine studies; Allgaier et al., 2013, Almeida and Almeida, 1999, Arthur et al., 1999, Cullum et al., 2006, D'ath et al., 1994, De Craen et al., 2003, Friedman et al., 2005b, Lee et al., 2013 and Licht-Strunk et al., 2005. It was unclear if translation of the gold-standard diagnostic reference was appropriate for the studies by Allgaier et al. and De Craen et al. It was unclear if the reference test was interpreted blind to the brief version of the GDS in the remaining studies. It was unclear for the study by Licht-Strunk et al. if both interpretation was blind and if translation was appropriate.

For the domain of flow/timing of study design, sixteen studies were rated as having an overall 'low' rate of bias. Seven studies were rated as having an overall 'high' rate of bias; for the studies by Abas et al., 1998, Cullum et al., 2006, D'ath et al., 1994, Licht-Strunk et al., 2005, and Malakouti et al., 2006 all participants did not receive the same test and all participants were not included in analysis. It was also unclear in the studies by Cullum et al., 2006 and D'ath et al., 1994 if the interval between administration of the brief version of the GDS and gold-standard reference test was less than two weeks. The interval between administration of the brief version of the GDS and gold-standard reference test was greater than two weeks for the study by Almeida and Almeida, 1999, and Arthur et al., 1999.

The six studies by Castello et al., 2006, Friedman et al., 2005b, Izal et al., 2010, Lee et al., 2013, Van Marwijk et al., 1995, and Wongpakaran et al., 2013, were rated as having an overall 'unclear' rate of bias for flow/timing of study because the interval between administration of the brief version of the GDS and gold-standard reference test was unclear.

Nine out of the twenty-nine primary studies were thought to have a participant selection that was not applicable to the target population of this review. The studies by Almeida et al. and Bae et al. were deemed as not having an applicable participant selection because they were based in secondary care, mental health settings, which led to high prevalence rates of major depression. The remaining seven studies by Abas et al., 1998, Broekman et al., 2011, Lee et al., 2013, Malakouti et al., 2006, McCabe et al., 2006, Rait et al., 1999, and Wongpakaran et al. 2013, were deemed not to have an applicable participant selection because the countries of setting and ethnicities are not comparable to the population of this review.

The applicability of the index test and reference test for all twenty-nine primary studies were rated positively.

All twenty-nine primary studies reported diagnostic data for the GDS-15; in eight samples (Blank et al., 2004$^{1-3}$, Bae and Cho, 2004, Castello et al., 2010, Gerety et al., 1994, Izal et al., 2010, Lyness et al., 1997, Neal and Baldwin, 1994, Van Marwijk et al., 1995) data regarding the GDS-15 was extracted from the original, 30-item GDS. The GDS-15 was administered directly to participants in the remaining twenty-one studies. All briefer versions of the GDS were extracted from either the original 30-item GDS (Castello et al., 2010, Izal et al., 2010, Van Marwijk et al., 1995) or the GDS-15 (Allgaier et al., 2013, Almeida and Almeida, 1999, Broekman et al., 2011, D'ath et al., 1994).

Study	Participant selection:				Index test:					Reference test:					Flow/timing:			
	Consecutive or random sample	Avoid case-control/ avoid artificially inflated base	Avoided inappropriate exclusions	Overall risk of bias	GDS interpreted blind to reference test	Threshold pre-specified or multiple cut-offs reported	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Reference test correctly classifies target condition	Reference test interpreted blind to GDS	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Interval of two weeks or less	All participants receive same reference test	All participants included in analysis?	Overall risk of bias
Abas et al. (1998)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✗	✗	High
Allgaier et al. (2013)	✓	✓	✓	Low	✓	✓	?	?	Unclear	✓	✓	?	✓	Unclear	✓	✓	✓	Low
Almeida and Almeida (1999)	?	✗	✓	High	?	✓	✓	✓	Unclear	✓	?	✓	✓	Unclear	✗	✓	✓	High
Arthur et al. (1999)	✓	✓	✓	Low	✗	✓	n/a	n/a	High	✓	?	n/a	n/a	Unclear	✗	✓	✓	High
Bae and Cho (2004)	✓	✗	✓	High	✓	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	Low
Bijl et al. (2006)	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	Low
Blank et al. (2004)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Broekman et al. (2011)	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	Low
Castello et al. (2010)	✓	✓	✗	Unclear	✓	✗	✓	✗	High	✓	✓	✓	✓	Low	?	✓	✓	Unclear
Cullum et al. (2006)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear	✓	?	n/a	n/a	Unclear	?	✗	✗	High
D'Ath et al. (1994)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear	✓	?	n/a	n/a	Unclear	?	✗	✗	High

Table 12: QUADAS-II A results

	Participant selection:				Index test:					Reference test:					Flow/timing:			
Study	Consecutive or random sample	Avoid case-control/ avoid artificially inflated base	Avoided inappropriate exclusions	Overall risk of bias	GDS interpreted blind to reference test	Threshold pre-specified or multiple cut-offs reported	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Reference test correctly classifies target condition	Reference test interpreted blind to GDS	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Interval of two weeks or less	All participants receive same reference test	All participants included in analysis?	Overall risk of bias
Davidson et al. (2009)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
D'Ath et al. (1994)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear	✓	?	n/a	n/a	Unclear	?	×	×	High
Davidson et al. (2009)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
De Craen et al. (2003)	✓	✓	✓	Low	✓	✓	?	?	Unclear	✓	✓	?	?	Unclear	✓	✓	✓	Low
Friedman et al. (2005)b	✓	✓	✓	Low	?	×	n/a	n/a	High	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
Gerety et al. (1994)	✓	✓	✓	Low	✓	×	n/a	n/a	High	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Izal et al. (2010)	✓	✓	✓	Low	✓	×	✓	✓	High	✓	✓	✓	✓	Low	?	✓	✓	Unclear
Julian et al. (2009)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Lee et al. (2013)	✓	?	✓	Unclear	?	✓	✓	✓	Unclear	✓	?	✓	✓	Unclear	?	✓	✓	Unclear
Licht-Strunk et al. (2005)	✓	✓	✓	Low	?	✓	✓	✓	Unclear	✓	?	?	?	Unclear	?	×	×	High

Table 12: QUADAS-II results A cont.

	Participant selection:				Index test:					Reference test:					Flow/timing:			
Study	Consecutive or random sample	Avoid case-control/ avoid artificially inflated base	Avoided inappropriate exclusions	Overall risk of bias	GDS interpreted blind to reference test	Threshold pre-specified or multiple cut-offs reported	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Reference test correctly classifies target condition	Reference test interpreted blind to GDS	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Interval of two weeks or less	All participants receive same reference test	All participants included in analysis?	Overall risk of bias
Lyness et al. (1997)	✓	✓	✓	Low	×	×	n/a	n/a	High	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Malakouti et al. (2006)	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	×	High	✓	×	×	High
Marc et al. (2008)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
McCabe et al. (2006)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Neal and Baldwin (1994)	✓	✓	✓	Low	✓	×	n/a	n/a	High	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Phelan et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Rait et al. (1999)	✓	✓	✓	Low	✓	×	n/a	n/a	High	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Van Marwijk et al. (1995)	✓	✓	✓	Low	✓	?	✓	✓	Unclear	✓	✓	✓	✓	Low	?	✓	✓	Unclear
Watson et al. (2004)	✓	✓	✓	Low	✓	?	n/a	n/a	Unclear	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Wongpakaran et al. (2013)	✓	✓	✓	Low	✓	✓	✓	✓	Low	✓	✓	✓	✓	Low	?	✓	✓	Unclear

Table 12: QUADAS-II results A cont.

Study	**Patient selection:** Applicability	**Index test:** Applicability	**Reference test:** Applicability
Abas et al. (1998)	✗	✓	✓
Allgaier et al. (2013)	✓	✓	✓
Almeida and Almeida (1999)	✗	✓	✓
Arthur et al. (1999)	✓	✓	✓
Bae and Cho (2004)	✗	✓	✓
Bijl et al. (2005)	✓	✓	✓
Blank et al. (2004)	✓	✓	✓
Broekman et al. (2011)	✓	✓	✓
Castello et al. (2010)	✓	✓	✓
Cullum et al. (2006)	✓	✓	✓
D'Ath et al. (1994)	✓	✓	✓
Davidson et al. (2009)	✓	✓	✓
De Craen et al. (2003)	✓	✓	✓
Friedman et al. (2005)b	✓	✓	✓
Gerety et al. (1994)	✓	✓	✓
Izal et al. (2010)	✓	✓	✓
Julian et al. (2009)	✓	✓	✓
Lee et al. (2013)	✗	✓	✓
Licht-Strunk et al. (2005)	✓	✓	✓
Lyness et al. (1997)	✓	✓	✓
Malakouti et al. (2006)	✗	✓	✓
Marc et al. (2008)	✓	✓	✓
McCabe et al. (2006)	✗	✓	✓
Neal and Baldwin (1994)	✓	✓	✓
Phelan et al. (2010)	✓	✓	✓
Rait et al. (1999)	✗	✓	✓
Van Marwijck et al. (1995)	✓	✓	✓
Watson et al. (2009)	✓	✓	✓
Wonkpakaran et al. (2013)	✗	✓	✓

Table 13: QUADAS-II results B

Narrative analysis

The same studies reported diagnostic data pertaining to brief GDS versions of 10 or fewer items. For example, the study by Van Marwijk et al. reported data for the GDS-1, GDS-4 and GDS-10.

It was only possible to perform a diagnostic meta-analysis for the GDS-15 and not other brief versions of the GDS. At least four studies are required to conduct a diagnostic meta-analysis using Stata. Other brief versions of the GDS (i.e. ten or fewer items) were not comprised of standardised items so while a measure in one study may have shared the same name with that used in another study they were essentially different measures (See Table 14); this is why it was not possible to undertake a diagnostic meta-analysis for the GDS-1, GDS-4 and GDS-10 despite there being more than four studies for each. For the GDS-5, GDS-7 and GDS-8 it was not possible to perform a diagnostic meta-analysis because there were too few studies. See Table 15.

***GDS-1*:** Four studies (Almeida and Almeida, 1999, Castello et al., 2010, Van Marwijk et al., 1995, D'ath et al., 1994) reported diagnostic data concerning the GDS-1. Not all studies used the same item to comprise the GDS-1 and so two versions of the GDS-1 have been identified: Almeida and Almeida, Castello et al., and Van Marwijk et al. used the item '*Are you basically satisfied with your life?*'; whereas D'ath et al. utilised the item '*Do you feel that your life is empty?*'

For the GDS-1 comprised of the item 'Are you basically satisfied with your life?' sensitivity ranged from 0.18 to 0.62. Almeida et al. and Castello et al. reported sensitivities of 0.62 (95% CI 0.45-0.76) and 0.48 (95% CI 0.45-0.67) respectively; 95% CI overlapped. See Table 15. Van Marwijk et al. reported a sensitivity of 0.18 (95% CI 0.07-0.36); there was some overlap in the 95% CI with Castello et al. but not Almeida and Almeida. Reported specificities for the three studies showed less variance; specificity ranged from 0.91 to 0.96. All 95% confidence intervals overlapped. The study by Almeida et al. had the widest 95% confidence interval of the three studies: 0.72-0.99. The remaining two studies showed narrower confidence intervals; Castello et al. 0.93-0.99 and Van Marwijk et al. 0.90-0.94).

For the study by D'ath et al. sensitivity was 0.59 (95% CI 0.42-0.74), which is comparable with the reported sensitivities of the other GDS-1 version. Specificity for D'ath et al. was 0.75 (95% CI 0.64-0.84), which is much lower than that reported in the other version.

Item number	GDS-15 item	Allgaier et al. (2013)	Almeida & Almeida (1999)	Broekman et al. (2011)	Castello et al. (2010)	D'ath et al. (1994)	Izal et al. (2010)	Van Marwijk et al. (1995)
1	Are you basically satisfied with your life?	GDS-4 GDS-8	GDS-1 GDS-4 GDS-10	GDS-7	GDS-1 GDS-4 GDS-10	GDS-4 GDS-10	GDS-5 GDS-10	GDS-1 GDS-4 GDS-10
2	Have you dropped many of your activities and interests?		GDS-4 GDS-10		GDS-4 GDS-10	GDS-10	GDS-10	GDS-4 GDS-10
3	Do you feel that your life is empty?	GDS-4 GDS-8		GDS-7		GDS-1 GDS-4 GDS-10	GDS-10	
4	Do you often get bored?	GDS-8	GDS-10	GDS-7	GDS-10		GDS-5	GDS-10
5	Are you in good spirits most of the time?	GDS-8	GDS-10	GDS-7	GDS-10			GDS-10
6	Are you afraid that something bad is going to happen to you?	GDS-4				GDS-4 GDS-10	GDS-10	
7	Do you feel happy most of the time?	GDS-4 GDS-8	GDS-4 GDS-10	GDS-7	GDS-4 GDS-10	GDS-4 GDS-10	GDS-10	GDS-4 GDS-10
8	Do you feel helpless?	GDS-8	GDS-10	GDS-7	GDS-10	GDS-10	GDS-5 GDS-10	GDS-10
9	Do you prefer to stay at home, rather than going out and doing new things?		GDS-4 GDS-10		GDS-4		GDS-5	GDS-4 GDS-10
10	Do you feel you have more problems with memory than most?				GDS-10	GDS-10	GDS-10	
11	Do you think it is wonderful to be alive?	GDS-8						
12	Do you feel pretty worthless the way you are now?		GDS-10		GDS-10		GDS-5	GDS-10
13	Do you feel full of energy?		GDS-10		GDS-10	GDS-10	GDS-10	GDS-10
14	Do you feel that your situation is hopeless?	GDS-8				GDS-10	GDS-10	
15	Do you think that most people are better off that you are?		GDS-10	GDS-7	GDS-10	GDS-10	GDS-10	GDS-10

Table 14: Items comprising brief versions of the GDS

***GDS-4*:** Five studies reported diagnostic data for the GDS-4. Again, like the GDS-1, two different versions of this brief version of the GDS have been identified. Studies have reported diagnostic data for the cut-off scores of 1 and 2. The GDS-4 is comprised of the same items for the studies by Allgaier et al. and D'ath et al. At a cut-off score of 1, sensitivity is 0.85 (95% CI 0.55-0.98) and 0.93 (95% CI 0.80-0.98) respectively. For the studies by Castello et al. and Van Marwijk et al., where the same items comprise the GDS-4, sensitivities show more variance; 0.84 (95% CI 0.66-0.95) and 0.61 (95% CI 0.42-0.77) respectively. There is overlap in confidence intervals with the exception of D'ath et al. and Van Marwijk et al., which do not overlap with each other. See Table 15.

In terms of specificity, there is more difference between the two versions of the GDS-4 at a cut-off score of 1; reported specificities for Allgaier et al. and D'ath et al. are 0.53 (95% CI 0.42-0.65) and 0.63 (95% CI 0.52-0.74) respectively. Reported specificities for Castello et al. and Van Marwijk et al. are more favourable than the other studies; 0.75 (95% CI 0.68-0.81) and 0.72 (95% CI 0.68-0.76) respectively. The reported specificities for Castello et al. and Van Marwijk et al. are more similar and show smaller confidence intervals than the other studies. Overall, with the exception of Allgaier et al., there is good overlap in the confidence intervals between all studies. See Table 15.

At a cut-off score of 2, there is no preference regarding reported diagnostic data for either version of the GDS-4. The studies by Allgaier et al. and D'ath et al. report sensitivities of 0.54 (95% CI 0.25-0.81) and 0.61 (0.65-0.91) respectively. Reported specificities are closer together and have narrower confidence intervals: 0.92 (95% 0.84-0.97) and 0.92 (95% CI 0.80-0.95) respectively. See Table 15.

Version	Cut-off score	Utilise same items	Study	Sensitivity (95% CI)	Specificity (95% CI)
1	n/a	1	Almeida & Almeida (1999)	0.62 (0.45-0.76)	0.91 (0.72-0.99)
			Castello et al. (2010)	0.48 (0.30-0.67)	0.96 (0.93-0.99)
			Van Marwijk et al. (1995)	0.18 (0.07-0.36)	0.92 (0.90-0.94)
		3	D'ath et al. (1994)	0.59 (0.42-0.74)	0.75 (0.64-0.84)
4	1	1, 3, 6 and 7	Allgaier et al. (2013)	0.85 (0.55-0.98)	0.53 (0.42-0.65)
			D'ath et al. (1994)	0.93 (0.80-0.98)	0.63 (0.52-0.74)
		1, 2, 7 and 9	Castello et al. (2010)	0.84 (0.66-0.95)	0.75 (0.68-0.81)
			Van Marwijk et al.(1995)	0.61 (0.42-0.77)	0.72 (0.68-0.76)
	2	1, 3, 6 and 7	Allgaier et al. (2013)	0.54 (0.25-0.81)	0.92 (0.84-0.97)
			D'ath et al. (1994)	0.61 (0.45-0.76)	0.89 (0.80-0.95)
		1, 2, 7 and 9	Almeida and Almeida (1999)	0.81 (0.65-0.91)	0.78 (0.56-0.93)
			Castello et al. (2010)	0.54 (0.36-0.73)	0.94 (0.90-0.97)
			Van Marwijk et al. (1995)	0.67 (0.48-0.82)	0.66 (0.62-0.70)
5	2	1, 4, 8, 9, and 12	Izal et al. (2010)	0.67 (0.46-0.84)	0.78 (0.72-0.84)
7	2	1, 3, 4, 5, 7, 8 and 15	Broekman et al. (2011)	0.93 (0.88-0.97)	0.91 (0.90-0.92)
8	5	1, 3, 4, 5, 7, 8, 11 and 14	Allgaier et al. 2013)	0.77 (0.46-0.95)	0.89 (0.80-0.95)
10	2	1, 2, 4, 5, 7, 8, 9, 12, 13 and 15	Van Marwijk et al. (1995)	0.67 (0.48-0.82)	0.66 (0.62-0.70)
	3	1, 2, 3, 6, 7, 8, 10, 13, 14 and 15	D'ath et al. (1994)	0.93 (0.80-0.98)	0.63 (0.52-0.74)
			Izal et al. (2010)	1.00 (0.88-1.00)	0.82 (0.76-0.86)
		1, 2, 4, 5, 7, 8, 9, 12, 13 and 15	Almeida and Almeida (1999)	0.92 (0.64.-0.99)	0.65 (0.53-0.75)
			Castello et al. (2010)	0.77 (0.59-0.90)	0.81 (0.75-0.86)
			Van Marwijk et al. (1995)	0.52 (0.34-0.69)	0.83 (0.80-0.86)
	4	1, 2, 4, 5, 7, 8, 9, 12, 13 and 15	Almeida and Almeida (1999)	0.85 (0.55-0.98)	0.79 (0.68-0.87)
			Castello et al. (2010)	0.65 (0.45-0.81)	0.89 (0.84-0.93)

Table 15: Diagnostic data for brief versions of the GDS

The study by Almeida and Almeida reported diagnostic data for the GDS-4, comprised of the same items as the version used by Castello et al. and Van Marwijk et al., at a cut-off score of 2. Reported sensitivities for this version of the GDS-4 are greater than the reported sensitivities for the Allgaier et al. and D'ath et al. version. Reported sensitivity for Almeida and Almeida was 0.81 (95% CI 0.65-0.91), for Castello et al. 0.54 (95% CI 0.36-0.73) and for Van Marwijk et al. 0.67 (95% CI 0.48-0.82); all confidence intervals overlapped. Reported specificity for Almeida and Almeida was 0.78 (95% CI 0.56-0.93), for Castello et al. 0.94 (95% CI 0.90-0.97) and for Van Marwijk et al. 0.67 (95% CI 0.62-0.70); again all confidence intervals overlapped. See Table 15.

***GDS-5*:** Only one study, by Izal et al., reported diagnostic data concerning the GDS-5. Reported sensitivity was 0.67 (95% CI 0.48-0.84) and specificity was 0.78 (95% CI 0.72-0.84).

***GDS-7*:** The study by Broekman et al. was the only study that reported diagnostic data concerning the GDS-7. Reported sensitivity was 0.93 (95% CI 0.88-0.97) and specificity was 0.91 (95% CI 0.90-0.92).

***GDS-8*:** Allgaier et al. was the only study that reported diagnostic data concerning the GDS-8. Reported sensitivity was 0.77 (95% CI 0.46-0.95) and specificity was 0.89 (95% CI 0.80-0.95).

***GDS-10*:** Two versions of the GDS-10 were identified – see Table 14. The studies by Almeida and Almeida, Castello et al. and Van Marwijk et al. used the same items; all three studies reported diagnostic data for a cut-off score of 3, but only Van Marwijk et al. reported diagnostic data for a cut-off score of 2. Van Marwijk et al. did not report data at a cut-off score of 4 unlike the other two studies.

At a cut-off score of 2, Van Marwijk et al. reported a sensitivity of 0.67 (95% CI 0.48-0.82) and a specificity of 0.66 (95% CI 0.62-0.70).

The reported sensitivities of the three studies at a cut-off score of 3 varied greatly (0.52 to 0.92); however, all 95% confidence intervals did overlap. Van Marwijk et al. reported the lowest sensitivity at 0.52 (95% CI 0.34-0.69). Almeida and Almeida reported the highest sensitivity at 0.92 (95% CI 0.64–0.99). The sensitivity reported by Castello et al. was 0.77 (95% CI 0.59-0.90).

The range of reported specificities for the three studies was narrower (0.65 to 0.83). Reported specificities of Castello et al. and Van Marwijk et al. were very similar; 0.81 (95% 0.75-0.86) and 0.83 (95% CI 0.80-0.86) respectively. Almeida and Almeida reported a much lower specificity of 0.65 (95% CI 0.53-0.75).

At a cut-off score of 4, Almeida and Almeida, again, reported a higher sensitivity and lower specificity than Castello et al. There was, however, overlap in the reported 95% confidence intervals. Reported sensitivities were 0.85 (95% CI 0.55-0.98) and 0.65 (95% CI 0.45-0.81) for Almeida and Almeida and Castello et al. respectively. Reported specificities were 0.79 (95% 0.68-0.87) and 0.89 (95% CI 0.84-0.93) respectively.

The studies by D'ath et al. and Izal et al. reported diagnostic data for another version of the GDS-10; only diagnostic data at a cut-off score of 3 were reported. Reported sensitivities for D'ath et al. and Izal et al. were 0.93 (95% CI 0.80-0.98) and 1.00 (95% CI 0.88-1.00) respectively. These results are more favourable than the sensitivity data reported by the other version of the GDS-10 used by the three studies discussed above. Specificity data reported by D'ath et al. and Izal. et al. is, however, similar to reported specificities of this other version. Reported specificity is 0.63 (95% CI 0.52-0.74) for D'ath et al. and 0.82 (95% CI 0.76-0.86) for Izal et al.

***GDS-15*:** 32 samples reported diagnostic data for the GDS-15. Unlike other brief versions of the GDS, the items comprising the GDS-15 are standardised. 22 samples reported diagnostic data at multiple cut-off scores. Ten samples only reported diagnostic data at a single cut-off score. The recommended cut-off score of the GDS-15 is 5 (Yesavage, 1986). Of the ten samples reporting diagnostic data at a single cut-off score, three did not reported diagnostic data at a cut-off score of 5 (Gerety et al., 1994, McCabe et al., 2006, Rait et al., 1999). Of the 22 samples reporting diagnostic data at multiple cut-off scores, five did not report diagnostic data at cut-off score of 5 (Allgaier et al., 2013, Blank et al., 2004 [1-3], Van Marwijk et al., 1995).

When data pertaining to multiple cut-off scores were report, consecutive cut-off scores were reported (e.g. diagnostic data were reported at a cut-score of 4, 5, 6, 7, etc.). Two of the 22 studies, however, did not report diagnostic data at consecutive cut-off scores; for example, Cullum et al. reported diagnostic data at a cut-off score of 5, 7, 8, 9 and 10 but not 6. Van Marwijk et al. reported diagnostic data at cut-off score of 4, 5, 6, 8, 9 and 10 but not 7.

Diagnostic data were found for a cut-off score of 1-13. Only three samples reported diagnostic data at a cut-off score of 1, 11, 12 and 13 and therefore meta-analysis was not possible; a minimum of four samples is required to perform meta-analysis. Table 16 shows diagnostic data for the GDS-15 at cut-off scores of 1, 11, 12 and 13.

Cut-off score	Study	Sensitivity (95% CI)	Specificity (95% CI)
1	Marc et al., 2008	0.95 (0.75-0.99)	0.16 (0.09-0.24)
11	Bae and Cho, 2004	0.65 (0.51-0.76)	0.88 (0.80-0.94)
	Malakouti et al., 2006	0.55 (0.32-0.76)	0.88 (0.80-0.91)
	Marc et al., 2008	0.15 (0.03-0.38)	0.98 (0.93-1.00)
12	Malakouti et al., 2006	0.36 (0.17-0.59)	0.89 (0.84-0.93)
	Marc et al., 2008	0.10 (0.12-0.32)	0.99 (0.94-1.00)
13	Malakouti et al., 2006	0.27 (0.11-0.50)	0.91 (0.86-0.95)
	Marc et al., 2008	0.00 (0.0-0.17)	100.0 (0.96-1.00)

Table 16: Diagnostic data for the GDS-15 at cut-off scores where meta-analysis was not possible

Marc et al. reported diagnostic data for a cut-off score of 1. Sensitivity was 0.95 (95% CI 0.75-0.99) and specificity was 0.16 (95% CI 0.09-0.24).

Three studies reported diagnostic data for a cut-off score of 11, which showed great variance. Bae and Cho reported a sensitivity of 0.65 (95% CI 0.51-0.76) and Malakouti et al. reported a sensitivity of 0.55 (95% CI 0.32-0.76). The 95% confidence intervals overlap. In comparison, Marc et al. reported a much lower sensitivity; 0.15 (95% 0.03-0.38). The 95% confidence interval of Marc et al. overlapped with Malakouti et al. but not that of Bae and Cho.

Reported specificities were more similar, especially for the two studies by Bae and Cho and Malakouti et al., which reported specificities of 0.88 (95% CI 0.80-0.94) and 0.88 (95% CI 0.80-0.91) respectively. The 95% confidence intervals of both studies were very close. Marc et al. reported a higher specificity of 0.98 (95% CI 0.93-1.00).

The studies by Malakouti et al. and Marc et al. report diagnostic data at a cut-off score of 12 and 13. At both cut-off scores, reported sensitivity was greater for Malakouti et al. in comparison to Marc et al. Whereas reported specificity for both cut-off scores was greater for Marc et al. compared to Malakouti et al. There was no overlap in confidence intervals for sensitivity for Malakouti et al. and Marc et al. at a cut-off score of 12 and 13.

Meta-analysis

It was possible to perform meta-analysis of the GDS-15 from a cut-off score of 2 to 10. See Table 17. At a cut-off score of 5, which is, as mentioned, the recommended cut-off score of the GDS-15 (Yesavage, 1986), meta-analysis found a pooled sensitivity of 0.89 (95% CI 0.80-0.94) and a pooled specificity of 0.77 (95% CI 0.65-0.86) from 23 studies. See Figure 3.

Effects of different cut-off scores:

In comparison to pooled diagnostic data at a cut-off score of 5, a cut-off score of 4 results in a similar sensitivity but much higher specificity: 0.88 (95% CI 0.67-0.96) and 0.86 (95% CI 0.68-0.94) respectively. The diagnostic odds ratio at a cut-off score of 4 was 42.05 (95% CI 17.42-101.49), which is much higher than that found for the recommended cut-off score of 5 (27.28 (95% CI 16.57-44.93)). Meta-analysis at a cut-off score of 4 involved fewer study participants; 7874 compared to 11468 at a cut-off score of 5.

Sensitivity and specificity, as discussed in Chapter 1, are inversely related; as sensitivity increases, specificity decreases and vice versa. When applied to the measures that use different cut-off scores, a rising cut-off score will lead to an increasing sensitivity and so decreasing specificity. Therefore, predictable changes in regards to sensitivity and specificity are expected with changes in cut-off score.

For the GDS-15, pooled diagnostic data are available for every cut-off score from 2 to 10. As expected, the lowest pooled sensitivity was found for a cut-off score of 10; (0.47; 95% CI 0.27-0.69). However, pooled sensitivity at a cut-off score of 2 (0.90; 95% CI 0.79-0.95) was lower than that found at a cut-off score of 3 (0.95; 95% CI 0.77-0.99). It should be noted that 95% confidence intervals do overlap however.

Pooled sensitivity at a cut-off score of 4 was lower than that found at a cut-off score of 5; 0.88 (95% CI 0.67-0.96) and 0.89 (95% CI 0.80-0.94) respectively, which again is not expected statistically.

The highest pooled specificity was found for a cut-off score of 10 and the lowest pooled specificity was found for a cut-off score of 2. Pooled specificity rises consecutively from a cut-off score of 5; however, pooled specificity at a cut-off score of 4 is greater than that observed at a cut-off score of 5; 0.86 (95% CI 0.68-0.94) and 0.77 (95% CI 0.65-0.86) respectively. There should not be a drop in pooled specificity between a cut-off score of 4 and 5.

As discussed in more detail subsequently, one likely explanation for this phenomenon is that studies are selectively reporting cut-off points on the basis of how well they perform in that particular sample.

Cut-off score	No. of studies	N	Prevalence of major depression (%)	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
2	4	1517	9.8	0.90 (0.79-0.95)	0.43 (0.35-0.51)	1.57 (1.41-1.74)	0.25 (0.12-0.46)	6.38 (3.34-12.20)
3	6	5849	10.7	0.95 (0.77-0.99)	0.68 (0.57-0.77)	2.96 (2.15-4.06)	0.07 (0.01-0.39)	42.04 (6.58-268.52)
4	10	7874	10.1	0.88 (0.67-0.96)	0.86 (0.68-0.94)	6.06 (2.78-13.25)	0.14 (0.05-0.39)	42.05 (17.42-101.49)
5	23	11468	11.5	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)
6	20	9886	11.8	0.79 (0.68-0.87)	0.83 (0.72-0.90)	4.53 (2.85-7.20)	0.26 (0.17-0.38)	17.61 (10.12-30.63)
7	12	8770	11.0	0.72 (0.55-0.85)	0.90 (0.80-0.95)	7.12 (4.09-12.39)	0.31 (0.19-0.51)	22.94 (13.58-38.74)
8	9	7541	10.3	0.70 (0.43-0.88)	0.91 (0.78-0.97)	7.84 (3.69-16.67)	0.33 (0.16-0.67)	23.90 (10.84-52.72)
9	8	3321	9.4	0.52 (0.30-0.73)	0.92 (0.83-0.96)	6.36 (4.08-9.91)	0.52 (0.34-0.79)	12.20 (8.09-18.39)
10	6	3127	9.2	0.47 (0.27-0.69)	0.94 (0.87-0.98)	8.11 (5.30-12.42)	0.56 (0.38-0.82)	14.47 (10.16-20.61)

Table 17: Pooled diagnostic data for the GDS-15

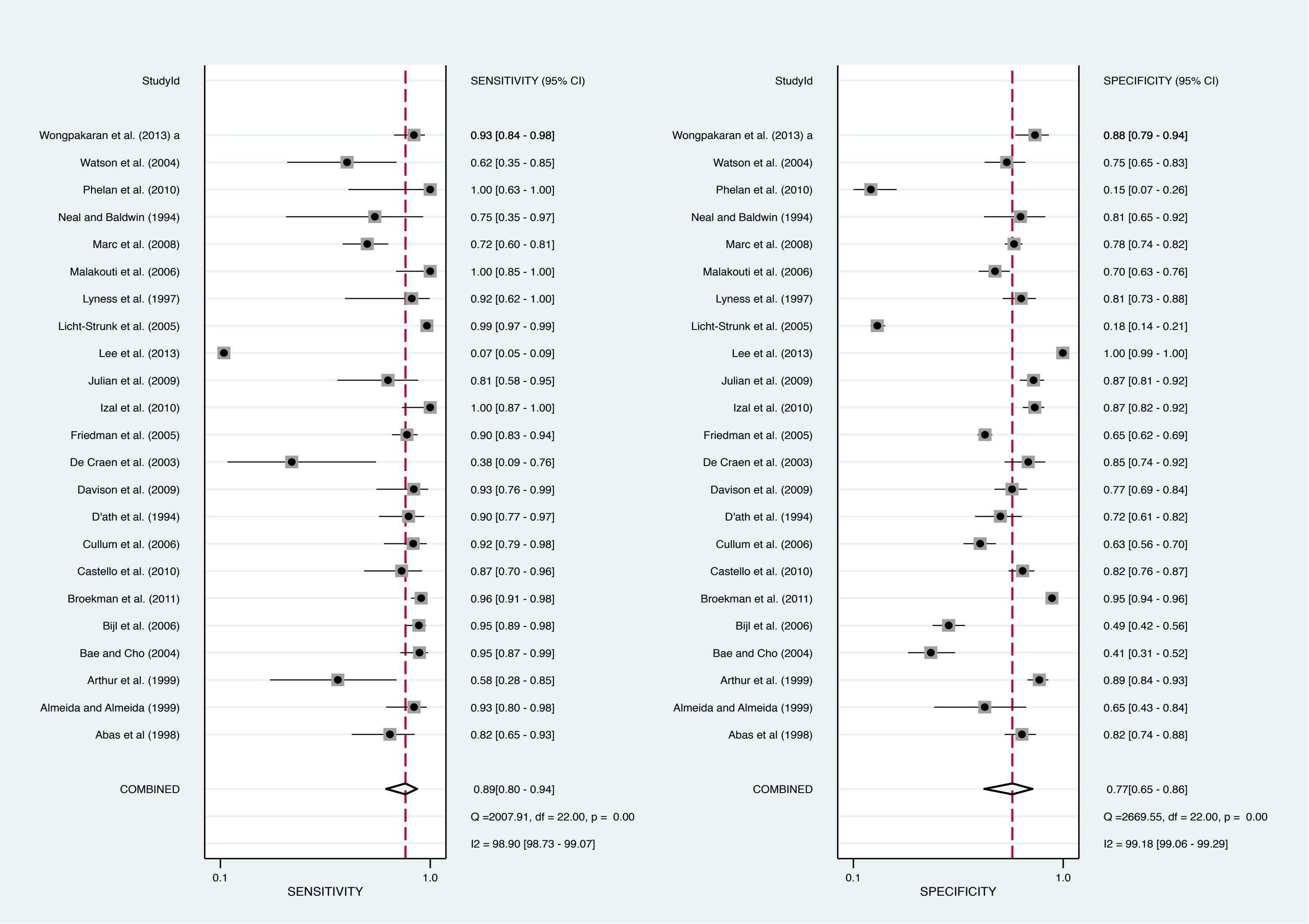

Figure 3: Forest plots of sensitivity and specificity data for primary studies of the GDS-15 at a cut-off score of 5

***Summary receiver operating characteristic (SROC) curves*:**

SROC curves were generated for cut-off scores of 4 – 6 for the GDS-15. See Figures 4 - 6. At a cut-off score of 4, area under curve (AUC) is 0.93 which is higher than the AUC found at cut-off scores of 5 and 6. The AUC was lowest at a cut-off score of 6; 0.87. The 95% confidence intervals for the AUC at the different cut-off scores overlap however suggesting that such differences are not significant.

An AUC of >0.90 suggests that a test has 'excellent' diagnostic accuracy. An AUC of >0.80 suggests a test has 'good' diagnostic accuracy.

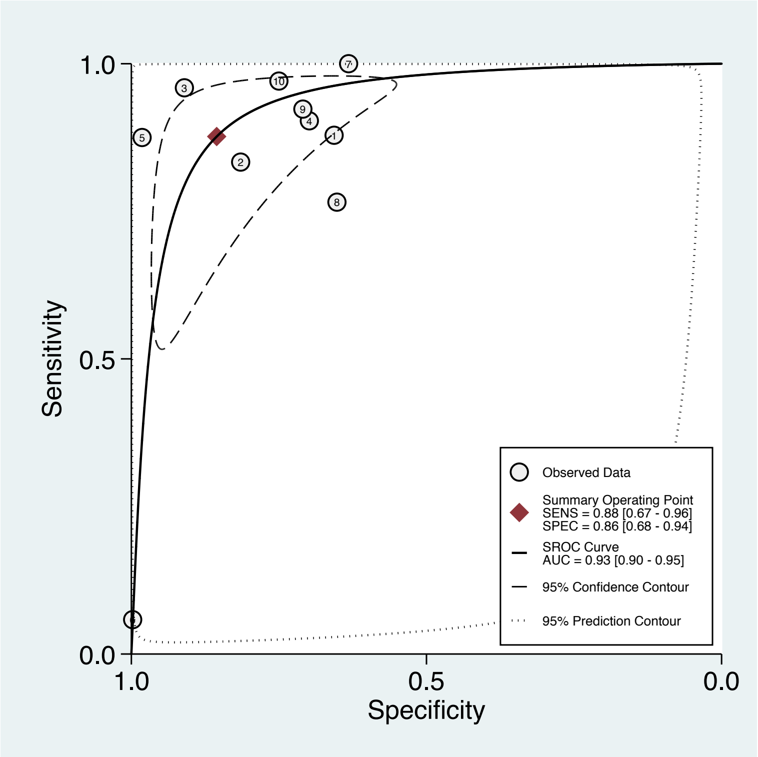

Figure 4: SROC for the GDS-15 at a cut-off score of 4

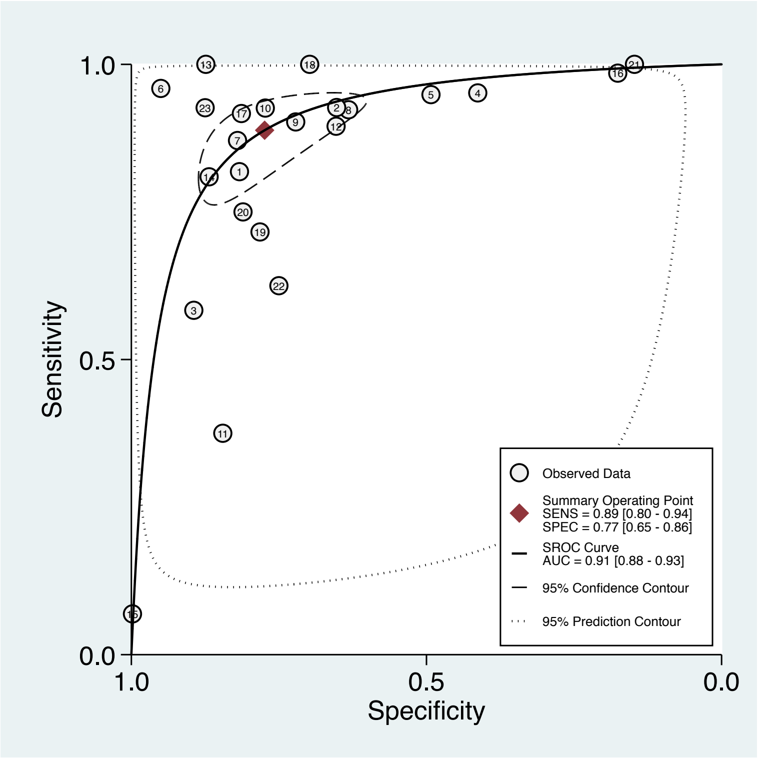

Figure 5: SROC for the GDS 15 at a cut-off score of 5

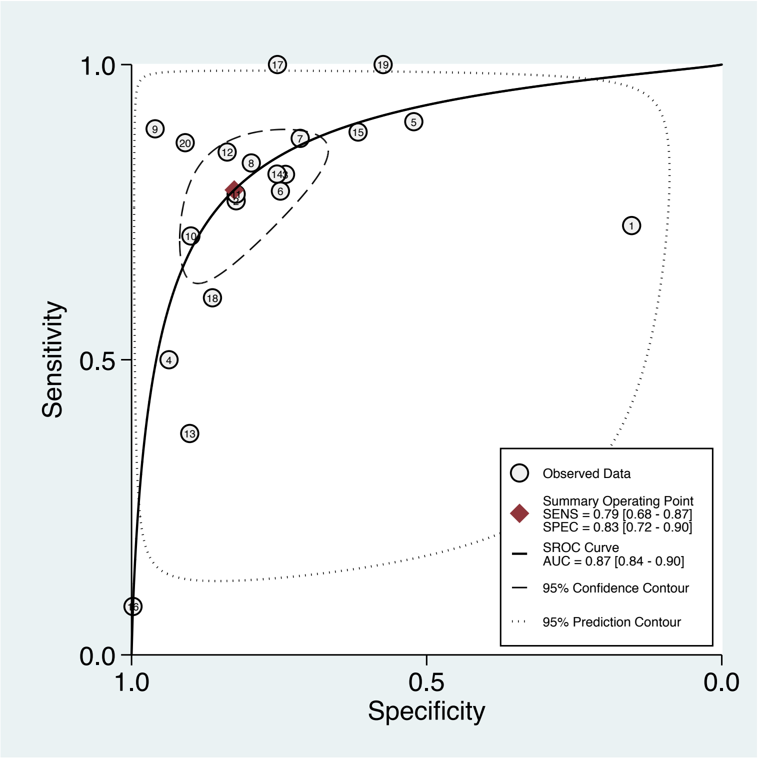

Figure 6: SROC for the GDS 15 at a cut-off score of 6

Heterogeneity:

Of the meta-analyses performed for the nine different cut-off scores of the GDS-15, between-study heterogeneity, as measured by the I^2 statistic, was significantly high for the majority of cut-off scores.

Between-study heterogeneity was high for pooled diagnostic data at the recommended cut-off score of 5 as reflected by the I^2 statistic, which was 76.7%. Cochrane describes such an I^2 statistic as possibly representing 'substantial' or 'considerable' heterogeneity.

Between-study heterogeneity was explored for different cut-off scores of the GDS-15. Where the I^2 statistic was high, studies were identified as 'outliers', at a specific cut-off score, if their diagnostic odds ratio fell outside the pooled diagnostic odds ratio. Such 'outliers' were removed and meta-analysis was re-run. See Table 18. No 'outliers' were identified at a cut-off score of 9.

At a cut-off score of 5, pooled sensitivity increased and pooled specificity decreased when 'outliers' were removed from meta-analysis; however, this pattern was not observed at any other cut-off score. Pooled diagnostic odds ratio fell at all cut-off scores, however.

Pooled sensitivity remained unchanged at a cut-off score of 8 when 'outliers' were removed from meta-analysis. Pooled sensitivity fell slightly at a cut-off score of 4. A larger fall (0.5) was observed at cut-off scores of 6 and 7. An increase in sensitivity was observed at a cut-off score of 3.

Pooled specificities remained unchanged at a cut-off score of 4 and 6 when 'outliers' were excluded from meta-analyses. Pooled specificities fell at other cut-off scores with an exception of a cut-off score of 3 where it increased slightly.

The study by Broekman et al. contributed repeatedly to between-study heterogeneity at the cut-off scores of 4 to 8 as shown by Table 18. The study by Marc et al. also repeatedly contributed to between-study heterogeneity at cut-off scores of 3 to 5 and at a cut-off score of 8. The I^2 statistic dropped greatly when 'outliers' were excluded and meta-analysis re-run. At the cut-off scores of 3, 6 and 7 the I^2 statistic fell from >86.0% to 0.0%. At a cut-off score of 8, I^2 statistic fell by half to 45.6%.

Reducing between-study heterogeneity by removing studies identified as ‘outliers’ had more impact on pooled sensitivities in comparison to pooled specificities for different cut-off scores.

Cut-off score	Sensitivity of all primary studies	Specificity of all primary studies	I^2 statistic (%)	'Outliers' excluded from meta-analysis	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)	New I^2 statistic (%)
3	0.95	0.68	89.9	Marc et al. (2008) Van Marwijk et al. (1995)	0.99 (0.03-1.00)	0.70 (0.58-0.80)	3.35 (2.28-4.92)	0.00 (0.00-47.01)	33.69 (7.59-198.80)	0.0
4	0.88	0.86	85.8	Broekman et al. (2011) Marc et al. (2008)	0.86 (0.59-0.96)	0.86 (0.64-0.96)	6.13 (2.52-14.89)	0.17 (0.06-0.49)	37.17 (18.45-74.90)	14.7
5	0.89	0.77	76.7	Broekman et al. (2011) De Craen et al. Marc et al. (2008) Watson et al.	0.90 (0.81-0.95)	0.75 (0.60-0.86)	3.64 (2.24-5.93)	0.14 (0.08-0.22)	26.84 (19.11-37.69)	8.1
6	0.80	0.83	88.1	Abas et al. (1998) Broekman et al. (2011) Wongpakaran et al. (2013)[1]	0.75 (0.63-0.84)	0.83 (0.73-0.90)	4.39 (3.10-6.23)	0.30 (0.21-0.42)	14.70 (11.60-18.63)	0.0
7	0.72	0.90	86.2	Broekman et al. (2011)	0.67 (0.51-0.81)	0.88 (0.77-0.94)	5.65 (3.53-9.04)	0.37 (0.25-0.55)	15.18 (11.62-10.85)	0.0
8	0.70	0.91	91.2	Broekman et al. (2011) Marc et al. (2008) Phelan et al. (2010)	0.70 (0.39-0.90)	0.89 (0.66-0.97)	6.17 (2.52-15.11)	0.34 (0.16-0.73)	18.36 (10.61-31.77)	45.6

Table 18: Pooled diagnostic data for the GDS-15 excluding 'outliers'

Subgroup analyses

As discussed subgroup analyses were pre-specified.

***a) Participant age*:**

Mean participant age for each primary study was used to classify primary studies into three subsets; young-old (65 – 74 years of age), middle-old (75 – 84 years of age) and old-old (≥85 years of age).

Four studies (Castello et al., 2010, Licht-Strunk et al., 2005, Malakouti et al., 2006 and Van Marwijk et al., 1995) have not been included in subgroup analysis as they do not describe mean study age.

Ten studies were classified as 'young-old' by age (Abas et al., 1998, Almeida and Almeida, 1999, Bae and Cho, 2004, Bijl et al., 2006, Broekman et al., 2011, Julian et al., 2009, Lee et al., 2013, Lyness et al., 1997, Rait et al., 1999, Wongpakaran et al., 2013^1). These ten studies total 7362 study participants.

Fourteen studies were classified as 'middle-old' (Arthur et al., 1999, Blank et al., 2004^1, Blank et al., 2004^2, Blank et al., 2004^3, Cullum et al., 2006, D'ath et al., 1994, Friedman et al., 2005b, Gerety et al., 1994, Izal et la., 2010, Marc et al., 2008, Neal and Baldwin, 1994, Phelan et al., 2010, Watson et al., 2004, Wongpakaran et al., 2013^2). These fourteen studies amount to 2487 study participants.

Four studies were classified as 'old-old' (Allgaier et al., 2013, Davison et al., 2009, De Craen et al., 2003, McCabe et al., 2006. It was not possible to undertake meta-analysis at a cut-off score of 5 due to there being an insufficient (i.e. less than four) number of primary studies.

At the recommended cut-off score of 5, pooled sensitivity is the same for 'young-old' and 'middle-old' studies (i.e. 0.87). See Table 19. Pooled specificity, however, is much higher for 'young-old' studies compared to 'middle-old' studies; 0.86 (95% CI 0.65-0.96) and 0.72 (95% CI 0.56-0.84) respectively. This results in a greater diagnostic odds ratio for 'young-old' studies compared to 'middle old' studies; 42.20 (95% CI 20.33-87.71) and 17.06 (95% CI 2.28-35.02) respectively. It is of note, however, that the confidence intervals of the estimates overlap.

Subgroup analysis		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
None		23	All reporting diagnostic data at a cut-off score of 5	11468	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)	76.7
Mean age of study participants	YO	9	Abas et al. (1998) Almeida and Almeida (1999) Bae and Cho (2004) Bijl et al. (2006) Broekman et al. (2011) Julian et al. (2009) Lee et al. (2013) Lyness et al. (1997) Wongpakaran et al. (2013)[1]	7362	0.87 (0.67-0.96)	0.86 (0.65-0.96)	6.31 (2.46-16.20)	0.15 (0.06-0.37)	42.20 (20.33-87.71)	81.2
	MO	9	Arthur et al. (1998) Cullum et al. (2006) D'ath et al. (1994) Friedman et al. (2005)b Izal et al. (2010) Marc et al. (2008) Neal and Baldwin (1994) Phelan et al. (2010) Watson et al. (2009)	2487	0.87 (0.74-0.94)	0.72 (0.56-0.84)	3.09 (2.00-4.80)	0.18 (0.09-0.35)	17.06 (2.28-35.02)	42.2

Table 19: Pooled diagnostic of subgroup analyses of mean age for the GDS-15 at a cut-off score of 5

Heterogeneity

The I^2 statistic for 'young-old' studies is considerably higher than that for 'middle-old' studies; 81.2% and 42.2% respectively. An I^2 statistic of 81.2% is suggestive of a 'substantial-considerable' level of heterogeneity. An I^2 statistic of 42.2% represents a level of heterogeneity that ranges from 'may not being important' to 'moderate'.

Between-study heterogeneity was explored for the subgroup analysis of mean patient age. 'Outliers' were identified if their diagnostic odds ratio fell outside the pooled diagnostic odds ratio and meta-analysis was re-run excluding them. One study was identified as an 'outlier' for both age classifications; when meta-analysis was re-run excluding these studies the I^2 statistic fell substantially, for example, for studies classified as 'young-old' the I^2 statistic fell from 81.2% to 4.1%. Table 20 shows the I_2 statistic and identified 'outliers' for 'young-old' and 'middle-old' studies.

When meta-analysis for 'young-old' studies was performed excluding the study by Broekman et al. pooled diagnostic data worsened; both sensitivity and specificity.
The pooled diagnostic odds ratio fell notably from 42.4 to 28.84. Less change in the pooled diagnostic odds was observed when the study by Izal et al. was excluded from meta-analysis of 'middle-old' studies.

Age	I^2 statistic (%)	Identified 'outliers'	Diagnostic data excluding 'outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
YO	81.2	Broekman et al. (2011)	0.85 (0.61-0.95)	0.84 (0.58-0.95)	5.27 (2.05-13.56)	0.18 (0.08-0.44)	28.84 (18.01-46.16)	4.1
MO	42.2	Izal et al. (2010)	0.87 (0.67-0.95)	0.69 (0.52-0.82)	2.78 (1.95-3.96)	0.20 (0.09-0.44)	14.19 (7.97-25.27)	7.2

Table 20: Pooled diagnostic data of subgroup analysis of mean participant age excluding 'outliers' at a cut-off score of 5

b) Study setting:

Primary studies were divided into subsets depending on study setting; primary care, secondary care, community and residential/nursing home. Izal et al. was a mixed setting and therefore has not been included subgroup analysis of study setting.

As discussed studies were classified as being based in primary care, secondary care or community based. It was further possible to divide community based studies into participants living independently in the community and participants living in nursing or residential homes. See Table 21 for pooled diagnostic data in accordance to study setting. Owing to an insufficient number of studies it was not possible to perform a meta-analysis of just nursing and residential homes.

The number of study participants in meta-analysis varied by setting; 3124 in primary care, 640 in secondary care, 7471 in all community based studies, and 6665 in studies where participant were living independently in the community.

At a cut-off score of 5, pooled sensitivity was similar between all community based studies and studies of participants living independently in the community; 0.78 (95% CI 0.45-0.94) and 0.79 (95% CI 0.25-0.98) respectively. Pooled sensitivity was higher for primary and secondary care based studies: 0.92 (95% CI 0.83-0.96) and 0.93 (95% CI 0.88-0.96) respectively.

Pooled specificities showed more variance. Pooled specificity for all community based studies and studies of participants living independently in the community were greater than pooled specificity for primary and secondary care studies. For example, pooled specificity for primary care studies was 0.63 (95% CI 0.42-0.80) whereas pooled specificity for studies of participants living independently in the community was 0.94 (95% CI 0.75-0.99). See Table 21. All 95% confidence intervals overlap, which suggests that differences are not statistically significant. The lowest diagnostic odds ratio was found for primary care studies (18.58 (95% CI 13.14-26.27)), then for secondary care studies (29.00 (95% CI 13.27-63.38)), followed closely by all community based studies (29.31 (95% CI 9.19-93.47)) and finally studies where participants live independently (56.71 (95% CI 10.32-311.38)). Again, all 95% confidence intervals overlap.

Subgroup analysis		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Study setting	Primary care	9	Abas et al. (1998) Arthur et al. (1998) Bijl et al. (2005) Castello et al. (2010) D'ath et al. (1994) Friedman et al. (2005)b Licht-Strunk et al. (2005) Lyness et al. (1997) Phelan et al. (2010)	3124	0.92 (0.83-0.96)	0.63 (0.42-0.80)	2.49 (1.54-4.03)	0.13 (0.08-0.21)	18.58 (13.14-26.27)	0.0
	Secondary care	5	Almeida & Almeida (1999) Bae and Cho (2004) Cullum et al. (2006) Neal and Baldwin (1994)) Wongpakaran et al. (2013)[1]	640	0.93 (0.88-0.96)	0.70 (0.53-0.83)	3.05 (1.85-5.04)	0.11 (0.07-0.17)	29.00 (13.27-63.38)	37.8
	All community	8	Broekman et al. (2011) Davison et al. (2009) De Craen et al. (2002) Julian et al. (2009) Lee et al. (2013) Malakouti et al. (2006) Marc et al. (2008) Watson et al. (2009)	7471	0.78 (0.45-0.94)	0.90 (0.74-0.96)	7.36 (3.22-16.83)	0.25 (0.09-0.72)	29.31 (9.19-93.47)	91.2
	Community independent living	5	Broekman et al. (2011) De Craen et al. (2002) Julian et al. (2009) Lee et al. (2013) Malakouti et al. (2006)	6665	0.79 (0.25-0.98)	0.94 (0.75-0.99)	12.81 (3.80-43.21)	0.23 (0.04-1.40)	56.71 (10.32-311.38)	89.1

Table 21: Pooled diagnostic of subgroup analyses of study setting for the GDS-15 at a cut-off score of 5

Heterogeneity

Heterogeneity varied by study setting, ranging from 0.0 – 91.2%. The lowest measure of heterogeneity was found for primary care based studies; the I^2 statistic was 0.0%. The highest I^2 statistic was for all community based studies; 91.2%. Heterogeneity was also at a 'substantial-considerable' level for studies of participants living independently in the community; 89.1%. The I^2 statistic for secondary care based studies was 37.8%, which represents heterogeneity at a level that may not be important though could represent heterogeneity at a moderate level. See Table 21.

No 'outliers' were identified for primary or secondary care based studies. Four studies were identified as 'outliers' for community based studies – see Table 22. Heterogeneity for community based studies dropped notably when 'outliers' were removed and meta-analyses re-run. The I^2 statistic fell from 91.2% to 0.0%. Pooled sensitivity remained relatively the same, whereas pooled specificity increased slightly. This result in the pooled diagnostic odds ratio increasing from 29.31 (95% 9.19-93.47) to 40.27 (95% CI 18.52-87.56). Differences are not statistically significant as the 95% confidence interval overlap. Two studies were identified as 'outliers' for studies where participants were living independently in the community. The I^2 statistic feel from 89.1% to 0.0% when these studies were excluded, however meta-analysis was not possible due to an insufficient number of studies remaining (i.e. three studies).

Setting	I^2 statistic (%)	Identified 'outliers'	Diagnostic data excluding 'outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
Primary care	0.0	None identified	n/a					
Secondary care	37.8	None identified	n/a					
All community	91.2	Broekman et al. (2011) De Craen et al. (2003) Marc et al. (2008) Watson et al. (2004)	0.77 (0.25-0.97)	0.92 (0.62-0.99)	10.06 (2.55-39.72)	0.25 (0.05-1.28)	40.27 (18.52-87.56)	0.0
Community independent living	89.1	Broekman et al. (2011) De Craen et al. (2003)	Insufficient number of studies to perform meta-analysis					0.0

Table 22: Pooled diagnostic data of subgroup analysis of study setting excluding 'outliers' at a cut-off score of 5

c) Country setting:

Primary studies were divided into subsets of Western and non-Western countries and meta-analysis re-run in accordance to country of study setting. See Table 23. At the recommended cut-off score of 5, pooled sensitivity was found to be slightly greater for studies from non-Western countries compare to studies from Western countries; 0.90 (95% CI 0.45-0.99) and 0.88 (95% CI 0.81-0.93) respectively. Fewer studies were included in meta-analyses for non-Western countries compared to Western countries; 5 and 18 respectively. However, the number of participants included in meta-analysis for non-Western countries (6708) was greater than that of Western countries (3130).

Pooled specificity was found to be considerably higher in studies from non-Western countries compare to studies from Western countries; 0.90 (95% CI 0.59-0.98) and 0.72 (95% CI 0.61-0.81) respectively. The pooled diagnostic ratio for studies from Western countries is considerably higher than that studies from non-Western countries; 79.66 (95% CI 19.52-325.14) and 19.09 (95% CI 13.14-27.75) respectively. The overlap of the 95% confidence intervals signify such differences are unlikely to be statistically significant.

<u>Heterogeneity</u>

Heterogeneity was greater for studies from non-Western countries; an I^2 statistic of 84.1% represents a 'substantial-considerable' level of heterogeneity. The I^2 statistic for Western countries was lower at 34.1%, which may suggest that heterogeneity is less marked.

For studies from Western countries, the study by Izal et al. was identified as an 'outlier'. The I^2 statistic fell from 32.4% to 19.2% when meta-analysis was re-run excluding this study, which suggests that heterogeneity may no longer be important. Both pooled sensitivity and specificity were found to fall slightly; 0.87 (95% CI 0.79-0.92) and 0.70 (95% CI 0.59-0.80) respectively. See Table 24. The pooled diagnostic odds ratio fell from 19.09 (95% CI 13.14-27.75) to 16.16 (95% CI 12.13-21.51).

For studies from non-Western countries, the study by Broekman et al. was identified as an 'outlier'. When meta-analysis was re-run, the I^2 statistic fell from 84.1% to 43.0%, which still represents moderate heterogeneity. Pooled sensitivity fell from 0.90 (95% CI 0.45-0.99) to 0.85 (95% CI 0.29-0.99) but pooled specificity remained unchanged. The pooled diagnostic odds ratio fell from 79.66 (95% CI 19.52-325.14) to 48.98 (95% CI 18.07-132.73). Again, findings were not statistically significant as the 95% confidence intervals overlapped.

Subgroup analysis		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Studies based in Western countries	Yes	18	Abas et al. (1998) Almeida and Almeida (1999) Arthur et al. (1999) Bijl et al. (2006) Castello et al. (2010) Cullum et al. (2006) D'ath et al. (1994) Davison et al. (2009) De Craen at al. (2003) Friedman et al. (2005)b Izal et al. (2010) Julian et al. (2009) Licht-Strunk et al. (2005) Lyness et al. (1997) Marc et al. (2008) Neal and Baldwin (1994) Phelan et al. (2010) Watson et al. (2004)	3130	0.88 (0.81-0.93)	0.72 (0.61-0.81)	3.13 (2.30-4.26)	0.16 (0.11-0.25)	19.09 (13.14-27.75)	32.4
	No	5	Bae and Cho (2004) Broekman et al. (2011) Lee et al. (2013) Malakouti et al. (2006) Wongpkaran et al. (2013)[1]	6708	0.90 (0.45-0.99)	0.90 (0.59-0.98)	9.27 (2.05-42.01)	0.12 (0.02-0.82)	79.66 (19.52-325.14)	84.1

Table 23: Pooled diagnostic of subgroup analyses of country of study for the GDS-15 at a cut-off score of 5

Western Country	**I^2 statistic (%)**	**Identified ‘outliers’**	**Diagnostic data excluding ‘outliers'**					
			Sensitivity (95% CI)	**Specificity (95% CI)**	**PLR (95% CI)**	**NLR (95% CI)**	**DOR (95% CI)**	**New I^2 statistic (%)**
Yes	32.4	Izal et al. (2010)	0.87 (0.79-0.92)	0.70 (0.59-0.80)	2.95 (2.19-3.96)	0.18 (0.13-0.27)	16.16 (12.13-21.51)	19.2
No	84.1	Broekman et al. (2011)	0.85 (0.29-0.99)	0.90 (0.41-0.99)	8.46 (1.23-58.43)	0.17 (0.02-1.22)	48.98 (18.07-132.73)	43.0

Table 24: Pooled diagnostic data for subgroup analysis country of study excluding ‘outliers’ at a cut-off score of 5

Sensitivity analyses

As discussed sensitivity analyses of domains of the QUADAS-II were pre-specified.

a) Patient selection:

For sensitivity analysis according to the QUADAS-II domain of patient selection, meta-analysis is only possible for studies rated as having a 'low' risk of bias – in total this amounts to 27 samples. Only three samples were rated as having a 'high' risk of bias for patient selection (Almeida and Almeida, 1999, Bae and Cho, 2004, McCabe et al., 2006). Two samples were rated as having an 'unclear' risk of bias (Castello et al., 2010, Lee et al., 2013). A minimum of four studies is required to perform a diagnostic meta-analysis and therefore meta-analysis was not possible for studies rated as having an overall 'high' or 'unclear' risk of bias 'patient selection'.

At the recommended cut-off score of 5, pooled sensitivity was found to be 0.90 (95% CI 0.84-0.95) and pooled specificity was 0.75 (95% CI 0.64-0.83) for studies rated as having an overall 'low' risk of bias. This results in a pooled diagnostic odds ratio of 27.65 (95% CI 15.11-50.57). Meta-analysis included 9089 participants. See Table 25.

Little difference is observed in comparison of pooled diagnostic data for studies rated as having an overall 'low' risk of bias for 'patient selection' with pooled diagnostic data from meta-analysis of all studies. Pooled sensitivity and diagnostic odds ratios remain relatively the same. However, pooled specificity decreases slightly; it falls from 0.77 (95% CI 0.65-0.86) to 0.75 (95% CI 0.64-0.83).

Heterogeneity

Three studies, that had a 'low' risk of bias, were identified as 'outliers' at a cut-off score of 5 (Broekman et al., 2011, De Craen et al., 2003, Watson et al., 2004). When meta-analysis was re-run the I^2 statistic fell from 80.6% to 22.1%. Pooled sensitivity increased slightly (from 0.90 (95% CI 0.84-0.95) to 0.92 (95% CI 0.87-0.95)) whereas pooled specificity fell more (from 0.75 (95% CI 0.64-0.83) to 0.71 (95% CI 0.57-0.81). The pooled diagnostic odds ratio remained relatively the same; 27.65 (95% CI 15.11-50.57) compared to 27.90 (95% CI 18.51-42.06) when meta-analysis was re-run excluding 'outliers'. As 95% confidence intervals overlap differences are not statistically significant. See Table 26.

Sensitivity analysis of ratings of risk of bias of QUADAS-II		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Domain	Rating									
None		23	All reporting diagnostic data at a cut-off score of 5	11468	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)	76.7
Bias rating of 'participant selection'	Low	19	Abas et al. (1998) Arthur et al. (1999) Bijl et al. (2006) Broekamn et al. (2011) Cullum et al. (2006) D'ath et al. (1994) Davison et al. (2009) De Craen et al. (2003) Friedman et al. (2005)b Izal et al. (2010) Julian et al. (2009) Licht-Strunk et al. (2005) Lyness et al. (1997) Malakouti et al. (2006) Marc et al. (2008) Neal and Baldwin (1994) Phelan et al. (2010) Watson et al. (2004) Wongpakaran et al. (2013)[1]	9089	0.90 (0.84-0.95)	0.75 (0.64-0.83)	3.56 (2.48-5.12)	0.13 (0.08-0.21)	27.65 (15.11-50.57)	80.6

Table 25: Pooled diagnostic data for sensitivity analysis for QUADAS-II domain of 'risk of bias of participant selection' at a cut-off score of 5

Bias rating of ‘participant selection’	I^2 statistic (%)	Identified ‘outliers’	Diagnostic data excluding ‘outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
Low	80.6	Broekman et al. (2011) De Craen et al. (2003) Marc et al. (2008) Watson et al. (2004)	0.92 (0.87-0.95)	0.71 (0.57-0.81)	3.15 (2.14-4.62)	0.11 (0.08-0.16)	27.90 (18.51-42.06)	22.1

Table 26: Pooled diagnostic data of sensitivity analysis of QUADAS-II domain of ‘risk of bias of participant selection’ excluding ‘outliers’ at a cut-off score of 5

b) *Index test*:

Sixteen samples were rated as having a 'low' risk of bias for the QUADAS-II domain of index test (Abas et al., 1998, Allgaier et al., 2013, Bae and Cho, 2004, Blank et al., 2004^1, Blank et al, 2004^2, Blank et al., 2004^3, Broekman et al., 2011, Davison et al., 2009, Izal et al., 2010, Julian et al., 2009, Malakouti et al., 2006, Marc et al., 2008, McCabe et al., 2006, Phelan et al., 2010, Wongpakaran et al., 2013^1, Wongpakaran et al., 2013^2). Ten of these studies reported diagnostic data at a cut-off score of 5 and were therefore included in meta-analysis, which resulted in 6115 study participants. See Table 27.

Ten samples were rated as having an 'unclear' risk of bias for the domain of index test (Almeida and Almeida, 1999, Bijl et al., 2006, Cullum et al., 2006, D'ath et al., 1994, De Craen et al., 2003, Lee et al., 2013, Licht-strunk et al., 2005, Neal and Baldwin, 1994, Van Marwijk et al., 1995, Watson et al., 2004). All but one study (Van Marwijk et al., 1995), amounting to 3842 study participants, report diagnostic data at a cut-off score of 5.

Six samples were rated as having an 'high' risk of bias for the domain of index test (Arthur et al., 1999, Castello et al., 2010, Friedman et al., 2005b, Gerety et al., 1994, Lyness et al., 1997, Rait et al., 1999). Four of these studies, resulting in a 1511 participants, reported diagnostic data at a cut-off score of 5 and were therefore included in meta-analysis. See Table 27.

When studies rated as having a 'high' or 'unclear' risk of bias for the domain of index test were excluded from meta-analysis, at a cut-off score of 5, pooled sensitivity increased from 0.89 (95% CI 0.80-0.94) to 0.93 (95% 0CI 0.86-0.97). See Table 27. Little change is observed in pooled specificity; pooled specificity of studies rated as having a 'low' risk of bias is 0.76 (95% CI 0.60-0.88); whereas pooled specificity from meta-analysis of all studies is 0.77 (95% CI 0.65-0.86).

The pooled diagnostic odds ratio rises considerably when studies rated as having a 'high' or 'unclear' risk of bias are excluded from meta-analysis; the diagnosis odds ratio of all studies and studies rated as having a 'low' risk of bias was 27.28 (95% CI 16.57-44.93) and 44.31 (95% I 15.79-124.30) respectively. See Table 27.

Heterogeneity

Heterogeneity is much greater for meta-analysis of studies rated as having a ‘low’ risk of bias compared to heterogeneity for all studies; the I^2 statistic is 87.8% and 76.7% respectively.

The samples by Broekman et al. and Marc et a. were identified as being ‘outliers’ for studies rated as having a ‘low’ risk of bias. See Table 28. When meta-analysis was re-run excluding these two studies the I^2 statistic fell from 87.8% to 42.5%; an I^2 statistic of 42.5% represents ‘moderate’ heterogeneity. Pooled sensitivity remained similar but pooled specificity fell from 0.76 (95% CI 0.60-0.88) to 0.72 (95% CI 0.52-0.86). This led to a reduction in the pooled diagnostic odds ratio; from 44.31 (95% CI 15.79-124.30) to 38.80 (95% CI 16.09-93.58).

Sensitivity analysis of ratings of risk of bias of QUADAS-II		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Domain	Rating									
None		23	All reporting diagnostic data at a cut-off score of 5	11468	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)	76.7
Bias rating of 'index test'	High	4	Arthur et al. (1999) Castello et al. (2010) Friedman et al. (2005)b Lyness et al. (1997)	1511	0.83 (0.71-0.91)	0.80 (0.70-0.88)	4.23 (2.86-6.26)	0.21 (0.13-0.36)	19.83 (11.39-34.50)	0.0
	Unclear	9	Almeida and Almeida (1999) Bijl et al. (2006) Cullum et al. (2006) D'ath et al. (1994) De Craen et al. (2003) Lee et al. (2013) Licht-Strunk et al. (2005) Neal and Baldwin (1994) Watson et al. (2004)	3842	0.82 (0.58-0.94)	0.76 (0.49-0.91)	3.39 (1.72-6.70)	0.23 (0.11-0.49)	14.52 (9.82-21.46)	27.5
	Low	10	Abas et al. (1998) Bae and Cho (2004) Broekman et al. (2011) Davison et al. (2009) Izal et al. (2010) Julian et al. (2009) Malakouti et al. (2006) Marc et al. (2008) Phelan et al. (2010) Wongpakaran et al. (20013)[1]	6115	0.93 (0.86-0.97)	0.76 (0.60-0.88)	3.92 (2.18-7.05)	0.09 (0.04-0.19)	44.31 (15.79-124.30)	87.8

Table 27: Pooled diagnostic data for sensitivity analysis of risk of bias ratings from domains of the QUADAS-II at a cut-off score of 5

Bias rating of 'index test'	I^2 statistic (%)	Identified 'outliers'	Diagnostic data excluding 'outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
High	0.0	None identified	n/a					
Unclear	27.5	None identified	n/a					
Low	87.8	Broekman et al. (2011) Marc et al. (2008)	0.94 (0.87-0.97)	0.72 (0.52-0.86)	3.30 (1.83-5.93)	0.09 (0.04-0.17)	38.80 (16.09-93.58)	42.5

Table 28: Pooled diagnostic data of sensitivity analysis of QUADAS-II domain of 'risk of bias of index test' excluding 'outliers' at a cut-off score of 5

c) *Reference test*:

For the QUADAS-II domain of reference test, 22 samples were rated as having a 'low' risk of bias. The study by Van Marwijk et al. was not included in meta-analysis as it only reported diagnostic data at a cut-off score of 2 and 3. One study was rated as having a 'high' risk of bias (Malakouti et al., 2006). Nine studies were rated as having an 'unclear' risk of bias (Allgaier et al., 2013, Almeida and Almeida, 1999, Arthur et al., 1999, Cullum et al., 2006, D'ath et al., 1994, De Craen et al., 2003, Freidman et al., 2005b, Lee et al., 2013, Licht-Strunk et al., 2005).

When meta-analysis was re-run excluding studies rated as having a 'high' or 'unclear' risk of bias pooled diagnostic data remained relatively unchanged. For example, the pooled sensitivity of studies rated as having a 'low' risk of bias for the domain of reference test was 0.90 (95% CI 0.84-0.94) and the pooled sensitivity of all studies is 0.89 (95% CI 0.80-0.94). Pooled specificities are 0.76 (95% CI 0.64-0.85) and 0.77 (95% CI 0.65-0.86) respectively. Meta-analysis of studies rated as having a 'low' risk of bias included 6730 study participants. See Table 29. The pooled diagnostic odds ratio did improve slightly, when meta-analysis was re-run including just studies rated as having a 'low' risk of bias, from 27.28 (95% CI 16.57-44.93) to 30.41 (95% CI 14.27-64.83).

Heterogeneity

Heterogeneity was deemed to be at a 'substantial to considerable' level for meta-analyses of studies rated as having a 'low' risk; the I^2 statistic was 87.8%. The I^2 statistic was lower for all studies (76.7%), though such a figure still represents a 'substantial to considerable' level of heterogeneity.

Three studies rated as having a 'low' risk of bias were rated as 'outliers'; Broekman et al., Marc et al. and Watson et al. Heterogeneity did improve when meta-analyses were re-run excluding these 'outliers'; the I^2 statistic fell from 87.8% to 26.0%. An I^2 statistic of 26.0% may still represents a 'moderate' level of heterogeneity. When meta-analysis was re-run pooled sensitivity increased slightly (from 0.90 (95% CI 0.84-0.94) to 0.92 (95% CI 0.88-0.95)), whereas pooled specificity fell slightly (from 0.76 (95% CI 0.64-0.85) to 0.73 (95% CI 0.58-0.84)). The pooled diagnostic odds ratio however did not change much at all. See Table 30.

Sensitivity analysis of ratings of risk of bias of QUADAS-II		No. of studie s	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Domain	Rating									
None		23	All reporting diagnostic data at a cut-off score of 5	11468	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)	76.7
Bias rating of 'reference test'	Unclear	8	Almeida and Almeida (1999) Arthur et al. (1999) Cullum et al. (2006) D'ath et al. (1994) De Craen et al. (2003) Friedman et al. (2005)b Lee et al. (2013) Licht-Strunk et al. (2005)	4534	0.81 (0.54-0.94)	0.80 (0.50-0.94)	3.99 (1.69-9.39)	0.24 (0.11-0.53)	16.87 (11.75-24.22)	0.0
	Low	14	Abas et al. (1998) Bae and Cho (2004) Bijl et al. (2006) Broekman et al. (2011) Castello et al. (2010) Davison et al. (2009) Izal et al. (2010) Julian et al. (2009) Lyness et al. (1997) Marc et al. (2008) Neal and Baldwin (1994) Phelan et al. (2010) Watson et al. (2004) Wongpakaran et al. (20013)[1]	6730	0.90 (0.84-0.94)	0.76 (0.64-0.85)	3.82 (2.45-5.96)	0.13 (0.08-0.21)	30.41 (14.27-64.83)	84.4

Table 29: Pooled diagnostic data for sensitivity analysis of risk of bias ratings from domains of the QUADAS-II at a cut-off score of 5

Bias rating of 'reference test'	I^2 statistic (%)	Identified 'outliers'	Diagnostic data excluding 'outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
Unclear	0.0	None identified	n/a					
Low	87.8	Broekman et al. (2011) Marc et al. (2008) Watson et al. (2004)	0.92 (0.88-0.95)	0.73 (0.58-0.84)	3.42 (2.14-5.45)	0.11 (0.08-0.16)	31.08 (17.95-53.85)	26.0

Table 30: Pooled diagnostic data of sensitivity analysis of QUADAS-II domain of 'risk of bias of reference test' excluding 'outliers' at a cut-off score of 5

d) *Flow/timing of study*:

For the QUADAS-II domain of flow/timing of study design, 19 samples were rated as having a 'low' risk of bias. Seven samples were rated as having a 'high' risk of bias (Abas et al., 1998, Almeida and Almeida, 1999, Arthur et al., 1999, Cullum et al., 2006, D'ath et al., 1994, Licht-strunk et al., 2005, Malakouti et al., 2006). Six samples were rated as having an 'unclear' risk of bias (Castello et al., 2010, Friedman et al., 2005b, Lee et al., 2003, Van Marwijk et al., 1995, Wongpakaran et al., 2013[1-2]. The studies by Van Marwijk et al. and Wongpakaran et al. have not been included in sensitivity analysis as they do not report diagnostic data at a cut-off score of 5.

Again, the pooled diagnostic data from meta-analysis excluding studies rated as having a 'high' or 'unclear' risk of bias remain relatively unchanged. For example, the pooled specificity of studies rated as having a 'low' risk of bias was similar to the pooled specificity of all studies; 0.75 (95% CI 0.60-0.86) and 0.77 (95% CI 0.65-0.86) respectively. See Table 31. Pooled sensitivities are closer; 0.90 (95% CI 0.79-0.95) for studies rated as having a 'low' risk of bias and 0.89 (95% CI 0.80-0.94) for all studies. The pooled diagnostic odds ratio of studies rated as having a 'low' risk of bias was 25.77 (95% CI 10.13-65.56) and the pooled diagnostic odds ratio of all studies was 27.28 (95% CI 16.57-44.93) The number of study participants included in meta-analysis varied considerably; 1992 – 6269. See Table 31.

Heterogeneity

Heterogeneity was found to be high for studies rated as having a 'low' risk of bias; the I^2 statistic was 86.9%. Four studies rated as having a 'low' risk of bias for the domain of study 'flow/timing' were identified as 'outliers'; Broekman et al., De Craen et al., Marc et al. and Watson et al. When meta-analysis was re-run, the I^2 statistic improved from 86.9% to 14.8%. Pooled sensitivity increased slightly, whereas pooled specificity fell. This resulted in an increase in the pooled diagnostic odds ratio; 25.77 (95% CI 10.13-65.56) to 28.95 (95% CI 14.11-59.42). See Table 32.

Sensitivity analysis of ratings of risk of bias of QUADAS-II		No. of studies	Studies included	N	Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	I^2 (%)
Domain	Rating									
None		23	All reporting diagnostic data at a cut-off score of 5	11468	0.89 (0.80-0.94)	0.77 (0.65-0.86)	3.93 (2.58-6.00)	0.14 (0.09-0.24)	27.28 (16.57-44.93)	76.7
Bias rating of 'flow/timings'	High	7	Abas et al. (1998) Almeida and Almeida (1999) Arthur et al. (1999) Cullum et al. (2006) D'ath et al. (1994) Licht-Strunk et al. (2005) Malakouti et al. (2006)	1922	0.91 (0.82-0.96)	0.67 (0.48-0.82)	2.77 (1.70-4.51)	0.13 (0.08-0.23)	20.80 (13.26-32.60)	0.0
	Unclear	4	Castello et al. (2010) Friedman et al. (2005)b Lee et al. (2013) Wongpakaran et al. (20013)[1]	3277	0.72 (0.25-0.95)	0.93 (0.59-0.99)	10.03 (2.23-45.17)	0.30 (0.08-1.15)	33.47 (16.17-69.27)	60.6
	Low	12	Bae and Cho (2004) Bijl et al. (2006) Broekman et al. (2011) Davison et al. (2009) De Craen et al. (2003) Izal et al. (2010) Julian et al. (2009) Lyness et al. (1997) Marc et al. (2008) Neal and Baldwin (1994) Phelan et al. (2010) Watson et al. (2004)	6269	0.90 (0.79-0.95)	0.75 (0.60-0.86)	3.57 (2.18-5.85)	0.14 (0.07-0.28)	25.77 (10.13-65.56)	86.9

Table 31: Pooled diagnostic data for sensitivity analysis of risk of bias ratings from domains of the QUADAS-II at a cut-off score of 5 cont.

Bias rating of 'flow/timing'	I^2 statistic (%)	Identified 'outliers'	Diagnostic data excluding 'outliers'					
			Sensitivity (95% CI)	Specificity (95% CI)	PLR (95% CI)	NLR (95% CI)	DOR (95% CI)	New I^2 statistic (%)
High	0.0	None identified	n/a					
Unclear	60.6	None identified	n/a					
Low	86.9	Broekman et al. (2011) De Craen et al. (2003) Marc et al. (2008) Watson et al. (2004)	0.93 (0.88-0.96)	0.68 (0.47-0.83)	2.91 (1.66-5.09)	0.10 (0.06-0.16)	28.95 (14.11-59.42)	14.8

Table 32: Pooled diagnostic data of sensitivity analysis of QUADAS-II domain of 'risk of bias of study flow/timing' excluding 'outliers' at a cut-off score of 5

Meta-regression

The effects of study characteristics on pooled summary estimates for the GDS-15, at a cut-off score of 5, were explored by meta-regression analysis of the logic diagnostic odds ratio. Meta-regression included seven explanatory variables; country of study setting (Western or non-Western), study healthcare setting (primary care or not), proportion of study participants female, participant mean age, administration of GDS-15 (self-administrated or not), use of an extracted GDS-15 (or not), and language of the GDS-15 (English or non-English). Background reading in addition to subgroup and sensitivity analyses influenced the selection of explanatory variables.

Meta-regression revealed that two of the explanatory variables were predictive of diagnostic accuracy; country (i.e. non-Western) (p=0.005) and language (i.e. non-English) (p=0.05). The remaining five variables were found not to influence diagnostic accuracy. See Table 33.

Study explanatory variable	Coefficient	Standard error	t	p value	95% confidence interval
Western country	-1.54	0.48	-3.21	0.005	-2.55 – -0.53
Primary care setting	-0.22	0.50	-0.44	0.66	-1.28 – 0.83
Proportion female	-0.02	0.03	-0.63	0.54	-0.09 – 0.05
Mean age	-0.07	0.04	-1.71	0.11	-0.16 – 0.02
Self-administration of GDS-15	0.16	0.62	0.26	0.80	-1.14 – 1.46
Extracted GDS-15	-0.04	0.67	-0.07	0.95	-1.45 – 1.36
English version of GDS-15	-0.95	0.45	-2.10	0.05	-1.89 – 0.00

Table 33: Results of meta-regression analysis

Publication bias

Publication bias was explored using a funnel plot. See Figure 7 for a funnel plot of the diagnostic data for the GDS-15 at a cut-off score of 5. There is some indication of asymmetry in the funnel plot, which may be suggestive of publication bias. Studies appear to be missing from the bottom right corner but more so from the bottom left corner.

Six studies fall outside the 95% confidence intervals, particularly so for one study, which can be seen near the top right corner.

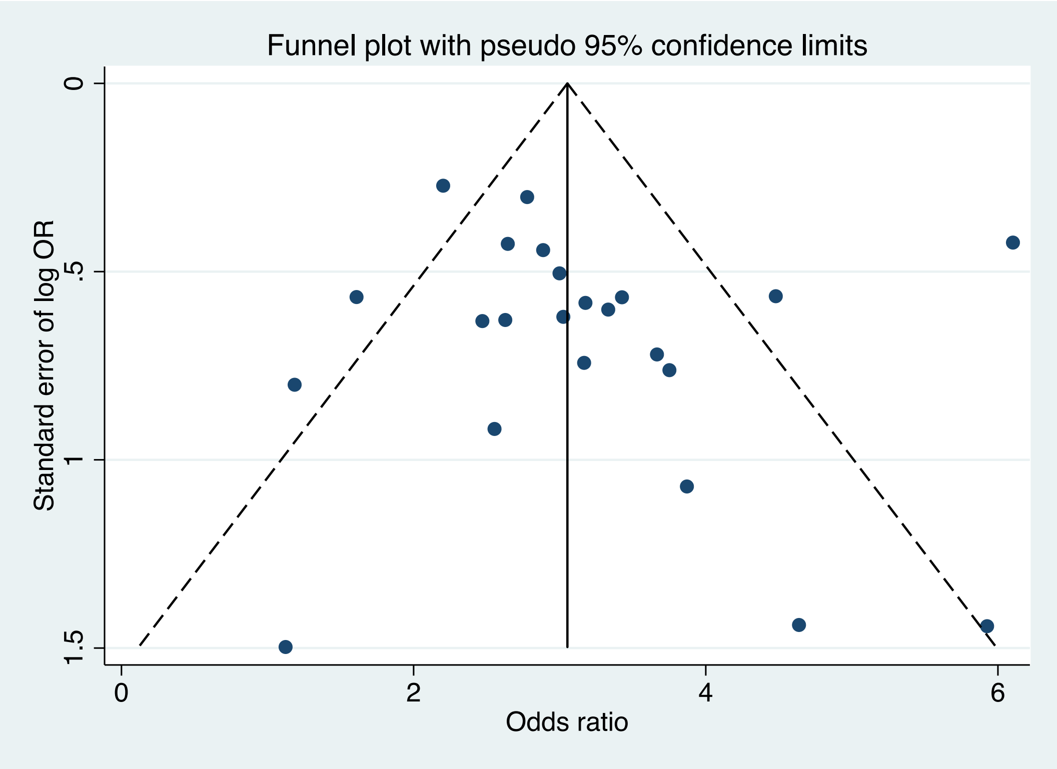

Figure 7: Funnel plot for studies reporting diagnostic accuracy of the GDS-15 at a cut-off score of 5

DISCUSSION

The aim of this systematic review was to establish the diagnostic accuracy of the GDS-15 and briefer versions.

Briefer versions of the GDS

In total, six different briefer versions of the GDS were identified in this review; GDS-1, GDS-4, GDS-5, GDS-7, GDS8 and GDS-10. Unlike the GDS-15, there are no set standardised items for briefer versions. As discussed previously for example, the items comprising the GDS-4 in the study by Allgaier et al. differed from the items comprising the GDS-4 in the study by Castello et al., so these two versions of the GDS-4 were in effect different rating scales.

Meta-analysis was not possible for briefer versions of the GDS due to an insufficient number of studies, which was secondary to different item compositions of the briefer version of the GDS in question and an inadequate number of studies for different reported cut-off scores.

Diagnostic accuracy of the GDS-15

This review established for the GDS-15, for the recommended cut-off score of 5, a pooled sensitivity of 0.89 (95% CI 0.80–0.94) and a pooled specificity of 0.77 (95% CI 0.65–0.86). The diagnostic odds ratio was 27.28 (95% CI 16.57– 44.93). Heterogeneity was established as being at a 'substantial' to 'considerable' level as reflected by an I^2 statistic of 76.7%. When meta-analysis was re-run excluding 'outliers' (Broekman et al., 2011, De Craen et al., 2003, Marc et al., 2008, Watson et al., 2004) pooled sensitivity only increased slightly to 0.90 (95% CI 0.81-0.95) and pooled specificity fell slightly to 0.75 (95% CI 0.60-0.86). The I^2 statistic fell to 8.1%.

Pooled diagnostic data at a cut-off score of 5 was compared to pooled diagnostic at other cut-off scores. Pooled diagnostic data at a cut-off score of 4 appears more favourable; sensitivity 0.88 (95% CI 0.67–0.96), specificity 0.86 (95% CI 0.68–0.94) and diagnostic odds ratio 42.05 (95% CI 17.42–101.49). However, meta-analysis at a cut-off score of 4 only included 10 studies. At a cut-off score of 4, the SROC curve shows an

AUC of 0.93 (95% CI 0.90-0.95), which is slightly higher than that found at a cut-off score of 5; 0.91 (95% CI 0.88-0.93).

At a cut-off score of 4, the I^2 statistic was 85.8%. When 'outliers' (Broekman et al., 2011 and Marc et al. 2008) were excluded the I^2 statistic fell to 14.7%. Again, there was little change in pooled diagnostic data.

Meta-analysis at cut-off scores of 5 and 6 included the highest number of primary studies; 23 and 20 respectively, which increases the accuracy of pooled diagnostic data at these cut-off scores. After a cut-off score of 6, the number of included studies (and so study participants) in meta-analysis at different cut-off scores continuously falls. The number of included studies in meta-analysis rises from 4, at a cut-off score of 2, to 10 at a cut-off score of 4.

At a cut-off score of 6, pooled sensitivity was lower than found at a cut-off score of 5 – 0.79 (95% CI 0.68–0.87) and 0.89 (95% CI 0.80–0.94) respectively. Whereas pooled specificity was higher – 0.83 (95% CI 0.72–0.90) and 0.77 (95% CI 0.65–0.86) respectively. Pooled diagnostic odds ratio dropped from 27.28, at a cut-off score of 5, to 17.61 at a cut-off score of 6. The AUC on the SROC curve at a cut-off score was 0.87 (95% CI 0.84-0.90).

The I^2 statistic was 88.1%. When meta-analysis was re-run excluding 'outliers' (Abas et al., 1998, Broekman et al., 2011, Wongpakaran et al., 2013[1]) pooled sensitivity fell to 0.75 (95% CI 0.63-0.84). Remaining pooled diagnostic data remained relatively the same. The I^2 statistic fell to 0.0%.

When examining pooled diagnostic data and SROC curves it should be noted that 95% confidence intervals do overlap, which suggests difference in findings are non-significant.

Findings of subgroup analyses

This review has found that the diagnostic performance of the GDS-15 is influenced by mean participant age, study setting and country of study. However, such findings are unlikely to be statistically significant because the 95% confidence intervals of subgroup analyses overlap.

Subgroup analysis revealed that diagnostic performance was better for ‘young-old’ participants (i.e. 65 – 74 years of age) compared to ‘middle-old’ participants (i.e. 75 – 84 years of age). When pooled diagnostic data, in accordance with mean age, are compared to pooled diagnostic data of all studies for the GDS-15 it can be observed that diagnostic performance improves when primary studies with a mean participant age ≥75 years are removed from meta-analysis.

Subgroup analysis of study setting has revealed that the diagnostic accuracy of the GDS-15 is lowest in a primary care setting followed by a secondary care setting. The diagnostic performance of the GDS-15 was better in studies based in a community setting; the greatest diagnostic accuracy was found for studies where participants were living independently in the community.

Results of subgroup analysis found pooled sensitivity and specificity were higher for non-Western countries compared to Western countries.

Findings of sensitivity analyses

Sensitivity analyses of risk of bias for the different domains of the QUADAS-II reveal that pooled diagnostic data remains relatively similar to the pooled diagnostic data of meta-analysis of all studies with the exception of the domain of the index test. Meta-analysis excluding studies rated as having a ‘high’ or ‘unclear’ risk of bias for the index test established that pooled sensitivity and pooled diagnostic odds ratio both increased compared to pooled data from meta-analysis of all studies.

Finding of meta-regression

Meta-regression revealed that country of study setting (i.e. non-Western) and language of the GDS-15 (i.e. non-English) were explanatory factors predictive of diagnostic accuracy. A non-Western setting and non-English GDS-15 are clearly linked factors.

Publication bias

Visual examination of the funnel plot is suggestive of the possibility of publication bias because there are fewer studies at the bottom of the funnel plot. Causes of an asymmetrical funnel plot include reporting bias, poor methodological quality, true heterogeneity, artefactual and chance (Sterne et al., 2011).

Limitations

Limitations of this review refer to the primary studies included and also refer directly to the review itself. These will be discussed in further detail in Chapter 4.

Limitations of primary studies – results of this review suggest the presence of selective reporting of cut-off scores (i.e. studies only reporting diagnostic data for a particular cut-off score if it performs well), which causes artificial inflation of diagnostic performance. Limitations of primary studies include concerns regarding methodological quality, particularly concerning risk of bias associated with the use and administration of the index tests. An important issue is that the validity of brief versions of the GDS for use in older adults with cognitive impairment is unknown. For briefer versions of the GDS (i.e. <15-items) there are no standardised items for inclusion, meaning, for example, the GDS-4 in one study was a different instrument to the GDS-4 in another study. It was therefore not possible to perform meta-analyses of briefer versions of the GDS with the exception of the GDS-15, which has standardised items.

Limitations related to the review itself – the protocol of this review was adhered to throughout; however it was not registered, which could introduce a source of bias. The possibility of bias is further introduced by only a single reviewer (Claire Pocklington) performing study selection and data extraction. A lot of time was invested in developing a comprehensive search strategy; however it remains possible that not all relevant studies were found.

CONCLUSION

This review aimed to establish the up-to-date diagnostic accuracy of the GDS-15 and briefer versions of the GDS. Unfortunately, it was not possible to perform meta-analysis for briefer versions of the GDS for two reasons; variations in item composition of briefer versions and reported cut-off scores, which led to an insufficient number of primary studies. Briefer versions of the GDS may offer more clinical appeal due to time restraints in clinical practice; they will take less time to administer in comparison to the GDS-15.

Meta-analysis was possible for the GDS-15. However, it is difficult to draw firm conclusions because our pooled results show evidence of selective reporting of cut-off scores; therefore, our findings should be interpreted cautiously.

CHAPTER 3

The clinical effectiveness of screening for depression in older adults: a systematic review

INTRODUCTION

The diagnostic accuracy of a depression screening tool is important if a screening programme is to be effective. The previous chapter has established the diagnostic accuracy of the GDS, which is a well-known screening tool for use in older adults. The value of screening is of no value, however, if screening does not lead to improved clinical outcomes.

The National Screening Committee guidance criteria for screening describe a need for evidence that a screening programme is more clinically effective than not screening (England, 2013, updated 2015).

Studies exploring the clinical effectiveness of screening for depression in older adults are limited in number. To date only one systematic review exploring the clinical effectiveness of screening for depression in older adults has been performed. That review identified only four primary studies for inclusion and had a number of limitations as discussed in Chapter 1; an exclusion of grey literature, and a search strategy limited to a primary care setting and English language papers only.

This chapter aims to establish if screening for depression in older adults is clinically effective. This chapter will establish up-to-date evidence regarding the clinical effectiveness of screening for depression.

Research question:

What is the clinical effectiveness of screening for depression in older adults?

METHOD

Protocol

A protocol was written in line with CRD guidance, which was registered with PROSPERO (registration number CRD42014010599). See Appendix 6 for the protocol.

Reporting

The Preferred Reporting of Items for Systematic Reivews and Meta-analyses (PRISMA) guidelines were used as a basis for reporting.

Search strategy

A) Search terms

Search terms that referred to the concepts of older adult, depression and screening were used. These three concepts were combined with the Boolean operator 'AND'. Like the first systematic review, search terms for the concept of older adult was comprised of just free text search terms only because subject heading search terms produced too many irrelevant results.

Again, like the first review, a combination of subject heading search terms (i.e. MeSH) and free text search terms were used for the concept of depression to identify all studies. Search terms for the concept of screening was also a combination of MeSH terms (i.e. Mass Screening/ and Diagnosis/) and several free-text terms. Free-text terms referred to case finding, screen, detect, predict, aware, identify and diagnosis. Subject heading search terms were exploded when used.

The syntax of the search strategy was customised to the different electronic databases used. See Appendix 7 for search strategies.

B) Electronic databases

The following electronic databases were searched; MEDLINE, EMBASE, Cumulative Index to Nursing and Allied Health (CLINAHL Plus), Cochrane Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts and Reviews of Effects (DARE) and the Health Technology Assessment (HTA).

The search strategy was performed from inception to May 2014. No limitations regarding language, or publication status or date range were applied.

C) Unpublished and grey literature

Unpublished and grey literature was included in the search strategy to reduce publication bias. The following unpublished and grey literature resources were searched: Conference proceedings via Web of Science, http://ethos.bl.uk, www.guideline.gov and www.opengrey.eu.

D) Additional search strategies

The clinical trials register, www.clinicaltrials.gov, was searched.

The reference list of the existing review by O'Conner et al. and the reference list of each primary study that fulfilled final inclusion criteria were manually checked to identify further eligible studies.

If data were missing or additional information was required authors were contacted.

Citation management

All identified citations were exported into the electronic reference and bibliotherapy managerial software package Endnote. Duplicate citations were deleted using both the command function and manually. The titles and abstracts of all identified citations were screened using Endnote.

Study selection

A) Inclusion and exclusion criteria

PICO criteria were developed for both the first (title and abstract) and second (full paper) stages. First sift criteria were more liberal than second sift criteria to reflect the limited information that may be available in a title and abstract. At the first stage sift if an abstract was not available, the criterion was judged against the title alone. Full papers were obtained of all studies that passed the first sift and these were judged against more detailed second sift criteria. Inclusion and exclusion criteria were piloted first. See Table 34.

As mentioned previously, background reading identified that some studies classify older adults as being 55 years of age or older. In order to not exclude such studies, inclusion criteria regarding age of population was 55 years of age or older.

B) *Study design*

Studies investigating the clinical effectiveness of screening, regardless of condition, are limited in number. A randomized controlled trial is the preferable study design to establish the clinical effectiveness of screening for depression. Ideally, only participants in the intervention group would have undergone screening. In some studies both participants in the intervention and control group have undergone screening, but the screening results of the control group have not influenced their management in anyway.

Hewitt et al. describe method of categorizing screening study designs (Hewitt et al., 2009). The model describes three different evidence levels. All studies, regardless of evidence level, are of a randomized controlled trial design. In level 1 evidence, allocation to study group is performed prior to screening and only the intervention group undergoes screening for the condition in question. The process of screening here is not coupled with care-as-usual or an enhancement of this.

Hewitt et al. describe two categories of level 2 evidence; *a* and *b*. In the level 2a study design, group allocation is performed prior to screening. The intervention group undergo screening and in addition treatment (care-as-usual) or enhanced care) is delivered. Those in the control group receive care-as-usual with no screening. In level 2b evidence, study participants undergo screening prior to study group allocation. The

PICO	Sift	Criteria
Population	First and second	• No restrictions in terms of ethnicity or country • No restrictions in terms of physical comorbidity and cognitive impairment • No restrictions regarding type of study design
Intervention	First	• Citation refers to use of depression screening
	Second	• Depression screening must have been implemented • Method of depression screening specified (i.e. rating scale used) • Depression screening implemented to intervention group only or both intervention and control group with outcome of screening for latter group not disclosed • No restrictions in terms of mode of screening administration • No restrictions in terms of person administrating the screening measure • Management intervention implemented following positive screen result described
Comparator	First	• Control group received care-as-usual (i.e. no enhancement from routine assessment and care administered)
	Second	• If screening occurred in control group, result of screening should not have been disclosed to participant and/or health professionals • If screening occurred in control group, result of screening should not have influenced diagnosis or management or depression
Outcome	First	• Measure of clinical effectiveness referred to
	Second	• Clinical effectiveness referred to symptom improvement – data regarding severity of depression at baseline and post-intervention/care-as-usual • No restrictions in terms of follow-up time period of outcome measure • No restrictions in terms of rating scale used to measure symptom improvement

Table 34: First and second sift inclusion criteria

results of screening for the intervention group are disclosed; whereas no feedback of screening results is disclosed for the control group. Again for the intervention group, screening leads to treatment delivery (care-as-usual or enhanced care). In studies of level 3 evidence all study participants are screened as study eligibility requires a positive screen result. Allocation to study group therefore happens post-screening. Those who screen negative are not included in the study.

Studies of a level 1 evidence present the highest level of methodological quality followed by level 2 then 3. The Hewitt model of classifying study design was adopted to organise the results of this review.

C) Screening of citations

The author (Claire Pocklington) screened the title and abstracts of all citations against inclusion criteria. If a study was identified as eligible, the full-paper was obtained for detailed assessment against second sift inclusion and exclusion criteria. If any uncertainty was encountered following assessment of the full-paper during the second sift of screening citations, this was discussed with a supervisor (Dean McMillan). If disagreements could not be resolved a second supervisor (Simon Gilbody) was involved. If two or more papers were identified with overlapping samples, all were included but reported as one study.

Data extraction

The author (Claire Pocklington) extracted all data and entered it into a standardised proforma. Extracted data were checked by the supervisor Dean McMillan. The following data were extracted:

1) Author, date of publication
2) Sample size
3) Study design
4) Descriptive characteristics of the setting (country, healthcare setting)
5) Descriptive characteristics of the sample (mean age, ethnicity, proportion female, cognitive status)
6) Descriptive characteristics of depression screening method use (tool used, administration mode, administered by whom)

7) Descriptive of clinical management for intervention and control group
8) Descriptive characteristics of outcomes measure (rating scale utilised and time point(s) of follow-up)

Study quality assessment

The Cochrane Collaboration's risk of bias tool was used to assess the methodological quality and risk of bias of all included primary studies. Primary studies were rated as having a 'high', 'unclear, or 'low' risk of bias across seven domains that capture all major causes of bias in randomised trials. The seven domains included 'sequence generation' (how participants were allocated to the intervention and control group); 'allocation concealment' (if it was possible to predict whether a participant was going to be allocated or had already been allocated to a group); 'blinding of participants and personnel' and 'blinding of outcome assessment' (all those actively involved in the study were unaware of group allocations); 'incomplete outcome data' (all data required for all participants were measured); 'selective reporting' (all data measured were reported); and 'other bias'.

Data synthesis

Data synthesis was primarily narrative. The standardised mean effect size (with confidence intervals) was calculated for all primary studies where possible.

Meta-analysis was not possible; however, the planned approach would have been to conduct a random-effect meta-analysis. The computer software programme RevMan 5.3 would have been used for this. Clinical heterogeneity would have been explored first. Statistical heterogeneity would have been assessed using the I^2. If high levels of heterogenetiy had been identified, outliers would have been identified by the visual examination of forest plots, and the analysis re-run excluding those outliers to examine the impact of this heterogeneity.

Subgroup and sensitivity analyses would have also been performed if meta-analysis was possible. Subgroup analysis of study setting (i.e. primary care, secondary care, non-clinical) and sensitivity analysis of different quality assessment criteria were planned. If non-randomised controlled trials were found and if there were a sufficient number of studies a sensitivity analysis would have also have been performed excluding them.

RESULTS

The search strategy identified 16,018 records, which resulted in 9482 post deduplication. The screening of titles and abstracts identified 77 records, which met initial inclusion criteria. Of these 77 records, 8 met second sift criteria and were included (See Figure 8).

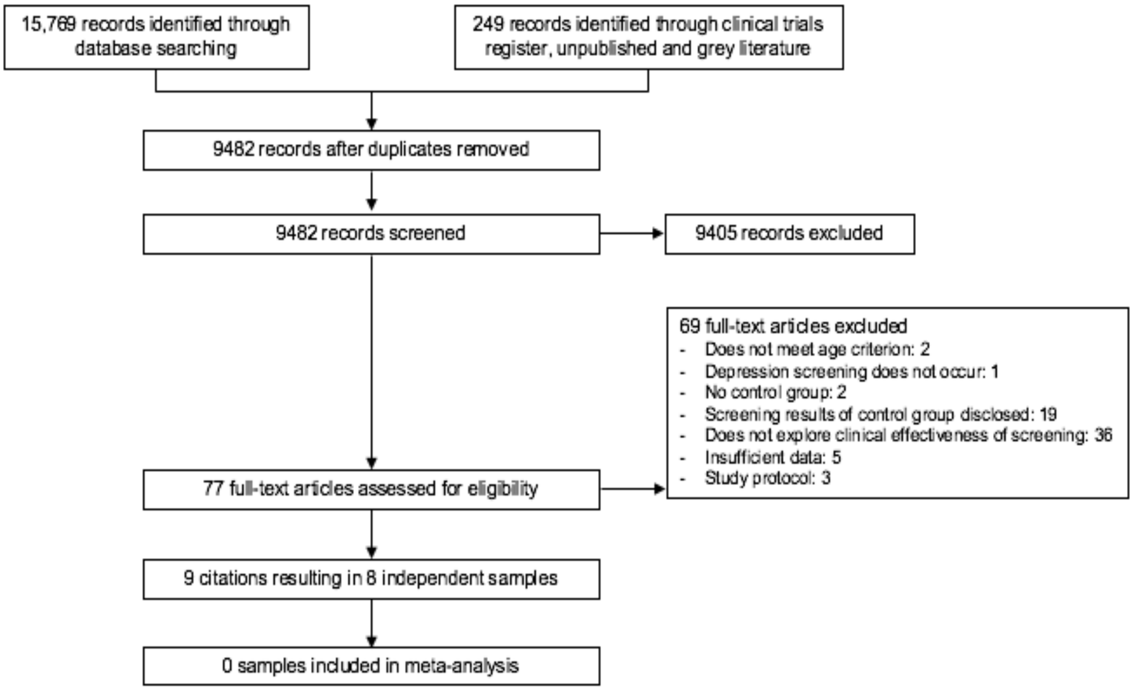
15,769 records identified through database searching
249 records identified through clinical trials register, unpublished and grey literature
9482 records after duplicates removed
9482 records screened
9405 records excluded
69 full-text articles excluded
- Does not meet age criterion: 2
- Depression screening does not occur: 1
- No control group: 2
- Screening results of control group disclosed: 19
- Does not explore clinical effectiveness of screening: 36
- Insufficient data: 5
- Study protocol: 3
77 full-text articles assessed for eligibility
9 citations resulting in 8 independent samples
0 samples included in meta-analysis

Figure 8: PRISMA diagram of study selection

The majority of records were excluded because they did not investigate clinical effectiveness. The second commonest reason for a record being excluded was that the results of screening for the control group were disclosed and therefore subsequent management would have been influenced. See Appendix 8 for a table listing specific reasons for exclusion for each study.

The search strategy identified the protocols for three studies. One of these protocols was for a study that did not meet inclusion criteria (van't Veer-Tazelaar et al., 2006). Data collection is still underway for a study by Imai et al. (Imai et al., 2013); if published in the future, it would likely add to the results of this review. The published data of the study protocol by Gitlin et al. was found and so this study was included in this review.

Two articles report data from one study (Bosmans et al., 2006, van Marwijk et al., 2008). These citations are treated as a single study in the analysis and will be referred to as van Marwijk et al, 2008.

The authors of six studies were contacted for more information by email. Outcome data was presented in graphical format for the studies by Callahan et al. and Joubert et al. and therefore the authors were contacted to request numerical data. The primary author of the Callahan et al. paper responded stating that he no longer had the original data due to the study being conducted in 1994. One of the authors of the Joubert et al. study replied stating that they did not have the original data but stated no reason why. The primary author of the Shah et al. study was contacted as findings were reported as 'non-significant' but no p-values were documented. No response was received. Authors of the studies by Gitlin et al., van Marwijk et al. and Whooley et al. were contacted by email as additional data was needed to adjust for the effect of cluster randomization; none of the authors replied however.

The eight studies amount to 812 participants; 448 in the intervention group and 464 in the control group. See Table 35 for study and sample characteristics.

Study	Sample characteristics	Study design/ Hewitt level	Screening details	Care received by intervention group	Care received by control group	Measured outcome	Follow-up points
Baldwin et al. (2004)	Country: UK Setting: acute medical wards, secondary care Age (yrs): Av. = 80.3 Female: 64.1% Intervention: 54, CAU: 60	Randomised controlled trial Hewitt level 2b	Scored ≥11 on the GDS and ≥180 on the OMC test	'Multi-facet nurse-led intervention', consisting of assessment, management (antidepressants and psychotherapy interventions) and liaison support (included patient education). Lasted for maximum of 6 weeks	'Usual care' – defined as care and treatment delivered by the acute ward staff, which could include antidepressant use and referral to psychiatry team/ psychiatrist	GDS score	6-8 weeks
Callahan et al. (1994)	Country: USA Setting: primary care Age (yrs): Av. = 68 Female: 68.8% Intervention: 100,CAU: 75	Randomised controlled trial Hewitt level 2b	Scored ≥16 on the CES-D and ≥15 on the HAM-D	Doctors given educational material and patient specific treatment recommendations. Two additional appointments with doctor given. Seen for a period of three-months.	Treatment at discretion of primary care doctor	HAM-D score	1, 3, 6 and 9 months
Cullum et al. (2007)	Country: UK Setting: inpatient ward, secondary care Age (yrs): Av. = 79.9 Female: 58.8% Intervention: 41, CAU: 45	Randomised controlled trial Hewitt level 2b	Scored ≥7 on the GDS-15	Liaison psychiatric nurse formulated management plan; addressed psychological and social needs. Need for antidepressant assessed. Regular 2-3 week follow-up and monitoring for a total of 12-weeks.	'Usual care' – if medical team diagnosed, management could include antidepressants, referral to MHS or to primary care for assessment/monitoring	GDS-15 score	12 weeks
Gitlin et al. (2013)	Country: USA Setting: community Age (yrs) Av. = 69.9 Female: 78.4% Intervention:106,CAU:102	Cluster randomised controlled trial Hewitt level 2b	Score ≥5 on PHQ-9 when performed twice over a two week interval	Multifactorial intervention involving unmet care needs assessment, patient education, stress reduction, behavioural activation. Improved access to social & medical services	Wait-list control – received intervention after 4-months	PHQ-9 score CES-D score	4 months

CAU – care-as-usual. CES-D - center for epidemiological studies depression scale. GDS - geriatric depression scale. HAM-D - Hamilton rating scale for depression. OMC test – 66 tem orientation-memory-concentration test. PHQ-9 – patient health questionnaire.

Table 35: Descriptive table of included studies

Study	Sample characteristics	Study design/ Hewitt level	Screening details	Care received by intervention group	Care received by control group	Measured outcome	Follow-up points
Joubert et al. (2013)	Country: Australia Setting: A&E, secondary care Age (yrs): Av. = 76.6 Female: 71.0% Intervention: 4, CAU: 4	Randomised controlled trial Hewitt level 3	Scored ≥2 on the GDS-4, then ≥5 on the GDS-15	'Comprehensive, integrated management plan' in line with national guidelines. Included patient education, problem-solving and counseling. One-off session delivered by a social worker. Primary care doctors informed and provided ongoing monitoring.	'Usual care' – not described	GDS-15 score	6 weeks
Shah et al. (2001)	Country: UK Setting: inpatient ward, secondary care Age (yrs): Av. = 85.0 Female: 57.0% Intervention: 14, CAU: 17	Randomised controlled trial Hewitt level 2b	Scored ≥11 on the GDS and ≥7 on the BAS-DEP	One-off psychogeriatric consultation, consisting of full-diagnostic work-up and management.	Not described	BAS-DEP, GDS and MADRS scores	10 weeks and 1 year
Van Marwijk et al. (2008)	Country: Netherlands Setting: primary care Age (yrs): Av. = 65.6 Female: 83.0% Intervention: 56, CAU: 67	Cluster randomised controlled trial Hewitt level 2b	Scored ≥5 on the GDS-15	Patient education, supportive counselling and pharmacological management in accordance with national guidelines (i.e. Paroxetine). Reviewed 2-weekly for initial 2-months then monthly for 4-months	'Usual care' – represented current actual practice	MADRS score and PRIME-MD score	2, 6 and 12 months
Whooley et al. (2000)	County: USA Setting: primary care Age (yrs): Av. = 75.8 Female: 60.5% Intervention: 97, CAU: 109	Cluster randomised controlled trial Hewitt level 2a	Scored ≥6 on the GDS-15	Primary care doctors given educational sessions about treatment & management instruction sheet specific to participant. Participants attended 6 wkly educational sessions (covered explaining, treatment options, coping mechanisms and prevention).	'Care as usual' – not described	GDS-15 score	2 years

BAS-DEP - brief assessment scale for depression. CAU – care-as-usual. GDS - geriatric depression scale. MADRS - Montgomery and Asberg depression rating scale. PRIME-MD - primary care evaluation of mental disorders.

Table 35: Descriptive table of included studies cont.

Overview of studies

Country:

Three studies were based in the UK (Baldwin et al., 2004, Cullum et al., 2007, Shah et al., 2001), three in the USA (Callahan et al., 1994, Gitlin et al., 2013, Whooley et al., 2000), one in the Netherlands (Bosmans et al., 2006, van Marwijk et al., 2008) and one in Australia (Joubert et al., 2013).

Setting:

Three studies were undertaken in primary care and four in secondary care. Of the studies based in secondary care, three were in inpatient wards; whereas one was in an emergency department. One study had a community setting, which was based in participants' homes and local senior centres (i.e. community centres specifically for older adults).

All of the studies set in the UK were based in secondary care.

Age:

The mean age range of the samples in the studies ranged from 65.5 – 85.0 years.
The secondary care studies had somewhat older aged samples than the primary care and community setting studies; the mean age of secondary care studies ranged from 76.6 – 85.0 years, whereas the mean age of primary care and community based setting studies ranged from 65.6 – 75.8 years.

Female:

For all eight studies, there were more females than males in the samples. The study by Shah et al. had the sample with the lowest proportion of female participants at 57.0%. The study with the highest proportion of female participants (83.0%) was by Van Marwijk et al. No relation between proportion of female participants in the sample and study setting or mean age was identified.

Ethnicity:

The studies based in the USA (Callahan et al., 1994, Gitlin et al., 2013, Whooley et al., 2000), describe the ethnicity of study participants. All participants were African-American in the Gitlin study and just over half of the sample were 'black' in the study by Callahan et al. The study by Whooley et al. describes the samples ethnicity in more detail; about a third were African-American, about a quarter were described as 'other' and the remainder were described as white.

Ethnicity is not described in the remaining five studies.

Cognitive status:

The studies by Shah et al. and Whooley et al. did not assess cognitive status; whereas in the studies by Joubert et al. and Van Marwijk et al. 'profound cognitive impairment' was an exclusion criterion. Van Marwijk et al. assessed cognition using the standardised mini-mental state examination (SMMSE) (Folstein et al., 1975); mean SMMSE was 25.6 for the intervention group and 26.6 for the control group.

The study by Baldwin et al. also assessed cognition using the SMMSE (Folstein et al., 1975). The mean SMMSE for the intervention and control group were 18.2 (SD 6.4) and 18.8 (SD 6.9) respectively. Participants were excluded from the study if they scored ≥10 on 6-item Orientation-Memory-Concentration (OMC) test (Katzman et al., 1983).

Callahan et al. assessed cognitive status using the short portable mental status questionnaire (SPMSQ) (Pfeiffer, 1975); making three or more errors suggested the presence of cognitive impairment. In accordance with the SPMSQ, 12.0% of the intervention group and 16.2% of the control group were found to have cognitive impairment.

Cognitive status was assessed using the abbreviated mental test score (AMTS) (Hodkinson, 1972) by Cullum et el. Cognitive status of the intervention and control groups were described as a percentage of participants scoring eight or more on the AMTS; this figure was 76% for the intervention group and 80% for the control group.

Sample size:

The sample size varied between the eight studies. The mean sample size was 119 (range 8 – 208). Gitlin et al. and Whooley et al. had sample sizes >200. The majority of studies ranged in sample size from 96 – 136 (Baldwin et al., 2004, Callahan et al., 1994, Cullum et al., 2007, Van Marwijk et al., 2008).

Screening process and depression rating scale used:

In total, eight different depression screening instruments were used. Three of the studies used more than one depression screening instrument (Callahan et al., 1994, Joubert et al., 2013, Shah et al., 2001). All but two studies (Callahan et al., 1994, Gitlin et al., 2013) administered a version of the Geriatric Depression Scale (GDS) to screen for depression. Two studies used the original 30-item GDS (Baldwin et al., 2004, Shah et al., 2001). The study by Shah et al. also administered the brief assessment scale for depression (BAS-DEP). Four studies used the 15-item version of the GDS (GDS-15) (Cullum et al., 2007, Joubert et al., 2013, Van Marwijk et al., 2008, Whooley et al., 2000). One study used an ultra-brief version of the GDS, the GDS-4 (Joubert et al., 2013).

The studies by Callahan et al. and Gitlin et al. were the only studies which did not administer a version of the GDS; instead Callahan et al. used the centre for epidemiological studies depression scale (CES-D) and the Hamilton rating scale for depression (HAM-D). Gitlin et al. used the patient-health-questionnaire (PHQ-9).

For the three studies that used more than one depression screening instrument there were two stages of screening; for participants to be included in the study they had to be identified as having a positive result on the first screening instrument, which meant they were eligible to be screened again using a second instrument; the participant must have had a positive result on the second instrument to have been included in the study.

Outcome measures and follow-up points:

The depression rating scales used as outcome measures varied between the studies. Seven studies used the same depression rating scale for screening purposes and to measure outcome (i.e. symptom improvement). The study by Shah et al. used three separate depression rating scales to measure outcome; the MADRS was used in addition to the GDS and BAS-DEP, which were used for screening purposes.

The study by Van Marwijk et al. measured outcome by the MADRS and PRIME-MD; neither of these were used for screening purposes. The GDS-15, which was used for screening, was used as a secondary outcome measure. It should be noted that the PRIME-MD is a diagnostic tool.

Follow-up time points for outcome measures varied between studies. The number of follow-up points also varied by study. Four studies (Baldwin et al., 2004, Cullum et al, 2007, Joubert et al. 2013, Whooley et al., 2000) measured outcome at one follow-up point only, which ranged from 6 weeks to 2 years. The remaining studies reported outcome measures at multiple follow-up points; for example, the study by Callahan et al. measured outcome at 1, 3, 6 and 9 months.

The study by Gitlin et al. utilised a cross-over design; after 4-months, participants in the control group received the intervention and outcomes were measured again at 8-months. Outcome measures at 8-months have not been included in this review.

Included studies reported and presented data differently; for example, some studies reported outcome measures as mean change in score from baseline; whereas other studies reported actual mean scores at follow-up points. The study by Callahan et al. presented mean HAM-D score for the four follow-up points in graphical format and so it was difficult to extract accurate figures. The authors were contacted and asked to provide these data but it is no longer available.

Study design:

Of the studies identified, five were randomized controlled trials (Baldwin et al., 2004, Callahan et al., 1994, Cullum et al., 2007, Joubert et al., 2013, Shah et al., 2001,) with the remainder being cluster-randomised controlled trials. The level of randomization in two cluster randomised controlled trials was at the level of the primary care practice or clinic (Van

Marwijk et al., 2008, Whooley et al., 2000). In the remaining cluster randomised controlled trial, individuals were first stratified in accordance with recruitment source (i.e. home or senior centre) and then block randomization occurred into the intervention or control groups (Gitlin et al., 2013).

Hewitt level of study design:

The majority of studies were classed as being of level 2b evidence. Table 36 shows the Hewitt evidence level rating of the eight studies included studies.

Evidence level rating		Study
1		none
2	a	Whooley et al. , 2000
	b	Baldwin et al., 2004 Callahan et al., 1994 Cullum et al., 2997 Shah et al., 2001 Van Marwijk et al., 2008
3		Joubert et al., 2013

Table 36: Hewitt evidence level rating of included studies

Quality assessment

The methodological quality of all eight primary studies was assessed using the Cochrane Collaboration's risk of bias tool. See Table 37. No studies were rated as having a high risk of bias for patient selection. The process of randomisation was not described in three studies and therefore they were rated as having an unclear risk of bias (Callahan et al., 1994, Joubert et al., 2013, Shah et al., 2001). Only two studies were rated as having a low risk of bias for 'allocation concealment'.

Blinding of study participants and personnel was associated with the highest risk of bias; participants were aware of which study group the participant had been placed in. Non-blinding of participants may have introduced performance bias. However, this is inevitable with psychological trials as it is typically impossible to achieve blinding of participants in such trials.

Five studies were rated as having an unclear risk of bias for 'blinding of outcome assessment' because it is unclear who administered outcome measures and whether this was conducted blind. Non-blinding of outcome assessment may have introduced a detection bias.

Ratings for 'incomplete reporting of outcome data' were variable. Two studies did not state how many participants had dropped out of the study so risk of bias was rated as high (Callahan et al., 1994, Joubert et al., 2013). Two studies describe drop-out rate overall and not specific to the intervention and control groups (Gitlin et al., 2013, Shah et al., 2001).

All studies were rated as having an unclear risk of bias for selecting reporting of results with the exception of the study by Gitlin et al., where the study protocol was available. Risk of 'other bias' was rated 'high' for all eight studies. With the exception of the studies by Van Marwijk et al. and Whooley et al., the remaining six studies commented that they were unpowered due to small sample sizes. In addition, the authors of the Baldwin study did not think they could exclude a 'cross-over effect'. Cluster randomisation in the study by Van Marwijk et al. did not ensure that sample characteristics were equally distributed; the authors specifically commented on the high proportion of female participants. Authors of the Whooley et al. study commented that the follow-up point would not have capturing any early benefits from screening. The authors felt that findings may not be generalizable due to the exclusion of people with physical disabilities or those who did not speak English participating in the study.

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants and personnel (performance bias)	Blinding of outcome assessment (detection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Other bias
Baldwin et al. (2004)	🟩	🟧	🟥	🟧	🟩	🟧	🟥
Callahan et al. (1994)	🟧	🟧	🟥	🟧	🟥	🟧	🟥
Cullum et al. (2007)	🟩	🟩	🟥	🟧	🟩	🟧	🟥
Gitlin et al. (2013)	🟩	🟩	🟥	🟩	🟧	🟩	🟥
Joubert et al. (2013)	🟧	🟧	🟥	🟧	🟥	🟧	🟥
Shah et al. (2001)	🟧	🟧	🟥	🟩	🟧	🟧	🟥
Van Marwijk et al. (2008)	🟩	🟧	🟥	🟩	🟩	🟧	🟥
Whooley et al. (2000)	🟩	🟧	🟥	🟧	🟩	🟧	🟥

Green = low | Red = high | Orange = unclear

Table 37: Quality assessment results for Cochrane Collaborations risk of bias tool

Narrative analysis

There is variation in how the primary studies report outcome data; the studies by Cullum et al. and Shah et al. present outcome data as change in score from baseline at follow-up point and therefore effect size was not calculated.

It has only been possible to calculate effect size for data presented by Baldwin et al., which presents actual outcome measures at follow-up points.

Outcome measures for Callahan et al. and Joubert et al. are presented in graphical format and without sufficient additional data to calculate effect sizes.

There is insufficient data to calculate effect size for the studies by Gitlin et al., Van Marwijk et al. and Whooley et al. as the additional data needed to adjust for the effect of cluster randomisation are not reported. See Table 38 for reported data.

Findings by Hewitt et al. level of study design

Outcome of studies classed as Hewitt level 1 model of study design

No studies were classified as having a level 1 model of study design. As discussed, this would have been the preferred type of study design.

Outcome of studies classed as Hewitt level 2a model of study design

Whooley et al. was the only studied identified as having a Hewitt level 2a model of study design. Though the mean change in GDS-15 score at the single follow-up point of 24-months was greater for the intervention group compared to the control group, see Table 38, the results of this study do not provide any evidence that screening for depression is clinically effective because findings are not statistically significant.

Outcome of studies classed as Hewitt level 2b model of study design

- Findings at a ≤4-month follow-up period

Six studies of level 2b evidence report outcome data at a follow-up of ≤4 months (Baldwin et al., 2004, Callahan et al., 1994, Cullum et al., 2007, Gitlin et al., 2013, Shah et al., 2001, Van Marwijk et al., 2008). See Table 38. Only two studies provide evidence that screening for depression in clinically effective however. An effect size of -0.28 (95% CI -0.65 –0.09) was established in favour of screening for the study by Baldwin et al. who reported outcome scores at 6-8 weeks using the GDS. Results reported by the primary study were statistically significant (*p* value = 0.04). See Table 38.

Screening was found to be clinically effective in the study by Gitlin et al. who reported outcome measures, using both the PHQ-9 and CES-D, at a four-month follow-up period.
Mean difference between the intervention and control group was 3.0 for the PHQ-9 and 3.5 for the CES-D. Differences between the intervention and control group were found to be statistically significant for both outcome measures (*p* = <0.001). See Table 38.
Thus the study provides evidence in support of screening for depression in older adults.

Reported data in the studies by Cullum et al. and Van Marwijk et al. show that outcome measures improved more for the intervention group, however observed differences were not statistically significant. See Table 38.

The studies by Callahan et al. and Shah et al. show outcome measures to be the same or better for the control group in comparison to the intervention group. The study by Callahan found no difference between HAM-D scores in the intervention and control group at the one or three-month follow-up point. Findings, however, were not statistically significant.

The findings of the Shah et al. study, which administered three different depression rating scales to measures outcome, are conflicting. For example, BAS-DEP and MADRS outcome scores show no difference between the intervention and control groups, however, GDS outcome score shows greater improvement in the intervention group. The findings of the Shah et al. study are not statistically significant for any of three outcome measures however. See Table 38.

- Findings at a 6-month follow-up period

The studies by Callahan et al. and Van Marwijk et al. report outcome data at a 6-month follow-up period. For the study by Van Marwijk et al. two outcome measures are utilised; MADRS and PRIME-MD score. Reported MADRS outcomes for the intervention and control groups show that screening is clinically effective; the difference in MARDS scores were found to be statistically significant (p = <0.05). Reported outcome in terms of PRIME-MD also show a greater reduction in score for the intervention group compared to the control group, however, such a difference was found not to be statistically significant. See Table 38.

Reported outcome for study by Callahan et al. provide no evidence that screening for depression in clinically effective; no difference in HAM-D score was observed between the intervention and control group. See Table 38.

Hewitt level	Follow-up period	Study	Follow-up point	Outcome measure	Outcome
2a	24 months	Whooley et al. (2000)	24 months	GDS-15	Intervention group mean change score: -2.4 (SD ±3.7) Control group mean change score: -2.1 (SD ±3.6) *p* value = 0.50
2b	≤4 months	Baldwin et al. (2004)	6-8 weeks	GDS	Mean difference between intervention and control group: -2 (95% CI -4.0 – -0.1) *p* value = 0.04 Calculated effect size -0.28 (-0.65 – 0.09)
		Callahan et al. (1994)	1 month	HAM-D	Mean score lower in control group *Difference not statistically significant*
			3 months		No difference in mean score between intervention and control group
		Cullum et al. (2004)	12 weeks	GDS-15	Intervention group mean change score: 4.6 (SD 3.86) Control group mean change score: 3.6 (SD 3.61) *Difference not statistically significant*
		Shah et al. (2001)	10 weeks	BAS-DEP	Intervention group median score and range: -5 (-12 – 3) Control group mean score and range: -5 (-13 – 2) *Difference not statistically significant*
				GDS	Intervention group median score and range: -4 (-14 – 8) Control group mean score and range: -3 (-7 – 6) *Difference not statistically significant*
				MADRS	Intervention group median score and range: -12 (-29 – -1) Control group mean score and range: -12 (-21 – 6) *Difference not statistically significant*

Table 38: Outcome data for studies by Hewitt evidence level of study design and by outcome measure follow-up time periods

Hewitt level	Follow-up period	Study	Follow-up point	Outcome measure	Outcome
2b	≤4 months	Van Marwijk et al. (2008)	2 months	MADRS	Intervention group mean score: 19.56 (SE ±3.32) (Baseline: 21.66 (SE ±2.86)) Control group mean score: 19.58 (SE ±3.49) (Baseline: 20.94 (SE ±2.48)) *Difference not statistically significant*
		Gitlin et al. (2013)	4 months	PHQ-9	Mean difference between intervention and control group: -3.0 (95% CI -4.7 – -1.4) *p* value = <0.001
		Gitlin et al. (2013)	4 months	CES-D	Mean difference between intervention and control group: -3.5 (95% CI -5.1 – -1.9) *p* value <0.001
	6 months	Callahan et al. (1994)	6 months	HAM-D	No difference in mean score between intervention and control group
		Van Marwijk et al. (2008)	6 months	MADRS	Intervention group mean score: 9.23 (SE ±2.84) Control group mean score: 11.45 (SE ±2.52) *p* value = <0.05
				PRIME-MD	Intervention group mean score: 2.80 (SE ±1.04) (Baseline: 6.10 (SE ±0.80)) Control group mean score: 3.99 (SE ±2.52) (Baseline: 6.33 (SE ±1.01)) *Difference not statistically significant*

Table 38: Outcome data for studies by Hewitt evidence level of study design and by outcome measure follow-up time periods cont.

Hewitt level	Follow-up period	Study	Follow-up point	Outcome measure	Outcome
2b	9 - 12 months	Callahan et al. (1994)	9 months	HAM-D	No difference in mean score between intervention and control group
		Shah et al. (2001)	12 months	BAS-DEP	Intervention group median score and range: -3 (-11 – 8) Control group mean score and range: -5 (-13 – 1) *Difference not statistically significant*
				GDS	Intervention group median score and range: -5 (-18– 1) Control group mean score and range: -1 (-9 – 3) *Difference not statistically significant*
				MADRS	Intervention group median score and range: -15 (-23–7) Control group mean score and range: -11 (-16 – -5) *Difference not statistically significant*
		Van Marwijk et al. (2008)	12 months	MADRS	Intervention group mean score: 10.80 (SE ±2.85) Control group mean score: 10.09 (SE ±2.50) *Difference not statistically significant*
				PRIME-MD	Intervention group mean score: 3.23 (SE ±1.04) Control group mean score: 3.74 (SE ±1.21) *Difference not statistically significant*
3	≤3 months	Joubert et al. (2013)	6 weeks	GDS-15	Intervention group mean score: 8.25 (Baseline: 7.75 (SD 3.30)) Control group mean score: 5.0 (Baseline: 5.50 (SD 2.65)) *Statistical significance of results unknown*

Table 38: Outcome data for studies by Hewitt evidence level of study design and by outcome measure follow-up time periods cont.

- Findings at a 9 – 12-month follow-up period

Three studies report outcome data at this follow-up period. See Table 38. The studies provide no evidence that screening for depression is clinically effective. Outcome reported by Callahan et al., like findings at the previous follow-up period, show no difference in HAM-D scores between the intervention and control groups.

For the study by Shah et al., greater symptom improvement was observed for GDS and MADRS scores for the intervention group compared to the control group. Outcome in accordance to BAS-DEP scores showed the opposite. However, findings were not statistically significant for any outcome measure used.

No statistical significant differences were observed in MADRS and PRIME-MD outcome scores for the intervention and control group for the study by Van Marwijk et al.

Outcome of studies classed as Hewitt level 3 model of study design

Joubert et al. was the only study that was classified as having a Hewitt level 3 model of study design. Findings do not support screening; symptoms were found to have worsened in the intervention group, whereas symptoms in the control group were found to have improved. It is not reported if the difference between outcome measure in the intervention and control groups were statistically significant, however. See Table 38.

Associations between screening process and outcome findings

The process of screening in the two studies by Callahan et al. and Joubert et al. was two-staged. The results of both of these studies do not support screening as being clinically effective.

Studies where the depression screening process involved just one stage – Cullum et al., Van Marwijk et al. and Whooley et al. – were associated with non-statistical significant findings where screening was found to be clinically effective.

Meta-analysis

As discussed, it was not possible to perform meta-analysis for outcome measures by Hewitt level of study design or by follow-up period due to an insufficient number of primary studies and insufficient data provided by primary studies.

Subgroup and sensitivity analyses

Meta-analyses for subgroup and sensitivity analyses were not possible due to an insufficient number of primary studies, data being no longer available (Carrahan et al., 1994) and missing data pertaining to studies that utilised a cluster randomised controlled design (Van Marwijk et al., 2008, Whooley et al., 2000).

Publication bias

It was not possible to explore for the presence of publication bias statistically due to insufficient data and an insufficient number of primary studies; at least 10 studies are required to undertake a funnel plot.

DISCUSSION

This aim of this systematic review was to establish the clinical effectiveness of screening for depression in older adults.

Findings of narrative analysis

Of the eight identified primary studies which explore the clinical effectiveness of screening for depression in older adults, only three studies have established significant findings in favour of the intervention group; Baldwin et al. (*p* value = 0.04), Gitlin et al. (*p* value = <0.001) and Van Marwijk et al. (*p* value = <0.05). Therefore, there is limited evidence to support screening for depression being clinically effective.

The three studies by Baldwin et al., Gitlin et al. and Van Marwijk et al. are all classed as a Hewitt level 2b of study design and they all utilised only one stage in the process of screening. Studies that used a two-stage screening process found screening to not be clinically effective.

The three studies that found statistically significant results in favour of screening varied in the outcome measures used (i.e. depression rating scale) and follow-up time periods that outcomes were measured; Baldwin et al. reported outcome measures at 6-8 weeks using the GDS, whereas Gitlin et al., reported outcome measures at 4-months using both the PHQ-9 and the CES-D. Van Marwijk et al. reported outcome measures using both the MADRS and PRIME-MD at three different follow-up points (≤4-months, 6-months and 12-months), however statistical significant findings in favour of screening were only found for the outcome measured by the MADRS at 6-months.

Limitations

As with the review in the previous chapter, limitations in this review apply to the included individual primary studies and the review itself.

***Limitations to primary studies*:** Clinical heterogeneity is high amongst the primary studies, particularly regarding the treatment received by the intervention groups as this varied considerably. The search strategy implemented and the depression rating scale or diagnostic tool used for outcome measure varied amongst the studies are other sources of clinical heterogeneity. It is unclear if all participants in the intervention groups actually received all aspects of treatment they were supposed to have had. Sample size and risk of bias also varied – risk of bias for several domains of the Cochrane risk of bias tool was unclear. Follow-up points of reported outcome measures varied between the studies; some outcome measures may have been reported too soon and some reported too late, which may have lead to the under-estimations of the clinical effectiveness of screening for depression.

Several studies commented that the sample size was too small meaning studies may have been statistically underpowered. In the three studies based in the USA there was high proportion of participants of an African-American/black ethnicity, which may affect how results can be generalized to the UK population.

Limitations with the review itself: There are two major limitations of this review; a) it has not been possible to calculate effect sizes for all primary studies because required insufficient data are not available; and b) meta-analysis has not been possible, again due to limited data. As meta-analysis has not been performed it is not possible to comment on statistical heterogeneity, in terms of overlapping confidence intervals or the calculated I_2 statistic.

Bias may have been introduced by a single reviewer, the author Claire Pocklington, performing study selection and data extraction independently. Ideally, more studies of a higher level of Hewitt study design would have been preferable (i.e. level 1 or 2ab).

Limitations will be discussed in more detail and explored further in Chapter 4.

CONCLUSION

This review aimed to establish the clinical effectiveness of screening for depression in older adults. Owing to insufficient data being available and an insufficient number of primary studies meta-analysis was not possible; therefore, the analysis has been narrative in nature.

This review has found there is limited evidence to support screening for depression in older adults being clinically effective.

CHAPTER 4

Discussion

INTRODUCTION

The aim of this dissertation was to explore the topic of screening for depression in older adults. The importance of this clinical topic has been described in Chapter 1. Depression is under-diagnosed in an older adult population, the reasons for which are multifactorial, and is associated with higher rates of morbidity and mortality, increased healthcare utilisation and increased economic costs compared to a younger population. A screening programme for depression in older adults could address diagnostic difficulties and adverse associations.

The first objective of this dissertation was to establish the diagnostic accuracy of brief versions of the GDS because they would be an obvious choice of tool to utilise in a screening programme. The GDS was designed specifically for use in older adults and is a widely known screening tool for depression with many derived briefer versions now in existence. As discussed in Chapter 1, existing evidence from the previous five systematic reviews is out-dated and these reviews had a number of methodological limitations: a focus on the original 30-item version rather than briefer versions that offer more clinical appeal, the presence of publication bias due to search strategies not including unpublished data, and an absence of any form of quality assessment in all but one review.

The second objective was to explore whether a screening programme would be justified by establishing if screening for depression is associated with better clinical outcomes. There is a lack of empirical evidence of the clinical effectiveness of screening for depression. The one existing review is out of date and has methodological limitations; a search strategy limited to a primary care setting, English language only and no inclusion of grey literature.

These objectives of this dissertation were achieved by conducting two separate systematic reviews, the results of which have been presented in Chapter 2 and 3 respectively.

Summary of results

The diagnostic accuracy of briefer versions of the GDS

Six briefer versions of the GDS were identified in addition to the GDS-15 in the first review. The six briefer versions of the GDS included the GDS-1, GDS-4, GDS-5, GDS-7, GDS-8 and GDS-10. Meta-analysis was not possible due to there being an insufficient number of studies; no set standardised items for briefer versions of the GDS contributed to this. Meta-analysis of the GDS-15 established a pooled sensitivity of 0.89 (95% CI 0.80-0.94) and a pooled specificity of 0.77 (95% CI 0.65-0.86) at the recommended cut-off score of 5. When 'outliers' were removed, the I^2 fell from 76.7% to 8.1%, statistic pooled sensitivity increased slightly whereas pooled specificity fell slightly; 0.90 (95% CI 0.81-0.95) and 0.75 (95% CI 0.60-0.86) respectively.

The clinical effectiveness of screening for depression in older adults

Of eight primary studies identified, three studies provide evidence that screening for depression in older adults in clinically effective. These studies found a statistically significant difference in reported outcome measures in favour of the intervention group. Some of the remaining studies show better outcome measures for the intervention group but findings are not statistically significant. Unfortunately, meta-analysis of primary studies was not possible.

STRENGTHS AND LIMITATIONS

The methodological quality of the two systematic reviews undertaken in this review and the methodological quality of all included primary studies will influence the interpretation of results and what conclusions can be drawn. Reflection and appraisal of the strengths and limitations of the two systematic reviews performed will enable appropriate interpretation of results.

Strengths

AMSTAR criteria

The methodological quality of a review should form the basis of whether results can be used to guide clinical practice (Sharif et al., 2013). The methodological quality of the systematic reviews conducted in this dissertation were assessed against A Measurement Tool to Assess Systematic Reviews (AMSTAR) criteria. AMSTAR facilitates conclusions to be made on how reliable results from systematic reviews are. AMSTAR criteria consist of 11 items, which leads to an overall summary score out of 11 being calculated; if an item is rated as 'yes' a score of 1 is given but if an item is rated as 'no' or 'can't answer' a score of 0 is given. If an item is 'not applicable' it is not included in the summary score. The following thresholds of summary score have been proposed; 8 – 11 is high quality, 4 – 7 is medium quality and 0 – 3 is low quality.

AMSTAR assessment of the review of the diagnostic accuracy of brief versions of the GDS

This review has a summary score of 8/11, indicating that the review is high quality. See Table 39 for assessment results. The first criteria the AMSTAR - Was 'a priori' design provided? – had to be rated as 'can't answer', in line with operationalisation criteria, because though the protocol is available it was not registered (see Appendix 2). As mentioned previously, the protocol was adhered to throughout the undertaking of the review.

Study selection and data-extraction were not duplicated; study selection and data-extraction were performed by the author (Claire Pocklington) alone. Any uncertainty was discussed with a supervisor (Dean McMillan) but if disagreement was encountered a further supervisor (Simon Gilbody) was involved. There were no 'conflict of interest' for this review and this is reported in Chapter 2. As 'conflict of interest' was not reported for included primary studies, the final AMSTAR criteria is rated as 'no'.

AMSTAR criteria	Rating			
	Yes	No	Can't answer	Not applicable
1. Was 'a priori' design provided?			✘	
2. Was there duplicate study selection and data-extraction?		✘		
3. Was a comprehensive literature search performed?	✘			
4. Was the status of publication (i.e. grey literature) used as an inclusion criteria?	✘			
5. Was a list of studies (included and excluded) provided?	✘			
6. Were the characteristics of the included studies provided?	✘			
7. Was the scientific quality of the included studies assessed and documented?	✘			
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	✘			
9. Were the methods used to combine the findings of studies appropriate?	✘			
10. Was the likelihood of publication bias assessed?	✘			
11. Was the conflict of interest included?		✘		
Summary score	8/11			

Table 39: AMSTAR assessment of review of the diagnostic accuracy of brief versions of the GDS

AMSTAR assessment of the review of the clinical effectiveness of screening for depression in older adults

The summary score of this review is 8/10, which indicates the review is high quality. See Table 40 for assessment results. The protocol for this review was registered with PROSPERO (registration number CRD42014010599). See Appendix 6 for the protocol. As with the other review, study selection and data-extraction were performed by the author (Claire Pocklington) only, therefore there was no duplication. The impact of this is discussed under 'Limitations'.

Publication bias was not assessed as analysis was narrative in nature. Again, 'conflict of interest' was not assessed in the included primary studies. There is no 'conflict of interest' for the review itself as reported in Chapter 3.

AMSTAR criteria	Rating			
	Yes	No	Can't answer	Not applicable
1. Was 'a priori' design provided?	✘			
2. Was there duplicate study selection and data-extraction?		✘		
3. Was a comprehensive literature search performed?	✘			
4. Was the status of publication (i.e. grey literature) used as an inclusion criteria?	✘			
5. Was a list of studies (included and excluded) provided?	✘			
6. Were the characteristics of the included studies provided?	✘			
7. Was the scientific quality of the included studies assessed and documented?	✘			
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	✘			
9. Were the methods used to combine the findings of studies appropriate?	✘			
10. Was the likelihood of publication bias assessed?				✘
11. Was the conflict of interest included?		✘		
Summary score	8/10			

Table 40: AMSTAR assessment of review of the clinical effectiveness of screening for depression

Quality assessment of primary studies

For the review of diagnostic accuracy of brief versions of the GDS

a) QUADAS-II results

The primary studies had a number of methodological issues as shown by the results of the QUADAS-II (see Table 12 and 13). Sensitivity analyses of QUADAS-II results for the GDS-15 has found that risk of bias associated with 'patient selection' was found not to influence pooled diagnostic data. Similar findings were found for QUADAS-II domain of 'flow/timing' of study design; pooled diagnostic data for a cut-off score of 4 – 6 remained relatively unchanged, however diagnostic performance fell considerably at a cut-off score of 7 and 8.

Bias regarding ‘reference test’ was found to impact upon pooled diagnostic data; pooled diagnostic performance fell, specifically pooled sensitivities. However, of most concern was the influence of bias associated with the ‘index test’ (i.e. the interpretation, reporting and, if necessary, translation of the GDS-15). Diagnostic performance improved, contrary to what would be expected, when primary studies rated as having a ‘high’ or ‘unclear’ risk of bias were removed from meta-analysis; pooled sensitivity increased and pooled specificity decreased across all cut-off scores. It would be expected that diagnostic performance would be exaggerated when meta-analysis included studies rated as having a ‘high’ or ‘unclear’ risk of bias.

b) Other methodological quality issues

The wide variation in gold-standard reference tests used in the primary studies, would have been a source of heterogeneity. A ‘gold-standard’ diagnostic test is associated with having an acceptable level of validity and reliability, however this ‘level’ will have varied between the gold-standard reference tests used.

Results of the funnel plot for primary studies reporting diagnostic data for the GDS-15 suggest the presence of publication bias and selective outcome reporting. However, publication bias and selective outcome reporting are just two sources of funnel plot asymmetry. Other causes include poor methodological quality of primary studies, true heterogeneity, artefactual occurrence and chance.

For the review of the clinical effectiveness of screening for depression

a) Cochrane Collaborations risk of bias results

Risk of bias in accordance with the Cochrane Collaborations risk of bias tool varied across the eight studies included in the second systematic review presented in Chapter 3. See Table 37. The presence of bias was especially high for ‘blinding of participants and personnel’, which would have significantly influenced results by inflation. Results may have also been influenced by the presence of a detection bias because ‘blinding of outcome assessment’ was rated as ‘unclear’ for five of the eight studies. The presence of bias was also high for the domain of ‘other bias’, which would have influenced results. The most common reason studies were rated as having a ‘high’ risk of bias for this

domain was due to a small sample size. Other reasons include the appropriateness of follow-up points of outcome measures and how generalizable results are.
The presence of selective reporting of results was unclear in seven of the studies as no study protocol was available, so it is unclear if they reported all intended outcomes.

The majority of studies however were rated as having a low risk of bias for 'random sequence generation' and so results were unlikely to have been influenced by process of participant randomisation.

Many studies were rated as having an 'unclear' risk of bias for the different domains of the Cochrane Collaborations risk of bias tool. The presence of an 'unclear' risk of bias may have inflated or deflated results. This does not necessarily mean that bias was present and that results of the study were influenced. This emphasises the importance of research authors to provide greater detail in the reporting of studies.

b) Other methodological quality issues

Three of the included studies (Whooley et al., 2000, van Marwijk et al., 2008, Gitlin et al., 2013) used a cluster randomized control design. Bias is introduced in cluster randomized control trials through recruitment bias, baseline imbalance, incorrect analysis and poor comparability with individual randomised trials (Cochrane Collaboration, 2011). In the studies by Van Marwijk et al., 2008 and Whooley et al., 2000 participant recruitment took place after cluster randomisation of primary care practices/physicians, meaning recruitment bias could be present. Baseline imbalances were present in all three of the cluster trials.

Two studies in particular (Joubert et al., 2013, Shah et al., 2001) had small sample sizes and so are likely to have been underpowered. This would have influenced findings considerably; the sample sizes may have been too small to detect change.

Limitations

Limitations of the review of diagnostic accuracy of brief versions of the GDS

a) Limitations of the primary studies

Results of this review suggest the selective reporting of cut-off scores for the GDS-15. This is suggested by two means. Firstly, not all studies report diagnostic data for the recommended cut-off score of 5; only 23 studies out of a total of 32 primary studies do.

Secondly, expected changes in sensitivity and specificity of the GDS-15 as cut-off score increases are not observed; as cut-off score increases sensitivity should fall; whereas specificity should rise. Table 17 illustrates that, as cut-off score increases, pooled sensitivity only continuously falls from a cut-off score of 5. Pooled sensitivity at a cut-off score of 2 (0.90 (95% CI 0.79-0.95) is lower than that found at a cut-off score of 3 (0.95 (95% CI 0.77-0.99) and pooled sensitivity at a cut-off score of 4 (0.88 (95% CI 0.67-0.96) is lower than that found at cut-off score of 5 (0.89 (95% CI 0.80-0.94) – these findings do not fit the expected pattern in relationship between increasing cut-off score and sensitivity. Table 17 also illustrates that pooled specificity only continuously rises from a cut-off score of 5. Pooled specificity at a cut-off score of 5 (0.77 (95% CI 0.65-0.86) was found to be lower than that found at a cut-off score of 4 (0.86 (95% CI 0.68-0.94). An interpretation of these findings is that study authors may have decided to only report a particular cut-off score if it performs well; if a diagnostic performance was poor at a particular cut-off score a decision may have been made not to report these data. Selective reporting of cut-off scores means that diagnostic performance (i.e. pooled diagnostic data) is artificially inflated, which places considerable limitations on the results of this review.

Statistical heterogeneity was found to be particularly high for all cut-off scores of the GDS-15; the I^2 statistic for all cut-off scores fell within the level that Cochrane guidance described as ‘substantial-considerable heterogeneity’. When meta-analysis was re-run, excluding ‘outliers’, at a cut-off score of 5, diagnostic data remained relatively unchanged; pooled sensitivity increased from 0.89 to 0.90 (95% CI 0.81-0.95), pooled specificity decreased from 0.77 to 0.75 (95% CI 0.60-0.86) and pooled diagnostic odds ratio fell from 27.28 to 26.84 (95% CI 19.11-37.69). Exclusion of ‘outliers’ led to the I^2 statistic falling from 76.7% to 8.1%. See Table 18.

The validity of the GDS-15 for use in older adults with cognitive impairment (i.e. dementia or mild cognitive impairment) is unknown. For the GDS-15, the item 'do you feel have more problems with memory than most?' may lack discriminatory capacity in individuals with a known cognitive impairment. The items 'have you dropped many of your interests and activities?' and 'do you have a prefer to stay at home, rather than going out and doing things?' may also not be indicative of depression in individuals with cognitive impairment. It was not possible to explore the impact of cognitive impairment on the diagnostic accuracy of the GDS-15 in this review for several reasons; not all studies measured cognitive function, when cognitive function was assessed different measures and thresholds were utilised, and in some studies cognitive impairment was an exclusion criterion.

There are no standardised items for briefer versions of the GDS; item composition varied between the primary studies. Therefore, for example, the GDS-4 in one study was a different rating scale from the GDS-4 in another study. Because studies did not use standardised items and because they did not report diagnostic data at the same cut-off scores it was not possible to perform meta-analyses.

b) Limitations of the review itself

Although a protocol was written for this review it was not registered on a database, such as PROSPERO. The protocol however was adhered to throughout the review being undertaken. It is important to register a protocol as it increases the transparency and reliability of a review. This is because a protocol protects against post-hoc decision-making and enables a reader to assess for the presence of selective reporting. The presence of a protocol also aims to reduce duplication of effort if someone else has already planned to undertake the review.

Inclusion and exclusion criteria for study selection were described in the protocol and were followed and applied. However, as study selection and data-extraction were performed by the author (Claire Pocklington) bias may have been introduced. Study selection and data-extraction were only performed by one person because this was an unfunded study and there were not the resources for two people to independently undertake these processes.

An extensive amount of time was spent developing the search strategy, which included grey literature, in order for all titles and abstracts of diagnostic accuracy studies to have been identified. However, it remains possible that relevant studies were not found. A reverse-citation search of the original, 1982 study article describing the development and validation of the GDS was not performed and so relevant studies may not have been found.

This review has not established pooled estimates of the diagnostic accuracy of briefer versions of the GDS because, as discussed, meta-analysis was not possible. This review classified primary studies based in an inpatient or outpatient setting as a secondary care setting. Therefore, such studies were pooled together in meta-analysis for subgroup analysis. It may have been more beneficial to have classified such studies separately because an inpatient setting suggests that study participants are more unwell (physically or mentally). In subgroup analysis of study setting, there was a classification of 'all community' based studies; this classification included primary studies where participants were living in nursing/residential homes and living independently in their own homes. There is clear clinical heterogeneity amongst these studies; therefore, findings of subgroup analysis of 'all community' based studies should be interpreted cautiously. The value of subgroup analysis of 'all community' based studies is questionable. It was possible, however, to further explore the impact of a community study setting by performing subgroup analysis of studies where participants were living independently in the community.

Primary studies utilised a wide-range of gold-standard reference tests.

This review classified older adults as 55 years or older. The decision to classify older adults as 55 years or older was influenced by background reading. Such a cut-off does not reflect the target older adult population in the UK because older adults are viewed as 65 years or older. This may have implications of the applicability of results to the target population of this review. Mean participant age ranged from 66.4 to 87.0 years. Some studies did not report mean participant age; Castello et al. reported that 59.5% of the sample were aged 60-69 years, Licht-Strunk et al. reported that 43.2% of the sample were aged 55-64 years and Malakouti et al. reported that 62.7% of the sample were aged 59-74 years.

Limitations of the review of the clinical effectiveness of screening for depression

a) Limitations of the primary studies

The sample size varied considerably between the seven primary studies (range: N = 8 – 206). Smaller studies are likely to be statistically under-powered and so may not be able to detect differences between treatment groups. In meta-analysis, sample size is taken into account, with larger sample sizes are assigned a greater weight. Meta-analysis was not possible in this review.

Follow-up points of outcome measures varied between the primary studies; there are no established standardised follow-up points. Three studies measured outcomes at one-month (Callahan et al., 1994) and 6-8-weeks (Baldwin et al., 2004, Joubert et al., 2013), which may not be enough time in the course of depression to detect change. The study by Whooley et al. had a single follow-up point of 24-months, which may be a follow-up period that is too long for several reasons; the effects of screening may have 'worn-off' by then, or the depressive episode may have gone into remission and a relapse or new episode could have occurred.

Three studies had a high proportion of study participants which were of an African-American/Black ethnicity (Callahan et al., 1994, Whooley et al., 2000, Gitlin et al., 2013), which may mean the findings of such studies are not generalizable to the UK population.

Outcome findings in the study by Shah et al. varied by the depression rating scale used. This raises questions about the validity of the depression rating scales used to measure outcome and the external consistency of rating scales to each other. It is known that the diagnostic accuracy of depression rating scales varies. Did the primary studies use the 'right' depression rating scale to measure outcome? Was the depression rating scale able to detect and measure the full extent of symptom change?

Several of the primary studies document that it was unclear if the treatment intervention had actually been implemented to all participants in the screening group.

A high degree of clinical heterogeneity was present. Treatment of depression for the intervention group varied amongst the included studies and the clinical effectiveness of such treatment would also have varied. Outcome measures may just reflect the clinical effectiveness of the treatment implemented rather than clinical effectiveness of screening. The effects of screening do not occur in isolation; screening facilitates treatment implementation.

b) Limitations of the review itself

Like the first review, study selection and data extraction were performed by the author (Claire Pocklington), which may have introduced bias. Like the first review, this was an unfunded study and there were not the resources for two people to independently perform study selection and data extraction. Inclusion and exclusion criteria, in accordance with the protocol, were adhered to throughout however.

Despite a considerable amount of time being spent developing an extensive search strategy, which included grey literature, it is possible that the search strategy did not identify all relevant studies.

It was only possible to calculate an effect size for one study. It was not possible to calculate effect size for remaining studies because they presented outcome data in different ways so that effect size could not be extracted, for example, some studies presented outcome measures as 'change score' (i.e. increase/decrease in score from baseline) whereas others presented actual score at follow-up point. Two studies presented outcome data in graphical format and so did not provide the actual figures. For the studies that presented data in graphical format, as a 'change score' or used a cluster-randomised controlled design it was not possible to calculate effect size because required data was not available.

Like the previous review, the age of participants in the review is lower than that of the target population (i.e. older adults in the UK are 65 years of age or older). Mean age of the included primary studies ranged from 65.6 – 85.0 years. This may have implications on the applicability of results to the target population.

INTERPRETATION OF RESULTS

Diagnostic accuracy of brief versions of the GDS

Diagnostic accuracy and cut-off scores of brief versions of the GDS

Data synthesis has established the diagnostic accuracy of the GDS-15 through meta-analysis. Due to a lack of a sufficient number of studies meta-analysis of briefer versions of the GDS was not possible and so analysis was narrative in nature. Reported sensitivities and specificities of different briefer versions of the GDS (i.e. GDS-4, GDS-5, etc.) vary amongst the primary studies; it unclear if such variations reflect true differences in diagnostic performance (i.e. that one particular briefer version of the GDS has superior diagnostic performance compared to others) or if differences are secondary to differences in study samples. It is therefore difficult to comment or conclude which briefer version of the GDS offers the best diagnostic performance for screening. However, briefer versions of the GDS take less time to administer and interpret, which may make them more appealing for use in clinical practice.

There is no clear guidance or recommendations for the diagnostic performance of a screening test in terms of sensitivity and specificity. The balance between sensitivity and specificity will vary between clinical contexts. However, as a general rule a screening programme requires a screening test to have a high sensitivity paired with at least moderate specificity.

Diagnostic meta-analysis has established the diagnostic accuracy of the GDS-15 at a range of different cut-off scores. Interpreting results in order to make a recommendation of which cut-off score provides the best diagnostic performance involves taking into account the impact and influence of the limitations of the review. Such limitations, especially the suggestive presence of the selective reporting of cut-off scores, may indicate that generated pooled diagnostic data maybe over-estimating the diagnostic performance of the GDS-15 at different cut-off scores. Therefore, results have to be interpreted cautiously.

Estimates of pooled diagnostic performance at different cut-off scores will have also been influenced by the established ‘substantial-considerable’ level of statistical heterogeneity.

At a cut-off score of 5, a pooled sensitivity of 0.89 (95% CI 0.80-0.94) and a pooled specificity of 0.77 (95% CI 0.65-0.86) were established, with a corresponding pooled diagnostic odds ratio of 27.28 (95% CI 16.57-44.93). When meta-analysis was re-run excluding 'outliers' heterogeneity improved considerably; the I^2 statistic fell from 76.7% to 8.1%. Pooled sensitivity increased to 0.90 (95% CI 0.81-0.95); whereas pooled specificity fell to 0.75 (95% 0.60-0.86). The pooled diagnostic odds ratio fell to 26.84 (95% CI 19.11-37.69).

Pooled diagnostic data at a cut-off score of 4 appears more favourable; pooled sensitivity 0.88 (95% CI 0.67-0.96), pooled specificity 0.86 (95% CI 0.68-0.94) and pooled diagnostic odds ratio 42.05 (95% CI 17.41-101.49). However, meta-analysis only included 10 studies at a cut-off score of 4 compared to 23 at a cut-off score of 5. The I^2 statistic was 85.8%. When 'outliers' were removed the I^2 statistic fell to 14.7%. Pooled sensitivity decreased to 0.86 (95% CI 0.59-0.96); whereas pooled specificity remained at 0.86. This resulted in a fall in the diagnostic odds ratio; 37.17 (95% CI 18.45-74.90).

Comparison of pooled diagnostic data at a cut-off score of 5 and 6 are more appropriate as a similar number of primary studies were included in meta-analysis; 23 and 20 respectively. At a cut-off score of 6, a lower pooled sensitivity of 0.79 (95% CI 0.68-0.87) and a higher pooled specificity of 0.83 (95% CI 0.72-0.90) were established. This resulted in a notably lower pooled diagnostic odds ratio of 17.61 (95% CI 10.12-30.63). When 'outliers' were removed from meta-analysis the I^2 statistic fell from 88.1% to 0.0%. Pooled sensitivity fell notably to 0.75 (95% CI 0.63-0.84), though pooled specificity remained unchanged at 0.83. The pooled diagnostic odds ratio fell to 14.70 (95% CI 11.60-18.63).

Pooled diagnostic data at other cut-off scores were less favourable compared to a cut-off score of 5. The findings of this review support existing evidence that the most appropriate cut-off score for the GDS-15 is 5; however, as discussed above, these results are limited by the potential for selective reporting of cut-off points.

Findings and relevance of subgroup analyses to interpreting results

a) ***Age***

The proportion of older adults >85 years of age is expected to rise by 106%, as discussed in Chapter 1 (The Kings Fund, 2014), and therefore a depression screening tool ideally should maintain diagnostic performance regardless of an individual's age. The GDS-15 may not be an appropriate screening tool to use in older adults aged ≥75 years old because pooled diagnostic data suggests that diagnostic performance of the GDS-15 deteriorates. For example, the pooled diagnostic odds ratio was 42.20 (95% CI 20.33-87.71) for studies classed as 'young-old' and 17.06 (95% CI 2.28-35.02) for studies classed as 'middle-old'. Meta-analysis of 'young-old' studies was associated with greater heterogeneity, however when this was explored observed differences in diagnostic performance in accordance to mean age remained. Meta-regression revealed age was not a significant predictive factor for diagnostic accuracy however.

b) ***Setting of screening***

Subgroup analysis of study setting has found that the diagnostic performance - specifically pooled specificity and diagnostic odds ratio - of the GDS-15 is lowest in a primary care setting. Pooled specificity falls from 0.77 to 0.63 (95% 0.42-0.80) and pooled diagnostic odds ratio falls from 27.28 to 18.58 (95% CI 13.14-26.26) when only primary care based studies are included in meta-analysis. This suggests that screening for depression in older adults using the GDS-15 would not be best delivered in primary care. This review has found that diagnostic performance of the GDS-15 is better in a community setting, especially when individuals are living independently, compared to primary or secondary care. For example, at a cut-off score of 5, the pooled diagnostic odds ratio was 18.58 (95% 13.14-26.27) for a primary care setting, 29.31 (95% CI 9.19-93.47) for a community based setting and 56.71 (95% CI 10.32-311.38) for a community independent living setting. Meta-regression analysis revealed that a primary care setting was not a significant predictive factor for diagnostic accuracy however.

It is unclear what the difference is between studies where participants were living independently in the community and studies based in primary care setting. Study participants living independently in the community may have utilised primary care services to the same extent as participants in primary care based studies.

Pooled diagnostic data for a secondary care setting is similar to pooled diagnostic data for all studies and more favourable than use of the GDS-15 in primary care. Secondary care may be a more opportune setting to delivering depression screening to older adults.

Findings of heterogeneity

The same studies were consistently identified as 'outliers'. The studies by Broekman et al., 2011 and Marc et al., 2008, were identified as 'outliers' at different cut-off scores for pooled diagnostic data and in all subgroup and sensitivity analyses. The study by Broekman et al. was the most frequent identified 'outlier'; it was the only study that used more than one language version of the GDS-15 and was the only study in a community setting, which specifically described study participants as social service users. The study by Marc et al. was the second most frequent identified 'outlier'; there was nothing different or unusual about this study compared to others. Both Broekman et al. and Marc et al. excluded participants with cognitive impairment, directly administered the GDS-15, used the DSM-IV SCID as the gold-standard diagnostic test and had 'low' overall ratings for each domain of the QUADAS-II.

Other identified 'outliers' through pooled diagnostic data, subgroup analyses and sensitivity analyses included Abas et al., 1998, De Craen et al., 2003, Izal et al., 2010, Phelan et al., 2010, Van Marwijk et al., 1995, Watson et al., 2004 and Wongpakaran et al., 2013.

Heterogeneity, as measured by the I^2 statistic, consistently improved when meta-analysis was re-run excluding 'outliers'. Pooled diagnostic data did not change considerably for cut-off scores of 4 – 6, which suggests that 'outliers' had diminutive impact upon diagnostic accuracy.

Comparison with existing evidence

Some primary studies, which were included in existing reviews did not meet the inclusion and exclusion criteria of this review and were therefore excluded. Comparison of the current results to existing reviews is difficult due to different approaches to statistical analysis, which have involved pooling data from different cut-off scores. For example, the review by Wancata et al., 2006, reports a mean sensitivity and specificity of the GDS-

15 regardless of cut-off score. Watson and Pignone, 2003, report the range in sensitivity and specificity reported by the five primary studies identified at a cut-off score of 3 to 5. Both reviews by Mitchell et al. (Mitchell et al., 2010a, Mitchell et al., 2010b) report pooled sensitivity and specificity regardless of cut-off score. See Table 41.

It is only possible to compare pooled diagnostic data from this review with one of the existing reviews; Dennis et al., 2012. Dennis et al. have performed meta-analyses of pooled diagnostic data at different cut-off scores thus allowing direct comparison with the findings of this review.

Review	Diagnostic data for a cut-off score of 5		
	Reported	**Sensitivity (%)**	**Specificity (%)**
Dennis et al., 2012	Meta-analysis of 14 primary studies at specific cut-off scores	79 (95% CI 70 – 86)	77 (95% CI 73 – 81)
Mitchell et al., 2010a	Meta-analysis of 15 primary studies regardless of cut-off score	83.4 (95% CI 79.7 – 88.4)	73.8 (95% CI 68.0 – 79.2)
Mitchell et al., 2010b	Meta-analysis of 10 primary studies regardless of cut-off score	81.3 (95% CI 77.2 – 85.2)	78.4 (95% CI 71.2 – 84.8)
Wancata et al., 2006	Mean of 21 studies regardless of cut-off score	0.805	0.750
Watson and Pignone, 2003	Range of 5 identified primary studies at cut-off scores of 3-5	82 – 100	72 – 82

Table 41: Diagnostic data reported in previous systematic reviews of the diagnostic accuracy of the GDS-15

The review by Dennis et al. identified only 14 primary studies; 9 less than this study. At a cut-off score of 5, pooled sensitivity in this review is higher than that found by Dennis et al.; 0.89 (95% CI 0.80–0.94) and 0.79 (95% CI 0.70–0.86) respectively. Pooled specificities however are the same; 0.77. The review by Dennis et al. found a pooled diagnostic odds ratio of 12.40 (95% CI 6.67–23.06), which is lower than the pooled diagnostic odds ratio found by this review; 27.28 (95% CI 16.57–44.93).

The review by Dennis et al. reports pooled diagnostic data at a cut-off score of 4 and 6 as well. At a cut-off score of 4, both reviews reported a pooled sensitivity of 0.88; however, pooled specificities varied; 0.64 (95% CI 0.57-0.70) vs. 0.86 (95% CI 0.68-0.94)

– 95% confidence intervals just overlap. See Table 42. Pooled diagnostic data for both reviews are more similar at a cut-off score 6.

Cut-off score	Sensitivity (95% CI)		Specificity (95% CI)		Diagnostic odds ratio (95% CI)	
	Dennis et al. (2012)	This review	Dennis et al. (2012)	This review	Dennis et al. (2012)	This review
4	0.88 (0.77-0.95)	0.88 (0.67-0.96)	0.64 (0.57-0.70)	0.86 (0.68-0.94)	13.61 (0.09-0.37)	42.05 (17.42-101.49)
5	0.79 (0.70-0.86)	0.89 (0.80-0.94)	0.77 (0.73-0.81)	0.77 (0.64-0.86)	12.40 (6.67-23.06)	27.28 (16.57-44.93)
6	0.74 (0.61-0.88)	0.79 (0.68-0.87)	0.81 (0.76-0.87)	0.83 (0.72-0.90)	12.62 (2.40-22.84)	17.61 (10.12-30.63)

Table 42: Comparison of diagnostic data of this review with the previous review by Dennis et al., 2012

The clinical effectiveness of screening for depression in older adults

The eight identified primary studies report outcome measures at a total of fourteen different follow-up points. Four studies (Whooley et al., 2000, Baldwin et al., 2004, van Marwijk et al., 2008, Gitlin et al., 2013), all of a Hewitt level 2 study design, found statistically significant evidence that screening is associated with better outcomes. These studies have reported outcome measures at varying follow-up points; 6-8 weeks, 4 months, 6 months and 24 months.

The remaining studies do not provide evidence that screening for depression in older adults is clinically effective. There is no statistically significant evidence that clinical outcomes are better for the intervention group in comparison to the control group.

Overall, evidence in favour of screening leading to better clinical outcomes, and so being clinically effective, is limited. An evidence gap regarding the clinical effectiveness of screening for depression in older adults remains.

Non-statistically significant findings were associated with smaller sample sizes (Shah et al., 2001, Cullum et al., 2007, Joubert et al., 2013). No other differences in regards to study setting or sample characteristics were associated with statistically significant or non-statistically significant findings.

Comparison with existing evidence

The one existing review by O'Connor et al. identified only four studies. The findings of the review informed decisions made by the US Preventative Service Task Force that screening for depression in older adults is not recommended. It concluded overall that screening for depression alone is not enough and commented that the delivered interventions used in management have multifactorial benefits. In summary, there is no benefit to screening if additional resources are not available to deliver additional care interventions. Such conclusions support the findings of this review.

This review identified a greater number of primary studies but not all the studies in the O'Connor review met the inclusion criteria of this review. One study which did not meet the inclusion criteria of this review was that by Rubenstein et al. This is because depression was not the only condition screened for; study participants underwent screening for falls/balance problems, urinary incontinence, depression, memory loss and functional impairment. Therefore, not all study participants will have screened positive for depression but would have been included in outcome measures relating to depression. No statistical analysis was performed by O'Connor et al.

IMPLICATIONS FOR FUTURE RESEARCH

Future research regarding the diagnostic accuracy of brief versions of the GDS

Meta-analysis was not possible for briefer versions of the GDS due to a lack of primary studies, variation in item composition of briefer GDS versions and variation in cut-off scores reported. There is a need for more primary studies to explore the diagnostic accuracy of briefer versions of the GDS. There are no standardised briefer versions of the GDS, i.e. item combination of specific briefer versions varies between studies. Future research needs to establish, which combination of items provides the best diagnostic performance and accuracy.

An important limitation identified in this review is the presence of selective reporting of cut-off scores, which artificially inflates the diagnostic accuracy of the GDS-15. Therefore, there is a need for future diagnostic accuracy studies of the GDS-15 to report all cut-off scores as this will enable future reviews to establish more accurate pooled diagnostic data, which can be interpreted with less caution.

Dementia is a significant issue in an older adult population and therefore the effect of cognitive impairment on the diagnostic accuracy of the GDS needs to be addressed in future research. Ideally future research studies, should measure cognitive function and explore this in data analysis.

Future research needs to further investigate the effect of age on diagnostic performance of the GDS-15 and upon briefer versions. In this review, meta-analysis was not possible for a mean age ≥85 years due to an insufficient number of studies. Researchers should state mean participant age, rather than describing sample age by proportions, to enable study inclusion in meta-analysis.

Subgroup analysis of the review revealed that future research needs to explore the influence of community or healthcare setting on diagnostic accuracy in more detail. For example, diagnostic accuracy should be established independently for outpatient and inpatient secondary care settings. Researchers need to have clear descriptions of what the differences are between study participants in community and primary care settings.

Future research regarding the clinical effectiveness of screening for depression in older adults

The review suggests that there is a need for a standardised methodological model to investigate the clinical effectiveness of screening, which is not just specific to screening for depression in older adults. Ideally, a study exploring the clinical effectiveness of screening should not simultaneously investigate the effectiveness of a new treatment. From this review, two methodological models for investigating the clinical effectiveness of screening for any condition are suggested;

1) Only intervention group screened. Consequent treatment in study, regardless of individual being in intervention or control group, must follow routine practice, guidelines, etc.
2) Both intervention and control group are screened but only results disclosed to those responsible for care of the intervention group. Consequent treatment in study, regardless of individual being in intervention or control group, must follow routine practice, guidelines, etc.

The former methodological model for study design would be preferential and would be akin to the Hewitt level 1 evidence model of study design.

There is a need for more research to establish the clinical effectiveness of screening for depression in older adults. Future studies exploring the clinical effectiveness of screening for depression should report data to facilitate inclusion in meta-analysis.

This dissertation has found some, but limited, evidence that screening for depression is clinically effective. This may have implications on healthcare resources, particularly economic concerns. A screening programme for depression will generate an increase in rate of detection, which could lead to increased prescribing of antidepressant medication, increased referral to psychological services and secondary care mental health services for examples, all of which have associated financial costs. However, earlier diagnosis may lead to improved clinical outcomes, which in turn could reduce healthcare costs. Future research therefore needs to explore the cost effectiveness of screening for depression in older adults. Current services may not have capacity to absorb an increase in the number of people diagnosed with depression and therefore future research would also have to explore this.

IMPLICATIONS FOR POLICY AND PRACTICE

In the UK, screening for depression in older adults is not policy or practice. The National Screening Committee has published guidance criteria regarding the viability, effectiveness and appropriateness of a screening programme. This guidance, as discussed in Chapter 1, cover the domains of the condition in question, the screening test, the treatment intervention, the effectiveness of the screening programme and implementation criteria. The guidelines are comprised of twenty separate criteria (England, 2013, updated 2015). See Appendix 1.

This dissertation cannot make recommendations about introducing a screening programme for depression in older adults because it only addresses and provides evidence towards some of these criteria. This dissertation addresses and provides evidence towards the following criteria;

- Criteria 4: *'There should be a simple, safe, precise and validated screening test.'* The GDS-15 could be an ideal screening test. It is simple and safe to administer. Findings of the first review (Chapter 2) have shown that it has acceptable diagnostic performance and validity in older adults. However, as discussed, findings of this review should be approached cautiously due to the selective reporting of cut-off scores.
- Criteria 5: *'The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.'* Findings of the review regarding the diagnostic accuracy of brief versions of the GDS, suggest that the GDS-15, at a cut-off score of 5, would provide acceptable diagnostic performance (high sensitivity and modest specificity) for a screening programme.
- Criteria 11: *'There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity'.* The findings of the second review of the clinical effectiveness of screening for depression in older adults (Chapter 3) provides some evidence that screening does improve clinical outcomes; however overall evidence is lacking.

To make a recommendation for a screening programme for depression in an older adult population evidence regarding all criteria of the National Screening Committee guidelines would have to be established. Chapter 1 of this dissertation presents why depression in older adults is an important clinical topic, which links to Criteria 1 of the National Screening Committee guideline: '*the condition should be an important health problem as judged by its frequency and/or severity*'. Older adults, and the wider public, should be made aware of the importance of this condition because individuals should be able to make an informed choice about whether to participate in screening (Criteria 19).

As discussed above, this dissertation provides evidence that the GDS-15 is *'a simple, safe, precise and validated screening test'* (Criteria 4) and that test values have been established (Criteria 5). Future research should explore if the GDS-15 is *'acceptable to the target population*' (Criteria 6). There also needs to be consensus on diagnostic procedures and treatment (Criteria 7). NICE guidance for the treatment of depression is available, which could address Criteria 9 and 10.

As well as the acceptability of the screening test being established, the acceptability of the whole screening programme (i.e. the screening test, diagnostic procedure and the treatment) to the older adult population, healthcare professionals and the wider public would have to be established, which would fulfil Criteria 12 of the National Screening Committee guidelines. Future research also has to establish if the benefits of screening for an individual would outweigh any (potential) harms (Criteria 13). The practicalities of administrating a screening programme for depression in older adults would also need to be explored (Criteria 17 and 18).

There is some evidence to support the clinical effectiveness of screening for depression in an older adult population, however, evidence is lacking so more research would have to be undertaken (Criteria 11). The cost effectiveness of screening for depression in older adults would also have to be explored (Criteria 14).

CONCLUSION

This dissertation presents and explains why depression in older adults is an important clinical topic currently and why it is going to become a more pressing issue in the future. As addressed, screening for depression in older adults could improve detection rates, lead to better clinical outcomes, reduced healthcare utilisation and, in turn, reduced healthcare costs.

The clinical effectiveness of a screening programme for depression in older adults is essentially dependent upon two factors; first, the diagnostic accuracy of the tool used to screen and secondly, the effectiveness of the treatment consequently implemented. There is an abundance of evidence available regarding effective treatments for depression in older adults, the details of which are beyond the scope of this dissertation.

This dissertation presents up-to-date evidence regarding the diagnostic accuracy of different brief versions of the GDS, which could become a first choice option for utilisation in a screening programme for depression in older adults. This dissertation has found that the GDS-15, at a cut-off 5, provides acceptable diagnostic performance as a screening tool. However, findings have to be interpreted cautiously because results may be biased due to selective reporting of cut-off scores. Time constraints and demand on busy clinical practice may require the use of a brief screening tool, such as briefer versions of the GDS. Unfortunately, this dissertation has not been able to produce statistical evidence regarding the diagnostic accuracy of briefer versions of the GDS. This is an area for future research.

This dissertation has found some evidence that screening for depression is associated with greater symptom improvement: however, evidence is limited. For policy and practice to be influenced and so for a screening programme for depression in older adults to be introduced, more evidence regarding the clinical effectiveness (and detrimental impacts) of screening need to be established. In addition, the cost effectiveness of screening for depression in older adults must be established.

REFERENCES

2002. Geriatric Depression Scale (GDS). *Occasional paper (Royal College of General Practitioners)*, 46-46.

ABAS, M. A., PHILLIPS, C., CARTER, J., WALTER, J., BANJEREE, S. & LEVY, R. 1998. Culturally sensitive validation of screening questionnaires for depression in older African-Caribbean people living in south London. *The British Journal of Psychiatry,* 173**,** 249-254.

ABRAMS, R. C., LACHS, M., MCAVAY, G., KEOHANE, D. J. & BRUCE, M. L. 2002. Predictors of self-neglect in community-dwellings elders. *American Journal of Psychiatry,* 159**,** 1724-30.

ALEXOPOULOS, G. S. 2005. Depression in the elderly. *The Lancet,* 365**,** 1961-1970.

ALEXOPOULOS, G. S., KIOSSES, D. N., HEO, M., MURPHY, C. F., SHANMUGHAM, B. & GUNNING-DIXON, F. 2005. Executive dysfunction and the course of geriatric depression. *Biological Psychiatry,* 58**,** 204-10.

ALEXOPOULOS, G. S., MEYERS, B. S., YOUNG, R. C., KALAYAM, B., KAKUMA, T., GABRIELLE, M., SIREY, J. A. & HULL, J. 2000. Executive dysfunction and long-term outcomes of geriatric depression. *Archives of General Psychiatry,* 57**,** 285-290.

ALLGAIER, A.-K., KRAMER, D., SARAVO, B., MERGL, R., FEJTKOVA, S. & HEGERL, U. 2013. Beside the Geriatric Depression Scale: The WHO-Five Well-being Index as a valid screening tool for depression in nursing homes. *International Journal of Geriatric Psychiatry,* 28**,** 1197-1204.

ALLGAIER, A. K., KRAMER, D., MERGL, R., FEJTKOVA, S. & HEGERL, U. 2011. Validity of the geriatric depression scale in nursing home residents: Comparison of GDS-15, GDS-8, and GDS-4. [German]

Validitat der Geriatrischen Depressionsskala bei Altenheimbewohnern: Vergleich von GDS-15, GDS-8 und GDS-4. *Psychiatrische Praxis,* 38**,** 280-286.

ALMEIDA, O. P. & ALMEIDA, S. A. 1999. Reliability of the Brazilian version of the geriatric depression scale (GDS) short form. *Arquivos de Neuro-Psiquiatria,* 57**,** 421-426.

AREAN, P. A. & AYALON, L. 2005. Assessment and treatment of depressed older adults in primary care. *Clinical Psychology: Science and Practice,* 12**,** 321-335.

AROMA, A., RAITASALO, A., REUNANEN, O., IMPIVAARA, M., HELIOVAARA, P. & KNEKT, P. 1994. Depression and cardiovascular diseases. *Acta Psychiatr Scand.* Supple, 377, 77-82.

ARTHUR, A., JAGGER, C., LINDESAY, J., GRAHAM, C. & CLARKE, M. 1999. Using an annual over-75 health check to screen for depression: validation of the short Geriatric Depression Scale (GDS15) within general practice. *International Journal of Geriatric Psychiatry,* 14**,** 431-9.

ASSOCIATION, A. P. 2013. *Diagnostic and statistical manual of mental disorders,* Arlington, VA, American Psychiatric Publishing.

BAE, J. N. & CHO, M. J. 2004. Development of the Korean version of the Geriatric Depression Scale and its short form among elderly psychiatric patients. *Journal of psychosomatic research,* 57**,** 297-305.

BALDWIN, R. C. 2000. Poor prognosis of depression in elderly people: causes and actions. *Annals of Medicine,* 32**,** 252-6.

BALDWIN, W., PRATT, H., GORING, H., MARRIOTT, A. & ROBERTS, C. 2004. Does a nurse-led mental health liaison service for older people reduce psychiatric morbidity in acute general medical wards? A randomised controlled trial. *Age & Ageing,* 33**,** 472-478.

BEHEYDT, L., SCHRIJVERS, D., DOCX, L., BOUCKAERT, F., HULSTIJN, W. & SABBE, B. 2014. Psychomotor retardation in elderly untreated depressed patients. *Frontiers in Psychiatry,* 5**,** 196.

BENNABI, D., VANDEL, P., PAPAXANTHIS, C., POZZO, T. & HAFFEN, E. 2013. Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic and therapeutic implications. *BioMed Research International*.

BERNABEI, V., MORINI, V., MORETTI, F., MARCHIORI, A., FERRARI, B., DALMONTE, E., DE RONCHI, D. & RITA ATTI, A. 2011. Vision and hearing impairments are associated with depressive-anxiety syndrome in Italian elderly. *Aging Mental Health,* 15(4), 467-74

BEYNON, R., LEEFLANG, M., MCDONALD, S., EISINGA, A., MITCHELL, R. L., WHITING, P. & GLANVILLE, J. M. 2013. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database Syst Rev,* 9.

BEYONDBLUE 2009. Depression in older age: A scoping study. *In:* HARALAMBOUS, B., LIN, X., DOW, B., JONES, C., TINNEY, C., BRYANT, C. (ed.) *beyondblue: the national depression initiative.* Australia: National Ageing Research Institute.

BIELIAUSKAS, L. A., AND LAUREN L. DRAG 2013. *Differential Diagnosis of Depression and Dementia,* New York, Springer.

BIJL, D., VAN MARWIJK, H. W., ADÉR, H. J., BEEKMAN, A. T. & DE HAAN, M. 2006. Test-characteristics of the GDS-15 in screening for major depression in elderly patients in general practice. *Clinical gerontologist,* 29**,** 1-9.

BIRRER, R. B. & VEMURI, S. P. 2004. Depression in later life: a diagnostic and therapeutic challenge. *American Family Physician,* 69.

BLANK, K., GRUMAN, C. & ROBISON, J. T. 2004. Case-Finding for Depression in Elderly People: Balancing Ease of Administration with Validity in Varied Treatment Settings. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences,* 59**,** 378-384.

BOSMANS, J., DE BRUIJNE, M., VAN HOUT, H., VAN MARWIJK, H., BEEKMAN, A., BOUTER, L., STALMAN, W. & VAN TULDER, M. 2006. Cost-Effectiveness of a Disease Management Program for Major Depression in Elderly Primary Care Patients. *Journal of general internal medicine,* 21**,** 1020-1026.

BROEKMAN, B. F., NITI, M., NYUNT, M. S. Z., KO, S. M., KUMAR, R. & NG, T. P. 2011. Validation of a brief seven-item response bias-free Geriatric Depression Scale. *The American Journal of Geriatric Psychiatry,* 19**,** 589-596.

BROWN, J. M., STEWARD, J. C., STUMP, T. E. & CALLAHAN, C. M. 2011. Risk of coronary heart disease events over 15 years among older adults with depressive symptoms. *The American Journal of Geriatric Psychiatry,* 19**,** 721-729.

BUTTERS, M. A., MULSANT, B. H., HOUCK, P. R., DEW, M. A., NEBES, R. D., REYNOLDS, C. F., III, BHALLA, R. K., MAZUMDAR, S., BEGLEY, A. E., POLLOCK, B. G. & BECKER, J. T. 2004a. Executive Functioning, Illness Course, and Relapse/Recurrence in Continuation and Maintenance Treatment of Late-life Depression: Is There a Relationship? *The American Journal of Geriatric Psychiatry,* 12**,** 387-394.

BUTTERS, M. A., WHYTE, E. M., NEBES, R. D., BEGLEY, A. E., DEW, M. A., MULSANT, B. H., ZMUDA, M. D., BHALLA, R., MELTZER, C. C. & POLLOCK, B. G. 2004b. The

Nature and Determinants of Neuropsychological Functioning in Late-LifeDepression. *Archives of General Psychiatry,* 61**,** 587-595.

CALLAHAN, C. M., HENDRIE, H. C., DITTUS, R. S., BRATER, D. C., HUI, S. L. & TIERNEY, W. M. 1994. Improving treatment of late life depression in primary care: a randomized clinical trial. *Journal of the American Geriatrics Society*.

CASTELLO, M. S., COELHO-FILHO, J. M. & CARVALHO, A. F. 2010. Validity of the Brazilian version of the Geriatric Depression Scale (GDS) among primary care patients. *International Psychogeriatrics,* 22**,** 63-66.

CENTRE FOR REVIEWS AND DISSEMINATION. 2008. Systematic Reivew: CRD guidance for undertaking reviews in healthcare. *CRD, University of York, York.*

CHAPMAN, D. P. & PERRY, G. S. 2008. Depression as a major component of public health for older adults. *Centres for Disease Control and Prevention,* 7.

CHERUBINI, A., NISTICO, G., ROZZINI, R., LIPEROTI, R., DI BARI, M., ZAMPI, E., FERRANNINI, L., AGUGLIA, E., PANI, L. & BERNABEI, R. 2012. Subthreshold depression in older subjects: An unmet therapeutic need. *The journal of nutrition, health & aging,* 16**,** 909-913.

COLE, M. G. & DENDUKURI, N. 2003. Risk factors for depression among elderly community subjects: a systematic review and meta-analysis. *American Journal of Psychiatry,* 160**,** 1147-1156.

COLLABORATION, C. 2011. 16.3.2 Assessing risk of bias in cluster-randomised trials. *In:* HIGGINS, J. P. & GREEN, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* The Cochrane Collaboration.

CONRADSSON, M., ROSENDAHL, E., LITTBRAND, H., GUSTAFSON, Y., OLOFSSON, B. & LÖVHEIM, H. 2013. Usefulness of the Geriatric Depression Scale 15-item version among very old people with and without cognitive impairment. *Aging & mental health,* 17**,** 638-645.

CRABB, R. & HUNSLEY, J. 2006. Utilization of mental health care services among older adults with depression. *Journal of Clinical Psychology,* 62**,** 299-312.

CUIJPERS, P., SMIT, F., OOSTENBRINK, J., DE GRAAF, R., TEN HAVE, M. & BEEKMAN, A. 2007. Economic costs of minor depression: a population-based study. *Acta Psychiatrica Scandinavica,* 115**,** 229-236.

CUIJPERS, P., VAN STRATEN, A. & SMIT, F. 2006. Psychological treatment of late-life depression: a meta-analysis of randomized controlled trials. *International Journal of Geriatric Psychiatry,* 21**,** 1139-49.

CULLUM, S., TUCKER, S., TODD, C. & BRAYNE, C. 2006. Screening for depression in older medical inpatients. *International Journal of Geriatric Psychiatry,* 21**,** 469-476.

CULLUM, S., TUCKER, S., TODD, C. & BRAYNE, C. 2007. Effectiveness of liaison psychiatric nursing in older medical inpatients with depression: a randomised controlled trial. *Age & Ageing,* 36**,** 436-42.

DATH, P., KATONA, P., MULLAN, E., EVANS, S. & CORNELIUS, K. 1994. SCREENING, DETECTION AND MANAGEMENT OF DEPRESSION IN ELDERLY PRIMARY-CARE ATTENDERS .1. THE ACCEPTABILITY AND PERFORMANCE OF THE 15-ITEM GERIATRIC DEPRESSION SCALE (GDS15) AND THE DEVELOPMENT OF SHORT VERSIONS. *Family practice,* 11**,** 260-266.

DAVISON, T. E., MCCABE, M. P. & MELLOR, D. 2009. An examination of the gold standard diagnosis of major depression in aged-care settings. *American Journal of Geriatric Psychiatry,* 17**,** 359-367.

DE CRAEN, A. J., HEEREN, T. & GUSSEKLOO, J. 2003. Accuracy of the 15-item geriatric depression scale (GDS-15) in a community sample of the oldest old. *International Journal of Geriatric Psychiatry,* 18**,** 63-66.

DEBETTE, S. & MARKUS, H. S. 2010. The clinical importance of white matter hyper intensities on brain magnetic resonance imaging; systematic review and meta-analysis. *BMJ,* 341.

DENNIS, M., KADRI, A. & COFFEY, J. 2012. Depression in older people in the general hospital: A systematic review of screening instruments. *Age and Ageing,* 41**,** 148-154.

DICTIONARY, O. 2015. *Oxford Dictionary of English*, Oxford University Press.

DJERNES, J. K. 2006. Prevalence and predictors of depression in populations of elderly: a review. *Acta Psychiatrica Scandinavica,* 113**,** 372-387.

DRAYER, R. A., MULSANT, B. H., LENZE, E. J., ROLLMAN, B. L., DEW, M. A., KELLEHER, K., KARP, J. F., BEGLEY, A., SCHULBERG, H. C. & REYNOLDS, C. F., 3RD 2005. Somatic symptoms of depression in elderly patients with medical comorbidities. *International Journal of Geriatric Psychiatry,* 20**,** 973-82.

EL-GABALAWY, R., MACKENZIE, C., THIBODEAU, M., ASMUNDSON, G. & SAREEN, J. 2013. Health anxiety disorders in older adults: conceptualizing complex conditions in late life. *Clinical Psychology Review,* 33**,** 1096-1105.

ENGLAND, P. H. 2013. *Population screening programmes - guidance. NHS population screening explained.* [Online]. Public Health England. Available: http://www.gov.uk/guidance/nhs-population-screening-explained [Accessed october 1st 2015].

ENGLAND, P. H. 2013, updated 2015. Criteria for appraising the viability, the effectiveness and appropriateness of a screening programme. *In:* ENGLAND, P. H. (ed.). London: Gov.uk.

EVANS, M. 1995. Detection and management of depression in the elderly physically ill patient. *Human Psychopharmacology: Clinical and Experimental,* 10**,** S235-S241.

EVANS, M., HAMMOND, M., WILSON, K., LYE, M. & COPELAND, J. 1997. Treatment of depression in the elderly: effect of physical illness on response. *International Journal of Geriatric Psychiatry,* 12**,** 1189-94.

EVANS, M. & MOTTRAM, P. 2000. Diagnosis of depression in elderly patients. *Advances in Psychiatric Treatment,* 6**,** 49-56.

EXCELLENCE, N. I. F. H. A. C. 2016. Depression in adults: recognition and management. *NICE guidelines [CG90]*.

EXCELLENCE, N. I. O. C. 2009. Depression in adults with a chronic physical health problem: treatment and management. *NICE*.

FINKELSTEIN, E. A., BRAY, J. W., CHEN, H., LARSON, M. J., MILLER, K., TOMPKINS, C., KEME, A. & MANDERSCHEID, R. 2003. Prevalence and costs of major depression among elderly claimants with diabetes. *Diabetes care,* 26**,** 415-420.

FISKE, A., WETHERELL, J. L. & GATZ, M. 2009. Depression in older adults. *Annual review of clinical psychology,* 5**,** 363-389.

FOLSTEIN, M., FOLSTEIN S. E. & MCHUGH, P. R. 1975. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189-198.

FRIEDMAN, B., CONWELL, Y., DELAVAN, R. R., WAMSLEY, B. R. & EGGERT, G. M. 2005a. Depression and suicidal behaviors in Medicare primary care patients under age 65. *Journal of general internal medicine,* 20**,** 397-403.

FRIEDMAN, B., HEISEL, M. J. & DELAVAN, R. L. 2005b. Psychometric Properties of the 15-Item Geriatric Depression Scale in Functionally Impaired, Cognitively Intact, Community-Dwelling Elderly Primary Care Patients. *Journal of the American Geriatrics Society,* 53**,** 1570-1576.

FUND, T. K. 2014. *Ageing Population* [Online]. Available: http://www.kingsfund.org.uk/time-to-think-differently/trends/demography/ageing-population [Accessed October 1st 2014].

GELLIS, Z. D., MCCRACKEN, S. G. Depressive Disorders in Older Adults. *Mental Health and Older Adults.* Chicago: Council on Social Work Education.

GERETY, M. B., WILLIAMS JR, J. W., MULROW, C. D., CORNELL, J. E., KADRI, A. A., ROSENBERG, J., CHIODO, L. K. & LONG, M. 1994. Performance of case-finding tools for depression in the nursing home: influence of clinical and functional characteristics and selection of optimal threshold scores. *Journal of the American Geriatrics Society,* 42**,** 1103-1109.

GITLIN, L. N., HARRIS, L. F., MCCOY, M. C., CHERNETT, N. L., PIZZI, L. T., JUTKOWITZ, E., HESS, E. & HAUCK, W. W. 2013. A Home-Based Intervention to Reduce Depressive Symptoms and Improve Quality of Life in Older African Americans: A Randomized Trial. *Annals of Internal Medicine,* 159**,** 243-252.

GOTHE, F., ENACHE, D., WAHLUND, L. O., WINBLAD, B., CRISBY, M., LOKK, J. & AARSLAND, D. 2012. Cerebrovascular diseases and depression: epidemiology, mechanisms and treatment. *Panminerva Medica,* 54**,** 161-170.

GROUP, N. Z. G. 2008. Identification of common mental disorders and management of depression in primary care. *Wellington: An Evidence-based Best Practice Guideline*.

HAMILTON, M. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, 23, 56-62.

HARPOLE, L. H., WILLIAMS, J. W., JR., OLSEN, M. K., STECHUCHAK, K. M., ODDONE, E., CALLAHAN, C. M., KATON, W. J., LIN, E. H., GRYPMA, L. M. & UNUTZER, J. 2005. Improving depression outcomes in older adults with comorbid medical illness. *General Hospital Psychiatry,* 27**,** 4-12.

HEGEMAN, J., KOK, R., VAN DER MAST, R. & GILTAY, E. 2012. Phenomenology of depression in older compared with younger adults: meta-analysis. *The British Journal of Psychiatry,* 2000**,** 275-281.

HEWITT, C. E., GILBODY, S.M., BREALEY, S., PAULDEN, M., PALMER, S., MANN, R, *et al.* Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis. *Health Technology Assessment*, 2009, 13(36), 77-80

HICKIE, I., SCOTT, E., WILHELM, K. & BRODATY, H. 1997. Subcortical hyperintensities on magnetic resonance imaging in patients with severe depression--a longitudinal evaluation. *Biological Psychiatry,* 42**,** 367-74.

HICKIE, I., SIMONS, L., NAISMITH, S., SIMONS, J., MCCALLUM, J. & PEARSON, K. 2003. Vascular risk to late-life depression: evidence from a longitudinal community study. *Australian & New Zealand Journal of Psychiatry,* 37**,** 62-65.

HIGGINS, J. P., GREEN, S. & (EDITORS) 2011. 9.5 Heterogeneity. *Cochrane Handbook for Systematic Reviews of Interventions.* The Cochrane Collaboration.

HILDEBRAND, C., TAYLOR, M. & BRADWAY, C. 2014. Elder self-neglect: the failure of coping because of cognitive and functional impairments. *Journal of the American Association of Nurse Practitioners,* 26**,** 452-462.

HODKINSON, H. M. 1972. Evaluation of a mental test for assessment of mental impairment in the elderly. *Age and Ageing*, 1(4), 233-238.

HUANG, C.-Q., DONG, B.-R., LU, Z.-C., YUE, J.-R. & LIU, Q.-X. 2010. Chronic diseases and risk for depression in old age: a meta-analysis of published literature. *Ageing research reviews,* 9**,** 131-141.

INAMURA, K., TSUNO, N., SHINAGAWA, S., NAGATA, K. & NAKAYAMA, K. 2015. Correlation between cognition and symptomatic severity in patients with late-life somatoform disorders. *Aging and Mental Health,* 19**,** 169-174.

IZAL, M., MONTORIO, I., NUEVO, R., PEREZ-ROJO, G. & CABRERA, I. 2010. Optimising the diagnostic performance of the Geriatric Depression Scale. *Psychiatry research,* 178**,** 142-146.

JONGENELIS, L., EISSES, A., POT, A., BEEKMAN, A. & RIBBE, M. 2002. Depression in long term care facilities: Validation and reliability of the geriatric depression scale in a dutch nursing home population. *Gerontologist,* 42**,** 256-256.

JOUBERT, L., LEE, J., MCKEEVER, U. & HOLLAND, L. 2013. Caring for depressed elderly in the emergency department: establishing links between sub-acute, primary, and community care. *Social Work in Health Care,* 52**,** 222-38.

JULIAN, L. J., GREGORICH, S. E., EARNEST, G., EISNER, M. D., CHEN, H., BLANC, P. D., YELIN, E. H. & KATZ, P. P. 2009. Screening for depression in chronic obstructive pulmonary disease. *Copd: Journal of Chronic Obstructive Pulmonary Disease,* 6**,** 452-458.

KANG, H., ZHAO, F., YOU, L., GIORGETTA, C., VENKATESH, D., SARKHEL, S. & PRAKASH, R. 2014. Pseudo-dementia: A neuropsychological review. *Annals of Indian Academy of Neurology,* 17**,** 147-154.

KATON, W. J. 2011. Epidemiology and treatment of depression in patients with chronic medical illness. *Dialogues in Clinical Neuroscience,* 13**,** 7-23.

KATON, W. J., LIN, E., RUSSO, J. & UNÜTZER, J. 2003. Increased medical costs of a population-based sample of depressed elderly patients. *Archives of General Psychiatry,* 60**,** 897-903.

KATON, W. J., SCHOENBAUM, M., FAN, M.-Y., CALLAHAN, C. M., WILLIAMS, J., HUNKELER, E., HARPOLE, L., ZHOU, X.-H. A., LANGSTON, C. & UNÜTZER, J. 2005. Cost-effectiveness of improving primary care treatment of late-life depression. *Archives of General Psychiatry,* 62**,** 1313-1320.

KATZMAN, R., BROWN, T., FUID, P., PECK, A., SCHECKTER, R. & SCHIMMEL, H. 1983. Validation of a short orientation-memory-concentration test of cognitive impairment. *American Journal of Psychiatry*, 140, 734-739.

KESSLER, R. C., BERGLUND, P., DEMLER, O., JIN, R., MERIKANGAS, K. R. & WALTERS, E. E. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry,* 62**,** 593-602.

KESSLER, R. C., MCGONAGLE, K. A., NELSON, C. B., HUGHES, M., SWARTZ, M. & BLAZER, D. G. 1994. Sex and depression in the National Comorbidity Survey. II: Cohort effects. *Journal of affective disorders,* 30**,** 15-26.

KORTE, J., BOHLMEIJER, E., CAPPELIEZ, P., SMIT, F. & WESTERHOF, G. 2012. Life review therapy for older adults with moderate depressive symptomatology: A pragmatic randomized controlled trial. *Psychological medicine,* 42**,** 1163-1173.

LEE, S. C., ET AL. 2013. The Use of the Korean Version of Short Form Geriatric Depression Scale (SGDS-K) in the Community Dwelling Elderly in Korea. *Journal of Korean Geriatric Psychiatry*.

LICHT-STRUNK, E., BEEKMAN, A. T. F., DE HAAN, M. & VAN MARWIJK, H. W. J. 2009. The prognosis of undetected depression in older general practice patients. A one year follow-up study. *Journal of affective disorders,* 114**,** 310-315.

LICHT-STRUNK, E., VAN DER KOOIJ, K. G., VAN SCHAIK, D. J., VAN MARWIJK, H. W., VAN HOUT, H. P., DE HAAN, M. & BEEKMAN, A. T. 2005. Prevalence of depression in older patients consulting their general practitioner in The Netherlands. *International Journal of Geriatric Psychiatry,* 20**,** 1013-1019.

LOCKWOOD, K. A., ALEXOPOULOS, G. S., KAKUMA, T. & VAN GORP, W. G. 2000. Subtypes of cognitive impairment in depressed older adults. *The American Journal of Geriatric Psychiatry,* 8**,** 201-208.

LUPPA, M., SIKORSKI, C., LUCK, T., EHREKE, L., KONNOPKA, A., WIESE, B., WEYERER, S., KÖNIG, H.-H. & RIEDEL-HELLER, S. 2012. Age-and gender-specific prevalence of depression in latest-life–systematic review and meta-analysis. *Journal of affective disorders,* 136**,** 212-221.

LYNESS, J. M., NOEL, T., COX, C., KING, D. A., CONWELL, Y. & CAINE, E. D. 1997. Screening for depression in elderly primary care patients: A comparison of the center for epidemiologic studies—depression scale and the geriatric depression scale. *Archives of Internal Medicine,* 157**,** 449-454.

MACKENZIE, C. S., GEKOSKI, W. L. & KNOX, V. J. 2006. Age, gender and the underutilization of mental health services: The influence of help-seeking attitudes. *Aging and Mental Health*, 10, 574-582.

MALAKOUTI, S. K., FATOLLAHI, P., MIRABZADEH, A., SALAVATI, M. & ZANDI, T. 2006. Reliability, validity and factor structure of the GDS-15 in Iranian elderly. *International Journal of Geriatric Psychiatry,* 21**,** 588-93.

MARC, L. G., RAUE, P. J. & BRUCE, M. L. 2008. Screening performance of the 15-item Geriatric Depression Scale in a diverse elderly home care population. *The American Journal of Geriatric Psychiatry,* 16**,** 914-921.

MATHERS, C. D. & LONCAR, D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine,* 3**,** e442.

MCCABE, M. P., DAVISON, T., MELLOR, D., GEORGE, K., MOORE, K. & SKI, C. 2006. Depression among older people with cognitive impairment: prevalence and detection. *International Journal of Geriatric Psychiatry,* 21**,** 633-644.

MCDOUGALL, F. A., KVAAL, K., MATTHEWS, F. E., PAYKEL, E., JONES, P. B., DEWEY, M. E., BRAYNE, C., MEDICAL RESEARCH COUNCIL COGNITIVE, F. & AGEING, S. 2007. Prevalence of depression in older people in England and Wales: the MRC CFA Study.[Erratum appears in Psychol Med. 2007 Dec;37(12):1796]. *Psychological medicine,* 37**,** 1787-95.

MEEKS, T., VAHIA, I., LAVRETSKY, H., KULKARNI, G. & JESTE, D. 2011. A tune in 'A minor' can be 'B major': A review of epidemiology, illness course, and public health implications of subthreshold depression in older adults. *Journal of Affective Disorders,* 129**,** 126-142.

MITCHELL, A. J., BIRD, V., RIZZO, M. & MEADER, N. 2010a. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: A meta-analysis of GDS30 and GDS15. *Journal of affective disorders,* 125**,** 10-17.

MITCHELL, A. J., BIRD, V., RIZZO, M. & MEADER, N. 2010b. Which version of the Geriatric Depression Scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. *The American Journal of Geriatric Psychiatry,* 18**,** 1066-1077.

MITCHELL, A. J., RAO, S. & VAZE, A. 2010c. Do primary care physicians have particular difficulty identifying late-life depression? A meta-analysis stratified by age. *Psychotherapy & Psychosomatics,* 79**,** 285-94.

MITCHELL, A. J. & SUBRAMANIAM, H. 2005. Prognosis of depression in old age compared to middle age: a systematic review of comparative studies. *American Journal of Psychiatry,* 162**,** 1588-1601.

MITCHELL, N., HEWITT, C., ADAMSON, J., PARROTT, S., TORGERSON, D., EKERS, D., HOLMES, J., LESTER, H., MCMILLAN, D., RICHARDS, D., SPILSBURY, K., GODFREY, C. & GILBODY, S. 2011. A randomised evaluation of CollAborative care and active surveillance for Screen-Positive EldeRs with sub-threshold depression (CASPER): study protocol for a randomized controlled trial. *Trials [Electronic Resource],* 12**,** 225.

MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G. & THE PRISMA GROUP. 2009. Preferred reporting items for systematic reviews and meta-analysis: The PRISMA statement. *Annals of Internal Medicine*, 151(4)

MONTGOMERY, S. A. & ASBERY, M. 1979. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134(4), 382-89

MONTORIO, I. & IZAL, M. 1996. The Geriatric Depression Scale: a review of its development and utility. *International psychogeriatrics / IPA,* 8**,** 103-12.

MYKLETUN, A., BJERKESET, O., OVERLAND, S. 2009. Levels of anxiety and depression as predictors of mortality: the HUNT study. *British Journal of Psychiatry.* 195: 118-125.

NEAL, R. M. & BALDWIN, R. C. 1994. Screening for anxiety and depression in elderly medical outpatients. *Age and Ageing,* 23**,** 461-464.

NYUNT, M. S. Z., FONES, C., NITI, M. & NG, T.-P. 2009a. Criterion-based validity and reliability of the Geriatric Depression Screening Scale (GDS-15) in a large validation sample of community-living Asian older adults. *Aging & mental health,* 13**,** 376-82.

NYUNT, M. S. Z., KO, S. M., KUMAR, R., FONES, C. C. & NG, T. P. 2009b. Improving treatment access and primary care referrals for depression in a national community-based outreach programme for the elderly. *International Journal of Geriatric Psychiatry,* 24**,** 1267-1276.

O'CONNOR, E. A., WHITLOCK, E. P., GAYNES, B. & BEIL, T. L. 2009. Screening for depression in adults and older adults in primary care: an updated systematic review.

OLIVER, M. I., PEARSON, N., COE, N. & GUNNELL, D. 2005. Help-seeking behaviour in men and wmen with common mental health problems: cross-sectional study. *The British Journal of Psychiatry*, 186(4), 297-301.

ORGANIZATION, W. H. 2001a. *The World Health Report. Mental Health: New understanding, new hope.*, WHO.

ORGANIZATION, W. H. 2001b. The World health report: 2001: Mental health: new understanding, new hope.

PETTICREW, M. P., SOWDEN, A. J., LISTER-SHARP, D. & WRIGHT, K. 2000. False-negative results in screening programmes: systematic review of impact and implications. *Health Technology Assessment,* 4.

PFEIFFER, E. 1975. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatric Society*, 23(10), 435-441.

PHELAN, E., WILLIAMS, B., MEEKER, K., BONN, K., FREDERICK, J., LOGERFO, J. & SNOWDEN, M. 2010. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *Bmc Family Practice,* 11.

PIGEON, W. R., HEGEL, M., UNÜTZER, J., FAN, M.-Y., SATEIA, M. J., LYNESS, J. M., PHILLIPS, C. & PERLIS, M. L. 2008. Is insomnia a perpetuating factor for late-life depression in the IMPACT cohort? *Sleep,* 31**,** 481.

POLENICK, C. A. 2013. Behavioral activation for depression in older adults: theoretical and practical considerations. *Association for Behavioral Analysis International,* 36**,** 35-55.

POMEROY, I. M., CLARK, C. R. & PHILP, I. 2001. The effectiveness of very short scales for depression screening in elderly medical patients†. *International Journal of Geriatric Psychiatry,* 16**,** 321-326.

PRAKASH, O., GUPTA, L. N., SINGH, V. B. & NAGARAJARAO, G. 2009. Applicability of 15-item Geriatric Depression Scale to detect depression in elderly medical outpatients. *Asian journal of psychiatry,* 2**,** 63-65.

RAIT, G., BURNS, A., BALDWIN, R., MORLEY, M., CHEW-GRAHAM, C., ST LEGER, A. & ABAS, M. 1999. Screening for depression in African-Caribbean elders. *Family practice,* 16**,** 591-595.

RAPP, M. A., DAHLMAN, K., SANO, M., GROSSMAN, H. T., HAROUTUNIAN, V. & GORMAN, J. M. 2005. Neuropsychological differences between late-onset and recurrent geriatric major depression. *American Journal of Psychiatry,* 162**,** 691-698.

RELEVO, R. 2012. Effective search strategies for systematic reviews of medical tests. *Journal of General Internal Medicine,* 27**,** 28-32.

RICHARDSON, W. S., WILSON, M. C., NISHIKAWA, J. & HAYWARD, R. S. A. 1995. The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123, A12-13

RINALDI, P., MECOCCI, P., BENEDETTI, C., ERCOLANI, S., BREGNOCCHI, M., MENCULINI, G., CATANI, M., SENIN, U. & CHERUBINI, A. 2003. Validation of the Five-Item Geriatric Depression Scale in Elderly Subjects in Three Different Settings. *Journal of the American Geriatrics Society,* 51**,** 694-698.

RODDA, J., WALKER, Z. & CARTER, J. 2011. Depression in older adults. *BMJ: British Medical Journal (Overseas & Retired Doctors Edition),* 343**,** 683-687.

RODIC, D., MEYER, A. H. & MEINISCHMIDT, G. 2015. The association between depressive symptoms and physical diseases in Switzerland: a cross-sectional general population study. *Frontier of Public Health,* 23.

SAWYER, P. 2012. Counselling Older Adults at Risk of Suicide: Recognizing Barriers, Reviewing Strategies, and Exploring Opportunities for Intervention. *Alabama Counseling Association Journal,* 38**,** 80-103.

SEITZ, D., PURANDARE, N. & CONN, D. 2010. Prevalence of psychiatric disorders among older adults in long-term care homes: a systematic review. *International Psychogeriatrics,* 22**,** 1025-1039.

SHAH, A., ODUTOYE, K. & DE, T. 2001. Depression in acutely medically ill elderly inpatients: a pilot study of early identification and intervention by formal psychogeriatric consultation. *Journal of Affective Disorders,* 62**,** 233-240.

SHAHPESANDY, H. 2005. Different manifestation of depressive disorder in the elderly. *Neuroendocrinology Letters,* 26**,** 691-5.

SHARIF, M. O., SHARIF-JANJUA, F. N., ALI, H. & AHMED, F. 2013. Systematic reviews explained: AMSTAR - how to tell the good from the bad and the ugly. *Oral Health Dental Management,* 12**,** 9-16.

SHEA, B. J., GRIMSHAW, J. M., WELLS, G. A., BOERS, M., ANDERSSON, N., HAMEL, C., PORTER, A. C., TUGWELL, P., MOHER, D. & BOILER, L. M. 2007. Development of AMSTAR: A measurement tool to assess the methodological quality of a systematic review. *BMC Medical Research Methodology*, 15, 7-10.

SHEEHAN, B. & BANERJEE, S. 1999. Review: somatization in the elderly. *International Journal of Geriatric Psychiatry,* 14**,** 1044-1049.

SMALBRUGGE, M., JONGENELIS, L., POT, A. M., BEEKMAN, A. T. & EEFSTING, J. A. 2008. Screening for depression and assessing change in severity of depression. Is the Geriatric Depression Scale (30-, 15-and 8-item versions) useful for both purposes in nursing home patients? *Aging and Mental Health,* 12**,** 244-248.

SMALL, G.W. 1991. Recognition and treatment of depression in the elderly. *The Journal of Clinical Psychiatry,* 52, 11-22.

SNEED, J. R., CULANG-REINLIEB, M. E., BRICKMAN, A. M., GUNNING-DIXON, F. M., JOHNERT, L., GARCON, E. & ROOSE, S. P. 2011. MRI signal hyperintensities and failure to remit following antidepressant treatment. *Journal of affective disorders,* 135**,** 315-20.

SNEED, J. R., KEILP, J. G., BRICKMAN, A. M. & ROOSE, S. P. 2008a. The specificity of neuropsychological impairment in predicting antidepressant non-response in the very old depressed. *International Journal of Geriatric Psychiatry,* 23**,** 319-23.

SNEED, J. R., RINDSKOPF, D., STEFFENS, D. C., KRISHNAN, K. R. R. & ROOSE, S. P. 2008b. The vascular depression subtype: evidence of internal validity. *Biological Psychiatry,* 64**,** 491-7.

SOLHAUG, H. I., ROMULD, E. B., ROMILD, U. & STORDAL, E. 2012. Increased prevalence of depression in cohorts of the elderly: an 11-year follow-up in the general population–the HUNT study. *International Psychogeriatrics,* 24**,** 151.

SPEER, D. C. & SCHNEIDER, M. G. 2003. Mental health needs of older adults and primary care: Opportunity for interdisciplinary geriatric team practice. *Clinical Psychology: Science and Practice,* 10**,** 85-101.

SPITZR, R. L., KROENKE, K., WILLIAMS, J. B. W. & THE PATIENT HEALTH QUESTIONNAIR PRIMARY CARE STUDY GROUP. 1999. Validation and utility of a self-reprot version of the PRIME-MD. *JAMA*, 1999, 282(18), 1737-1744.

STERNE, J. A. C., SUTTON, A. J., IOANNIDIS, J. P. A., TERRIN, N., JONES, D. R., LAU, J., CARPENTER, J., RUCKER, G., HARBOR, R. M., SCHMID, C. H. S., TETZLAFF, J., DEEKS, J. J., PETERS, J., MACASKILL, P., SCHWARZER, G., DUVAL, S., ALTMAN, D. G., MOHER, D. & HIGGINS, J. P. T. 2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analysis of randomised controlled trials. *BMJ,* 342.

THOMSON REUTERS. 2016, EndNote Web, *EndNote,* Version X7.

TSUNO, N. & HOMMA, A. 2009. What is the association between depression and Alzheimer's disease? *Expert review of neurotherapeutics,* 9**,** 1667-1676.

UNUTZER, J. & SCHOENBAUM, M. 2009. Healthcare costs associated with depression in medically ill fee-for-service Medicare participants. *Journal of the American Geriatric Society,* 57**,** 375-584.

VAN MARWIJK, H. W., WALLACE, P., DE BOCK, G. H., HERMANS, J., KAPTEIN, A. & MULDER, J. 1995. Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the geriatric depression scale. *British Journal of General Practice,* 45**,** 195-195.

VAN MARWIJK, H. W. J., ADER, H., DE HAAN, M. & BEEKMAN, A. 2008. Primary care management of major depression in patients aged >55 years: Outcome of a randomised clinical trial. *British Journal of General Practice,* 58**,** 680-686.

VAN SOMEREN, E. 2000. Circadian and sleep disturbances in the elderly. *Experimental gerontology,* 35**,** 1229-1237.

VAN'T VEER-TAZELAAR, N., VAN MARWIJK, H., VAN OPPEN, P., NIJPELS, G., VAN HOUT, H., CUIJPERS, P., STALMAN, W. & BEEKMAN, A. 2006. Prevention of anxiety and depression in the age group of 75 years and over: a randomised controlled trial testing the feasibility and effectiveness of a generic stepped care programme among elderly community residents at high risk of developing anxiety and depression versus usual care. *BMC Public Health,* 18**,** 186.

WALTHER, S., HOFLE, O., FEDERSPIEL, A., HORN, H., HUGLI, S., WIEST, R., STRIK, W. & MULLER, T. J. 2012. Neural correlates of disbalanced motor control in major depression. *Journal of Affective Disorders,* 136**,** 124-133.

WANCATA, J., ALEXANDROWICZ, R., MARQUART, B., WEISS, M. & FRIEDRICH, F. 2006. The criterion validity of the Geriatric Depression Scale: a systematic review. *Acta Psychiatrica Scandinavica,* 114**,** 398-410.

WATSON, L. C., LEWIS, C. L., KISTLER, C. E., AMICK, H. R. & BOUSTANI, M. 2004. Can we trust depression screening instruments in healthy ‘old-old’adults? *International Journal of Geriatric Psychiatry,* 19**,** 278-285.

WATSON, L. C. & PIGNONE, M. P. 2003. Screening accuracy for late-life depression in primary care: a systematic review. *Journal of Family Practice,* 52**,** 956-956.

WEEKS, S. K., MCGANN, P. E., MICHAELS, T. K. & PENNINX, B. W. 2003. Comparing Various Short-Form Geriatric Depression Scales Leads to the GDS-5/15. *Journal of Nursing Scholarship,* 35**,** 133-137.

WEYERER, S., EIFFLAENDER-GORFER, S., KÖHLER, L., JESSEN, F., MAIER, W., FUCHS, A., PENTZEK, M., KADUSZKIEWICZ, H., BACHMANN, C. & ANGERMEYER, M. C. 2008. Prevalence and risk factors for depression in non-demented primary care attenders aged 75 years and older. *Journal of affective disorders,* 111**,** 153-163.

WHITING, P., RUTJES, A. W., REITSMA, J. B., BOSSUYT, P. M. & KLEIJNEN, J. 2003. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3, 25

WHITING, P. F., RUTJES, A. W., WESTWOOD, M. E., MALLETT, S., DEEKS, J. J., REITSMA, J. B., LEEFLANG, M. M., STERNE, J. A. & BOSSUYT, P. M. 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine,* 155**,** 529-536.

WHOOLEY, M. A., STONE, B. & SOGHIKIAN, K. 2000. Randomized Trial of Case-Finding for Depression in Elderly Primary Care Patients. *Journal of general internal medicine,* 15**,** 293-300.

WILSON, J. M. G. & JUNGNER, G. 1968. Principles and Practice of Screening for Disease. *World Health Organization Public Health Papers,* 34.

WIN, S., PARAKH, K., EZE-NILLAM, C. M., GOTTDIENER, J. S., KOP, W. J. & ZIEGEISTEIN, R. C. 2011. Depressive symptoms, physical inactivity and risk of cardiovascular mortality in older adults: the cardiovascular health study. *Heart,* 97**,** 500-505.

WONGPAKARAN, N., WONGPAKARAN, T. & VAN REEKUM, R. 2013. The use of GDS-15 in detecting MDD: a comparison between residents in a Thai long-term care home and geriatric outpatients. *Journal of clinical medicine research,* 5**,** 101.

YESAVAGE, J. A. 1986. The use of self-rating depression scales in the elderly. *Poon, Leonard W [Ed]***,** 213-217.

YESAVAGE, J. A., BRINK, T., ROSE, T. L., LUM, O., HUANG, V., ADEY, M. & LEIRER, V. O. 1983. Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research,* 17**,** 37-49.

YESAVAGE, J. A. & SHEIKH, J. I. 1986. Geriatric Depression Scale (GDS) recent evidence and development of a shorter version. *Clinical Gerontologist,* 5.1-2**,** 165-173.

ZUNG, W.W. 1965. A self-rating depression scale. *Archives of General Psychiatry*, 12, 63-70.

APPENDICES

APPENDIX 1

The National Screening Committee guidance: Criteria for appraising the viability, effectiveness and appropriateness of a screening programme (Public Health England,

1. The condition

1. The condition should be an important health problem as judged by its frequency and/or severity. The epidemiology, incidence, prevalence and natural history of the condition should be understood, including development from latent to declared disease and/or there should be robust evidence about the association between the risk or disease marker and serious or treatable disease.
2. All the cost-effective primary prevention interventions should have been implemented as far as practicable.
3. If the carriers of a mutation are identified as a result of screening the natural history of people with this status should be understood, including the psychological implications.

2. The test

4. There should be a simple, safe, precise and validated screening test.
5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.
6. The test, from sample collection to delivery of results, should be acceptable to the target population.
7. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.
8. If the test is for a particular mutation or set of genetic variants the method for their selection and the means through which these will be kept under review in the programme should be clearly set out.

3. The intervention

9. There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care. Evidence relating to wider benefits of screening, for example those relating to family members, should be taken into account where available. However, where there is no prospect of benefit for the individual screened then the screening programme shouldn't be further considered.
10. There should be agreed evidence based policies covering which individuals should be offered interventions and the appropriate intervention to be offered.

4. The screening programme

11. There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (such as Down's syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.
12. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.
13. The benefit gained by individuals from the screening programme should outweigh any harms for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.
14. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically

balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

5. Implementation criteria

15. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.
16. All other options for managing the condition should have been considered (such as improving treatment or providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.
17. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.
18. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.
19. Evidence-based information, explaining the purpose and potential consequences of screening, investigation and preventative intervention or treatment, should be made available to potential participants to assist them in making an informed choice.
20. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.

6. References

- Department of Health, Screening of pregnant women for hepatitis B and immunisation of babies at risk. London: Dept of Health, 1998 (Health Service Circular : HSC 1998/127).
- Wilson JMG, Jungner G. Principles and practice of screening for disease. Public Health Paper Number 34. Geneva: WHO, 1968.
- Cochrane AL. Holland WW. Validation of screening procedures. Br Med Bull. 1971, 27, 3.
- Sackett DL, Holland WW. Controversy in the detection of disease. Lancet 1975;2:357-9.
- Wald NJ (Editor). Antenatal and Neonatal screening. Oxford University Press, 1984.
- Holland WW, Stewart S. Screening in Healthcare. The Nuffield Provincial Hospitals Trust, 1990.
- Gray JAM. Dimensions and definitions of screening. Milton Keynes: NHS Executive Anglia and Oxford, Research and Development.
- Angela Raffle/Muir Gray Screening Evidence and Practice, Oxford University Press 2007.

APPENDIX 2

Protocol for the systematic review of the diagnostic accuracy of brief versions of the GDS

PROSPERO dataset guidance

Title & timescale

1. Review title*

The diagnostic accuracy and validity of different versions of the Geriatric Depression Scale (GDS) for older adults

2. Original language

English

3. Anticipated or actual start date*

February 2014

4. Anticipated completion date*

July 2014

5. Stage of review at time of submission*

Searches about to be performed

Review team details

6. Named contact*

Claire Pocklington

7. Named contact email*

cp945@york.ac.uk

8. Named contact address

Mental Health & Addiction Research Group
Department of Health Sciences
University of York
Heslington
York
YO10 5DD

9. Named contact phone number

+44 1904 321112

10. Organisational affiliation of the review*

Department of Health Sciences, University of York
http://www.york.ac.uk/healthsciences/research/
Centre for Review and Dissemination, University of York
http://www.york.ac.uk/inst/crd/
Hull York Medical School, University of York
http://www.hyms.ac.uk/

11. **Review team members & their organisational affiliations**

Title, first name, last name of all working directly on review =
Dr Claire Pocklington , Research Fellow[1]; Dr Dean McMillan, Senior Lecturer[1, 2]

[1] Department of Health Sciences, University of York
[2] Hull York Medical School, University of York

12. Funding sources/sponsors*

None

13. Conflict of interest*

None known

14. Collaborators

Names and affiliation of any individuals or organisations working on the review not listed in team.

None

Review methods

15. Review question(s)* *State the question(s) to be addressed / review objectives. Please complete a separate box for each question.*

1. What is the diagnostic test accuracy and validity of the Geriatric Depression Scale for older adults?

16. Searches*

Details of sources to be searched, and any restrictions (eg. language and publication period). Full search strategy not required but may be added as a link or attachment.

Searches of published and unpublished literature will be performed to identify diagnostic test accuracy studies of the geriatric depression scale in older adults.

The following databases will be searched: MEDLINE, MEDLINE In-Process, PsycINFO, EMBASE, Cumulative Index to Nursing & Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), and the Health Technology Assessment (HTA) database. The search strategy will involve no language or date limits and no filters on study design. Unpublished and grey literature will also be searched. The reference list of all studies included will be examined to identify other relevant studies for inclusion. Experts will be contacted if required to locate further studies.

17. URL to search strategy

Insert link or pdf but recognise that this means your search strategy is publically accessible

18. Condition or domain being studied*

Short description of disease being studied

Depression is the commonest mental illness in those aged over 65 years. Despite this it is often under-recognised and consequently under treated. Incidence and prevalence are expected to rise in the future due to increase in both life expectancy and population size. Depression is often more difficult to diagnosis in older adults due to differences in symptomatology and the comorbidity of physical illnesses. Better diagnostic strategies would lead to improved clinical and economic outcomes.

19. Participants/population*

Summary criteria for the participants or populations being studied, including details of inclusion and exclusion criteria

The population of interest is older adults. Older adults are classified as 55 years of age or older.

20. Intervention(s), exposure(s)*

Full & clear descriptions of the nature of the interventions or exposures to be reviewed, including details of inclusion and exclusion criteria

The intervention is the Geriatric Depression Screen (GDS). Several different versions exist according to the number of items included. All different versions of the GDS will be included e.g. GDS-30, GDS-15, GDS-12, etc. There will be no exclusions in regarding the administration of the GDS.

21. Comparator(s)/control*

Where relevant, give details of the alternatives against which the main topic of the review will be compared, including details of inclusion and exclusion criteria.

The comparator will be gold standard diagnostic interviews developed from the Diagnostic and Statistical Manual (DSM) or International Classification of Disease (ICD) diagnostic criteria for depression. Examples of gold standard diagnostic interview instruments include CIDI, CIS, DIS, GMS, MINI, PAS, Prime, PSE, SADS, SCAN and SCID. There will be no exclusions in regarding the administration of the gold standard diagnostic interview.

22. Types of study to be included initially*

Include details of study designs to be included. If there are no restrictions on the type of study to be included, this should be stated.

There will be no restrictions regarding the type of study design included. Studies selected will ideally by randomised controlled trials. Where appropriate non-randomised controlled trails will be included. The quality of all studies included will be assessed using the Cochrane Collaborations' tool for assessing the risk of bias.

23. Context

Summary details of the setting and other characteristics which help define the inclusion/exclusion criteria.

There will be no exclusion criteria regarding country or setting for included studies. Studies from primary care settings, secondary care settings and non-clinical settings will be included.

24. Primary outcome(s)*

Give most important primary outcomes

2x2 contingency tables will be used to calculate diagnostic test accuracy estimates; this will include measures of sensitivity and specificity, positive and negative likelihood ratios, and diagnostic odds ratios. Only studies where data to construct a 2x2 contingency table can be extracted will be included.

25. Secondary outcome(s)*

List any additional outcomes that will be addressed, if there are no secondary outcomes, enter None.
None

26. Data extraction (selecting & coding)*

Give procedure for selecting studies for review and extracting data, including the number of researchers involved and how discrepancies will be resolved. List the data to be extracted.

Selected studies will fulfil search criteria that will be outlined by a PICO checklist. Studies fulfilling the search criteria will be identified by one researcher and initial selection for inclusion will be based on the abstract of the paper. Any uncertainties regarding paper inclusion will be discussed with another researcher. A third reviewer will be involved if there is any disagreement. If data is missing from any included studies the authors will be contacted for further information.

The following data will be extracted to a standardised proforma:

1) descriptive characteristics of the sample and setting (country, setting, age of sample, gender of sample, sample size, proportion depressed);
2) descriptive characteristics of the screening tool used (mode of administration, who administered, language);
3) descriptive characteristics of the gold standard (type of gold standard, whether DSM or ICD diagnoses);
4) quality assessment criteria (see below);
5) data to construct 2x2 contingency tables

27. Risk of bias (quality assessment)*

State whether and how risk of bias will be assessed, how the quality of individual studies will be assessed, and whether and how this will influence the planned synthesis.

The QUADAS-II will be used to assess study quality (Whiting et al., 2011). If a sufficient number of eligible studies are identified quality criteria will be used to inform sensitivity analyses. If there are a sufficient number of studies to conduct a sensitivity analysis we will run an analysis that will exclude studies that did not ensure blinding (of the results of the index test to the reference test or vice versa).

Reference: Whiting P. F., Rutjes A. W. S., Westwood M. E., Mallett S., Deeks J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Patrick, M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155,* 529-536.

28. Strategy for data synthesis*

Give planned general approach to be used, eg. whether data will be aggregate or at the level of individual participants, and whether a quantitative or narrative (descriptive) synthesis is planned. Where appropriate a brief outline of analytic approach should be given.

OVERVIEW OF ANALYSIS STRATEGY

It is hoped that there will be a sufficient number of comparable studies to perform a diagnostic meta-analysis.

CALCULATION OF DIAGNOSTIC TEST ACCURACY VALUES

As mentioned sensitivity and specificity, positive and negative likelihood rations and diagnostic odds ratios will be calculated from 2x2 contingency tables. 95% confidence intervals will be included.

ASSESSING HETEROGENEITY

Between study heterogeneity will be assessed using I^2 statistic of the pooled diagnostic odds ratio.

PRE-PLANNED COMPARISONS

The diagnostic accuracy of different item number versions of the GDS will be compared.

CRITERIA FOR CONDUCTING META-ANALYSES AND DETAILS OF META-ANALYSIS

A minimum of four studies is required to conduct a diagnostic meta-analysis. If there are four comparable studies and if heterogeneity in not deemed to be substantial bivariate diagnostic meta-analysis will be used to generate pooled estimates of sensitivity and specificity. Summary Receiver Operating Characteristics (sROC) will be calculated to produce 95% confidence interval ellipses within ROC space.

ANALYSIS OF PUBLICATION BIAS

Depending on the number studies identified funnel plots will be constructed to examine the potential role of publication bias.

29. Analysis of subgroups or subsets*

Give any planned analysis of subsets within the review. 'None planned' is a valid response

If there are a sufficient number of studies subgroup analysis of setting (i.e. primary care, secondary care, non-clinical), presence of cognitive impairment and age will be performed.

General information

30. Type of review

Diagnostic test accuracy systematic review

31. Language

English

32. Country

England

33. Other registration details

List of places where the review is registered

None

34. Reference and/or URL for published protocol

Give citation & link or upload pdf of published protocol in CRD pdf format

Will be added

35. Dissemination plans

Brief details about communicating essential messages to appropriate audiences

To be published in peer-reviewed journal

Presented findings at conferences focusing on older adult mental health, primary care or secondary care

36. Key words

One word per box, create separate new box for each new word

Diagnostic meta-analysis, screening, depression, diagnostic test accuracy, mental health, geriatric depression scale

37. Details of any existing review of same topic by same authors: (not a required field)

38. Current review status*

Should be updated when review is completed and when it's published – select from drop down box

39. Any other information

Provide any further information relevant to the registration of the review

40. Details of final report/publication(s)

Leave empty until review published. Give full citation for the report and URL where available

APPENDIX 3

Search strategy for the systematic review of the diagnostic accuracy of brief versions of the GDS

MEDLINE search strategy:

1. older$.ti,ab.
2. elder$.ti,ab.
3. geriatri$.ti,ab.
4. 1 or 2 or 3
5. Limit 4 to (humans and yr="1982-Current")
6. exp Depression/
7. exp Depressive Disorder/
8. (depressive or depression or depressed).ti,ab.
9. (melancholi$ or dysphori$ or dysthymi$).ti,ab.
10. 6 or 7 or 8 or 9
11. Limit 10 to (humans and yr="1982-Current")
12. "geriatric depression scale".ti,ab.
13. "GDS$".ti,ab.
14. 12 or 13
15. Limit 14 to (humans and yr="1982-Current")
16. 5 and 11 and 15

APPENDIX 4

Table of excluded studies from the systematic review of the diagnostic accuracy of brief versions of the GDS

Study	Reason for exclusion	Further information
Abolfotouh et al. (2001)	Insufficient information to construct a 2*2 table	
Adams et al. (2011)	Not major depression	
Agrell et al. (1989)	Inadequate reference test	
Ali et al. (2005)	Insufficient information to construct a 2*2 table Not major depression	Reference test not applied to all of sample
Allan et al. (2013)	Inadequate reference test	
Allen-Burge et al. (1994)	Inadequate reference test	
Almeida et al. (1999) *Reliability*	Insufficient information to construct a 2*2 table	
Appelros et al. (2004)	Inadequate reference test	
Arean et al. (2001)	Insufficient information to construct a 2*2 table	
Arvanti et al. (2005)	Insufficient information to construct a 2*2 table	
Baillon et al. (2014)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Baker et al. (1991)	Inadequate reference test	
Baker et al. (1997)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Balogun et al. (2010)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Balogun et al. (2011)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Banerjee et al. (2008)	Insufficient information to construct a 2*2 table	
Bidzan et al. (2002)	Inappropriate index test	GDS 30-item version used
Bieliauskas et al. (2011)	Inappropriate index test	GDS 30-item version used
Blancarte et al. 1993)	Inadequate reference test	
Brodaty et al. (1996)	Inappropriate index test	GDS 30-item version used
Brody et al. (2001)	Insufficient information to construct a 2*2 table	
Burke et al. (1989)	Inappropriate index test	GDS 30-item version used
Burke et al. (1991)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Burke et al. (1994)	Inadequate reference test	
Buz Delgado et al. (1996)	Inadequate reference test	
Calleo et al. (2011) ??	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Carrete et al. (2001)	Inadequate reference test	
Chang et al. (2011)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Cheng et al. (2004)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Cheng et al. (2005)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Ciadella et al. (1992)	Inappropriate index test	GDS 30-item version used
Clement et al. (1999)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Costa et al. (2006)	Inappropriate index test	GDS 30-item version used

Study	Reason for exclusion	Further information
Cwikel et al. (1989)	Inadequate reference test	
De Azpiazo et al. (1988)	Inappropriate index test	GDS 30-item version used
De Sousa et al. (2007)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Djernes et al. (2004)	Inadequate reference test	
Ertan et al. (2005)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Ertan et al. (2009)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Espino et al. (1996)	No reference test	Previous clinical diagnosis of depression used as reference
Evans et al. (1993)	Inappropriate index test	GDS 30-item version used
Falck et al. (1999)	Inappropriate index test	GDS 30-item version used
Fernandez-San Martin et al. (2002)	Inappropriate index test	GDS 30-item version used
Ferraro et al. (1997)	Inadequate reference test	
Filbert et al. (2012)	Inappropriate index test	GDS 30-item version used
Galaria et al. (2000)	Inadequate reference test	
Gerritsen et al. (2007)	Duplication in sample	Jongenelis et al. 2007
Gilley et al. (1997)	Inadequate reference test	
Graham et al. (2004)	Unclear reference test Insufficient information to construct a 2*2 table	
Gottfries et al. (1997)	Inadequate reference test	
Greenberg et al. (2004)	Inadequate reference test	
Harper et al. (1990)	Inappropriate index test Inadequate reference test	GDS 30-item version used
Harralson et al. (2002)	Insufficient information to construct a 2*2 table	
Hedberg et al. (2010)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Heisel et al. (2003)	Inadequate reference test	
Heisel et al. (2010)	Suicide.	
Heiser et al. (2004)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Hoyl et al. (1999)	Not major depression	
Ihara et al. (1998)	Insufficient information to construct a 2*2 table	
Jackson et al. (1993)	Inappropriate index test	GDS 30-item version used
Johnson et al. (1995)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Kafonek et al. (1989)	Inadequate reference test	
Kallenbach et al. (2006)	Not a study of GDS diagnostic accuracy	

Study	Reason for exclusion	Further information
Kee et al. (1996)	Conference proceedings. Insufficient information.	
Khattri et al. (2006)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Noyes et al. (2011)	Insufficient information to construct a 2*2 table	
Oiji et al. (1998)	Insufficient information to construct a 2*2 table	Purpose of study was not to establish diagnostic accuracy of GDS
Olivera et al. (2011)	Inadequate reference test	
O'Neill (2002)	Inappropriate index test	GDS 30-item version used
O'Riordan et al. (1990)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Ortega Orcos et al. (2007)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Paradela et al. (2005)	Not major depression.	All mood disorders included.
Parmalee et al. (1989)	Inadequate reference test	
Pendelton et al. (2008)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Pocinho et al. (2009)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Pomeroy et al. (2001)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Prado-Jean et al. (2011)	Insufficient information to construct a 2*2 table	
Prakash et al. (2009)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Ramos Brieva et al. (1991)	Inappropriate index test	GDS 30-item version used
Ramsay et al. (1991)	Inappropriate index test	GDS 30-item version used
Rao et al. (2001)	Inadequate reference test	
Rapp et al. (1988)	Inappropriate index test	GDS 30-item version used
Rinaldi et al. (2003)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Robison et al. (2002)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Roeckeman et al. (2012)	Not major depression.	
Roger et al. (2009)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Rovner et al. (1997)	Inappropriate index test	GDS 30-item version used
Royall et al. (1996)	Inadequate reference test	
Rubin et al. (2001)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria
Sagduyu et al. (1997)	Inappropriate index test	GDS 30-item version used
Sanchez-Garcia et al. (2008)	Inadequate reference test	
Schreiner et al. (2003)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Shah et al. (1992)	Inappropriate index test	GDS 30-item version used
Sharma et al. (2011)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Sharma et al. (2013)	Inappropriate index test	GDS 30-item version used
Singh et al. (2013)	Insufficient information to construct a 2*2 table	

Study	Reason for exclusion	Further information
Smalbrugge et al. (2005)	Insufficient information to construct a 2*2 table	
Smalbrugge et al. (2008)	Insufficient information to construct a 2*2 table	
Snowdon (1990)	Inappropriate index test	GDS 30-item version used
Soety et al. (2001)	Inadequate reference test	
Sokoya et al. (2003)	Insufficient information to construct a 2*2 table	
Sorensen et al. (1998)	Inadequate reference test	Clinical diagnosis according to ICD diagnostic criteria
Tang et al. (2003)	Not major depression	All mood disorders included
Tang et al. (2004) *detecting*	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Tang et al. (2005)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Teixeira et al. (2009)	Inappropriate index test	GDS 30-item version used
Teng et al. (2008)	Insufficient information to construct a 2*2 table	
Thompson et al. (2011)	Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Tumas et al. (2008)	Inadequate reference test Sample does not meet age criterion	Sample includes individuals less than 55 years of age
Van Warwijk et al. (1997)	Inappropriate index test	GDS 30-item version used
Vargas et al. (2007)	Inappropriate index test	GDS 30-item version used
Varma et al. (2008)	Inappropriate index test	GDS 30-item version used
Watson et al. (2004)	Inappropriate index test	GDS 30-item version used
Weintraub et al. (2004)	Insufficient information to construct a 2*2 table	
Weintraub et al. (2006)	Insufficient information to construct a 2*2 table	
Weintraub et al. (2007)	Not major depression	All mood disorders included
Wichowicz et al. (2004)	Inappropriate index test	GDS 30-item version used
Williams et al. (2009)	Sample does not meet age criterion Insufficient information to construct a 2*2 table	Sample includes individuals less than 55 years of age
Williams et al. (2012) *A comparison*	Overlap in sample	Copy of Williams 09 sample
Williams et al. (2012) *short and sweet*	Overlap in sample Sample does not meet age criterion	Copy of Williams 09 sample Sample includes individuals less than 55 years of age
Wonkpakaran et al. (2013) *level*	Inappropriate index test	GDS 30-item version used
Wonkpakaran et al. (2014)	Inappropriate index test	GDS 30-item version used
Wynkoop et al. (1999)	Inadequate reference test Insufficient information to construct a 2*2 table	
Yang et al. (2012)	Inappropriate index test	GDS 30-item version used
Yusuf et al. (2013)	Insufficient information to construct a 2*2 table	
Zalsman et al. (2008)	Inadequate reference test	Clinical diagnosis according to DSM diagnostic criteria

APPENDIX 5

Journal article published in The Internal Journal of Geriatric Psychiatry

Available at: http://onlinelibrary.wiley.com/doi/10.1002/gps.4407/full

APPENDIX 6

Protocol for the systematic review of the clinical effectiveness of screening for depression in older adults

PROSPERO dataset guidance

Title & timescale

1) Review title*

The clinical effectiveness of screening for depression in older adults

2) Original language

English

3) Anticipated or actual start date*

May 2014

4) Anticipated completion date*

July 2014

5) Stage of review at time of submission*

Review team details

6) Named contact*

Claire Pocklington

7) Named contact email*

cp945@york.ac.uk

8) Named contact address

Mental Health & Addiction Research Group
Department of Health Sciences
University of York
Heslington
York
YO10 5DD

9) Named contact phone number

+44 1904 321112

10)Organisational affiliation of the review*

Department of Health Sciences, University of York
http://www.york.ac.uk/healthsciences/research/
Centre for Review and Dissemination, University of York
http://www.york.ac.uk/inst/crd/
Hull York Medical School, University of York
http://www.hyms.ac.uk/

11) Review team members & their organisational affiliations

Title, first name, last name of all working directly on review =
Dr Claire Pocklington , Research Fellow[1]; Dr Dean McMillan, Senior Lecturer[1, 2]

[1] Department of Health Sciences, University of York
[2] Hull York Medical School, University of York

12)Funding sources/sponsors*

None

13)Conflict of interest*

None known

14)Collaborators

Names and affiliation of any individuals or organisations working on the review not listed in team.

None

Review methods

15) **Review question(s)*** *State the question(s) to be addressed / review objectives. Please complete a separate box for each question.*

What is the clinical effectiveness of screening for depression in older adults?

16)Searches*

Details of sources to be searched, and any restrictions (e.g. language and publication period). Full search strategy not required but may be added as a link or attachment.

Searches of published and unpublished literature will be performed to identify studies that address clinical effectiveness in terms of patient outcome. The following databases will be searched: MEDLINE, MEDLINE In-Process, PsycINFO, EMBASE, Cumulative Index to Nursing & Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), and the Health Technology Assessment (HTA) database. No date limit will be applied to the search strategy. The search strategy will involve no language limits or filters on study design. Unpublished and grey literature will also be searched. The reference list of all studies included will be examined to identify other relevant studies for inclusion. Experts will be contacted if required to gain further information or locate further studies.

17)URL to search strategy

Insert link or pdf but recognise that this means your search strategy is publically accessible

18)Condition or domain being studied*

Short description of disease being studied

Depression is the commonest mental illness in those aged over 65 years. Despite this it is often under-recognised and consequently under treated. Incidence and prevalence are expected to rise in the future due to increase in both life expectancy and population size. Depression is often more difficult to diagnosis in older adults due to differences in symptomatology and the comorbidity of physical illnesses. Depression screening could lead to improvement in rates of detection and diagnosis as well as treatment and associated clinical outcomes. The clinical effectiveness of depression screening for older adults is not well established.

19)Participants/population*

Summary criteria for the participants or populations being studied, including details of inclusion and exclusion criteria

The population of interest is older adults. Older adults are classified as 55 years of age or older.

20)Intervention(s), exposure(s)*

Full & clear descriptions of the nature of the interventions or exposures to be reviewed, including details of inclusion and exclusion criteria

The intervention of interest is the process and results of screening for depression, where the results of which will influence subsequent management.

The intervention of interest is the process of screening for depression, which the results of will direct management. The intervention group will receive some form of enhanced care secondary to a screening process.

Studies comparing screened people against non-screened people, where randomisation has occurred first, are preferable because they provide the most direct evidence of the impact screening has on clinical effectiveness. It is expected that the number of such studies identified will be low in number and for that reason eligible studies for inclusion will include studies where both arms have undergone screening but the screening results will only be disclosed for the intervention group.

No restrictions will be made in terms of mode of screening administration (e.g. telephone or face-to-face), the person administering the measure (e.g. clinician, researcher or self-administered), or setting (e.g. primary care, secondary care or a none-clinical setting).

21)Comparator(s)/control*

Where relevant, give details of the alternatives against which the main topic of the review will be compared, including details of inclusion and exclusion criteria.

The comparator will be 'care as usual' (CAU). This could include no screening for the control group or the screening results not being disclosed to the individual or health professionals responsible for their care. Screening results of the control group should not influence normal practice in terms of identification (diagnosis) or management. CAU will involve no enhancement from routine care being administered.

22)Types of study to be included initially*

Include details of study designs to be included. If there are no restrictions on the type of study to be included, this should be stated.

Ideally, included studies will be randomised controlled trails. Where appropriate non-randomised controlled trials will be included.

23)Context

Summary details of the setting and other characteristics which help define the inclusion/exclusion criteria.

There will be no exclusion criteria regarding country or setting for included studies.

24)Primary outcome(s)*

Give most important primary outcomes

The primary outcome will be symptom improvement. Symptom improvement will be change in symptom count or rating of severity as measured by a rating or screening tool, e.g. Geriatric Depression Scale (GDS), Hamilton Rating Scale for Depression (HAMD), Montgomery-Asberg Depression Rating Scale (MADRS).

25)Secondary outcome(s)*

List any additional outcomes that will be addressed, if there are no secondary outcomes, enter None.

None

26)Data extraction (selecting & coding)*

Give procedure for selecting studies for review and extracting data, including the number of researchers involved and how discrepancies will be resolved. List the data to be extracted.

Search criteria will be outlined by a PICO checklist. Studies fulfilling the search criteria will be identified by one reviewer and initial selection for inclusion will be based on the title and abstract of the paper. Any uncertainties regarding paper inclusion will be discussed with another reviewer. A third reviewer will be involved if there is any disagreement. If data is missing from any included studies the authors will be contacted so the information can be requested.

The following data will be extracted to a standardised proforma for each review:
Author, date of publication
Country, language
Study design
Descriptive characteristics of the setting
Descriptive characteristics of the sample – age, proportion female, ethnicity, cognitive status, physical comorbidity
Descriptive characteristics of depression screening method used – tool used, administration mode, administered by who
Outcome measure used

27)Risk of bias (quality assessment)*

State whether and how risk of bias will be assessed, how the quality of individual studies will be assessed, and whether and how this will influence the planned synthesis.

The quality of all studies included will be assessed for risk of bias using the Cochrane risk of bias assessment tool.

28)Strategy for data synthesis*

Give planned general approach to be used, eg. whether data will be aggregate or at the level of individual participants, and whether a quantitative or narrative (descriptive) synthesis is planned. Where appropriate a brief outline of analytic approach should be given.

OVERVIEW OF ANALYSIS STRATEGY

It is hoped that there will be a sufficient number of comparable studies to perform a meta-analysis. If a sufficient number of studies are not identified a narrative summary will be performed.

ASSESSING HETEROGENEITY

Between study heterogeneity will be assessed using the I^2 statistic of pooled effect size.

ANALYSIS OF PUBLICATION BIAS

Depending on the number studies identified funnel plots will be constructed to examine the potential role of publication bias.

29)Analysis of subgroups or subsets*

Give any planned analysis of subsets within the review. 'None planned' is a valid response

If there are a sufficient number of studies subgroup analysis of setting (i.e. primary care, secondary care, non-clinical) and age will be performed. If there are a sufficient number of studies non-randomised controlled studies will be excluded for a sensitivity analysis to be performed.

General information

30)Type of review

Systematic review of clinical and cost effectiveness of depression screening

31)Language

English

32)Country

England

33)Other registration details

List of places where the review is registered

None

34)Reference and/or URL for published protocol

Give citation & link or upload pdf of published protocol in CRD pdf format

35)Dissemination plans

Brief details about communicating essential messages to appropriate audiences

To be published in peer-reviewed journal
Presented findings at conferences focusing on older adult mental health, primary care or secondary care

36)Key words

One word per box, create separate new box for each new word
Screening, case finding, depression, mental health, clinical effectiveness, cost effectiveness, outcomes

37)Details of any existing review of same topic by same authors: (not a required field)

38)Current review status*

Should be updated when review is completed and when it's published – select from drop down box

39)Any other information

Provide any further information relevant to the registration of the review

40)Details of final report/publication(s)

Leave empty until review published. Give full citation for the report and URL where available

APPENDIX 7

Search strategy for the systematic review of the clinical effectiveness of screening for depression in older adults

MEDLINE search strategy:

1. exp Mass Screening/

2. "casefinding".ti,ab.

3. "case finding".ti,ab.

4. screen$.ti,ab.

5. detect$.ti,ab.

6. predict$.ti,ab.

7. aware$.ti,ab.

8. identif$.ti,ab.

9. diagnos$.ti,ab.

10. exp Diagnosis/

11. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10

12. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9

13. older.ti,ab.

14. elder$.ti,ab.

15. geriatric$.ti,ab.

16. 13 or 14 or 15

17. "randomized controlled trial".ti,ab.

18. "controlled clinical trial".ti,ab.

19. randomized.ti,ab.

20. randomly.ti,ab.

21. trial.ti,ab.

22. groups.ti,ab.

23. exp Clinical Trial/

24. exp Randomized Controlled Trial/

25. 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24

26. exp Depression/

27. (depression or depressive or depressed).ti,ab.

28. (melancholi$ or dysphori$ or dysthymi$).ti,ab.

29. 26 or 27 or 28

30. 11 and 16 and 25 and 29

APPENDIX 8

Table of excluded studies from the systematic review of the clinical effectiveness of screening for depression in older adults

Study	Reason for exclusion
Alexopoulos et al. (2005)	All participants screened for depression and results for the ‘usual care’ (control) group disclosed to participants’ physicians and therefore influenced subsequent management. Some participants known to have depression.
Alexopoulos et al. (2009)	All participants screened for depression and results for the ‘usual care’ (control) group disclosed to participants’ physicians and therefore influenced subsequent management. Some participants already known to have depression. Overlap in sample with Alexopoulos et al. (2005).
Arthur et al. (2002)	All participants screened for depression and results for control group disclosed to general practitioner and therefore influenced subsequent management.
Banerjee et al. (1996)	All participants screened for depression and results for control group disclosed to general practitioner and therefore influenced subsequent management.
Bartels et al. (2004)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management. Investigated depression, anxiety and at-risk alcohol use.
Bijl et al. (2003)	Insufficient information – results incomplete. Overlap in sample with van Marwijk et al. (2008)
Blanchard et al. (1995)	All participants screened for depression and results for the control group disclosed to participants general practitioner and therefore influenced subsequent management.
Bogner et al. (2005)	All participants screened for depression and results for the ‘usual care’ (control) group disclosed to participants physicians and influenced subsequent management. Overlap of sample with Alexopoulos et al. (2005).
Bruce et al. (2004)	All participants screened for depression and results for the ‘usual care’ (control) group disclosed to participants physicians and influenced subsequent management. Overlap in sample with Alexopoulos et al. (2005).
Burke et al. (2002)	Not a study of the clinical effectiveness of screening. Commentary about preventing functional decline.
Cervera-Enguix et al. (2004)	Participants known to have depression. Positive screen result part of eligibility criteria and not intervention. Aimed to determine effectiveness of extended release Venlafaxine
Challis et al. (2004)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for a number of conditions, therefore not all participants would have a positive screen for depression.
Chew-Graham et al. (2007)	Participants already known to have depression. All participants screened for depression and results for the ‘usual care’ (control) group disclosed and influenced subsequent management.
Conwell et al. (2009)	Not a study of the clinical effectiveness of screening. Editorial.
Cutchin et al. (2009)	Participants screened for functional decline. Participants not screened for depression.
Dalby et al. (2008)	Does not investigate the clinical effectiveness of depression screening - explores appropriateness of antidepressant prescriptions.
Daniels et al. (2011)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for a number of conditions that reflect frailty, therefore not all participants would have a positive screen for depression.

Study	Reason for exclusion
Davison et al. (2013)	Not specifically focused on clinical effectiveness of depression screening. Interventions involved residential home staff being trained in recognition of depression and depression screening. Outcome of interest was recognition of depression in residential homes.
Dozeman et al. (2011)	Explores depression prevention rather than clinical effectiveness of depression screening.
Dozeman et al. (2012)	Explores depression prevention rather than clinical effectiveness of depression screening.
Ell et al. (2007)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management.
Emery et al. (2012)	No control group for comparison to determine clinical effectiveness of screening, which was part of the intervention.
Fischer et al. (2002)	Explores cost and health utilisation, instead of clinical effectiveness, in participants who screen positive for depression.
Gallo et al. (2005)	Explored mortality rates. Overlap in sample with Alexopoulos et al. (2005).
Gallo et al. (2007)	All participants screened for depression and results for the 'usual care' (control) group disclosed to participants physicians and influenced subsequent management. Overlap in sample with Alexopoulos et al. (2005).
Gallo et al. (2013)	All participants screened for depression and results for the 'usual care' (control) group disclosed to participants physicians and influenced subsequent management. Overlap in sample with Alexopoulos et al. (2005).
Gilbody et al. (2008)	Not a study of the clinical effectiveness of screening. Commentary about collaborative care model.
Gitlin et al. (2012)	Study protocol. Participants in control group informed of screening result, provided with support in community and encouraged to inform their primary care physician
Hebert et al. (2001)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for the presence of functional decline, therefore not all participants would have a positive screen for depression.
Imai et al. (2013)	Protocol for study. Data collection still underway.
Jeong et al. (2013)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management – i.e. 'usual care' group were prescribed Citalopram.
Jutkowitz et al. (2010)	Conference poster. Insufficient information.
Kasckow et al. (2014)	Sample does not meet age criterion of 55 years of age or older – age of participants in study 50 years of age or older
Knight et al. (2008)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management.
Kominski et al. (2001)	All participants screened for depression and results for control group disclosed and therefore influenced subsequent management.
Konnert et al. (2009)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management. Does not investigate the clinical effectiveness of screening – explores effectiveness of cognitive behavioural therapy.
Lam et al. (2010)	Does not investigate the clinical effectiveness of screening - explores effectiveness of 'brief problem-solving treatment'.

Study	Reason for exclusion
Levkoff et al. (2004)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management. Investigated depression, anxiety and at-risk alcohol use. Overlap in sample with Bartels et al. (2004).
Luptak et al. (2008)	Does not investigate the clinical effectiveness of depression screening – explores improvement in detection rates and communication with primary care
McCabe et al. (2013)	Not specifically focused on clinical effectiveness of depression screening. Overlap in sample with Davison et al. (2013)
McCusker et al. (1996)	Does not investigate the clinical effectiveness of depression screening – explores improves in detection rate by physicians.
McCusker et al. (2003)	Participants screened for identification of being 'at risk' – not screened for depression at start of study.
McMillan et al. (2009)	Not a study of the clinical effectiveness of screening. Commentary about another study.
Oyama et al. (2005)	Explores suicidal ideation therefore does not specifically focus on depression. Outcome of interest suicide prevention.
Oyama et al. (2006) outcomes	Explores suicidal ideation therefore does not specifically focus on depression. Outcome of interest suicide prevention.
Oyama et al. (2006) preventing	Explores suicidal ideation therefore does not specifically focus on depression. Outcome of interest suicide prevention.
Oyama et al. (2006) local	Explores suicidal ideation therefore does not specifically focus on depression. Outcome of interest suicide prevention. Depression is not the only cause of suicide.
Oyama et al. (2008)	Meta-analysis of effects of interventions using depression screening on suicide rate.
Oyama et al. (2010)	Explores suicide epidemiology as well as effects of depression screening on suicide rate. Depression is not the only cause of a suicide.
Pickett et al. (2014)	All participants screened for depression and screening results for both groups disclosed and therefore influenced subsequent management. Authors specifically comment that they do not know to what extent this influenced care received.
Pizzi et al. (2011)	Conference poster. Insufficient information.
Quijano et al. (2007)	No control group. Therefore no comparison group available to determine clinical effectiveness of screening.
Rabins et al. (2000)	Screening not specific to depression and clinical effectiveness not explored.
Raue et al. (2010)	Explores suicidal ideation therefore does not specifically focus on depression. Overlap in sample with Alexopoulos et al. (2005).
Raue et al. (2012)	Conference poster. Does not investigate the clinical effectiveness of depression screening – explores older adults' views about involvement in care.
Reuben et al. (1995)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for the presence of at least one of 13 conditions, therefore not all participants would have a positive screen for depression.
Reuben et al. (1999)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for the presence of a number of conditions.
Reuben et al. (2012)	Conference poster. Insufficient information.

Study	Reason for exclusion
Reuben et al. (2013)	Not specifically focused on clinical effectiveness of depression screening. Participants screened for the presence of at least one of four conditions (falls, urinary incontinence, dementia and depression), therefore not all participants would have a positive screen for depression.
Rubenstein et al. (2007)	All participants screened for depression and screening results for both groups disclosed and influenced subsequent management.
Sirey et al. (2008)	Study of depressive symptoms and suicidal ideation epidemiology. Does not investigate the clinical effectiveness of depression screening.
Soon et al. (2002)	Not specifically focused on clinical effectiveness of depression screening. Outcome of interest were frequency of physicals mental health consultations and antidepressant use.
Unutzer et al. (2006)	Explores suicidal ideation therefore does not specifically focus on depression.
Van der Aa et al. (2013)	Sample does not meet age criterion of 55 years of age or older – age of participants in study 50 years of age or older
Van der Weele et al. (2011)	Explores cost effectiveness of depression screening rather than clinical effectiveness.
Van't Veer-Tazelaar et al. (2006)	Protocol for Van't Veer-Tazelaar et al. (2010) and Van't Veer-Tazelaar et al. (2011)
Van't Veer-Tazelaar et al. (2009)	Does not investigate the clinical effectiveness of screening - explores effectiveness of stepped-care prevention of anxiety and depression
Van't Veer-Tazelaar et al. (2010)	Explores cost effectiveness of stepped-care prevention of anxiety and depression. Overlap in sample with Van't Veer-Tazelaar et a. (2010).
Van't Veer-Tazelaar et al. (2011)	Explores stepped-care prevention of anxiety and depression. Overlap in sample with Van't Veer-Tazelaar et a. (2010).
Williams et al. (2008)	Not a study of the clinical effectiveness of screening. Commentary about Gallo et al. (2007)