An Analysis of the Differences in the Canadian and US Guidelines for Depression Screening Among Adults in the General Population: Potential Impact of These Different Recommendations at the Population Level

By

Alex Donald

MPH Candidate

Simon Fraser University

For

Supervisor: Michel Joffres

Summer 2016

Introduction

Depression is one of the most common conditions within Western culture, with approximately 4.7% of the population experiencing it within the past year (Public Health Agency of Canada). It is a mental disorder that lowers ones mood, negatively affecting one's ability to carry out normal activities as well as inhibiting one's sense of well being (Public Health Agency of Canada). Depression exists over a spectrum of severity and duration. Its most notable form, Major Depressive Disorder (MDD), is characterized by a pervasive and persistent low mood and a loss of interest or pleasure in daily activities for 2 weeks or more (Joffres et al., 2013). MDD's effect on ones emotional state can be associated with impaired functioning in one's social, occupational and educational life.

Despite its high prevalence in society, MDD, and other forms of depression, often go undiagnosed. It has been estimated that only 42% of those with depression are diagnosed (Joffres et al, 2012). With nearly 60% of cases missed, mass screening for depression in primary care has been suggested as a prevention tool that may help clinicians identify cases that would otherwise be missed. By detecting the disease earlier, in theory, this would help prevent some of the negative health effects associated with depression at a population level. This would be because patients could be diagnosed and treated before the depressive symptoms become more severe (Joffres et al, 2012, O'Connor et al, 2016).

Mass screening in a primary care setting, in theory, has the potential to improve the quality of life for those undiagnosed with depression. However, opinions about whether mass screening is an appropriate preventative measure are split. In 2013, the Canadian Task Force of Preventative Health Care (CTFPHC) conducted a systematic evidence review of the benefits and harms of screening for adults within the general population and various population groups who were at high-risk for depression. The CTFPHC determined there is *insufficient* evidence for mass screening to be implemented at the primary care level (Joffres et al, 2012). This recommendation *differs* from the United States Preventative Services Task Force (USPSTF) guidelines (O'Connor et al, 2016). The USPSTF also recommended against screening at a primary care level, however, it did recommend that mass screening should be conducted within the general public and for the high risk group of pregnant and post-partum women in the presence of specialized mental health services (O'Connor et al, 2016). This difference in guidelines is interesting, given that Canada and the U.S. have similar population demographics and prevalence of depression. One would assume their recommendations would be similar, yet this is not the case. A major reason for this difference in recommendations can be attributed to how the systematic evidence reviews were conducted by the CTFPHC and USPSTF.

The difference in guidelines for depression screening between the CTFPHC and USPSTF raises a cause for concern. While screening can serve as a prevention tool, it also has the potential of being harmful to patients and unnecessary costly. The CTFPHC, in its evidence review, identified there is little evidence showing the benefits of screening, and stated the potential harms of unnecessary treatment and excess cost make screening an unsuitable prevention tool (Joffres et al, 2012). This recommendation has been greeted with approval from critics and other health care organizations, in contrast to the USPSTF decision to have mass depression screening as long as mental health services are present (Thombs & Zeigelstein, 2013). The USPSTFs decision has been subject to question, as many believe the evidence used to support its recommendation to screen with specialized services is inappropriate (Thombs & Zeigelstein, 2014).

The following paper looks to address this difference in recommendations and evidence review methodology for depression screening from the CTFPHC and USPSTF in primary care amongst the general adult population. In addition to analyzing mass screening for the general adult population will be a supplementary analysis of recommendations for pregnant and post-partum women, which also differed between the CTFPHC and USPSTF. Through an analysis of the most current evidence reviews for both task forces, this paper will demonstrate why the CTFPHC recommendation for depression screening is the most appropriate. The paper will begin with an overview of depression and its population-level impact within Western society, followed by a discussion of the rationale for using depression screening as a preventative tool. It will then analyze the methodology of the review design and methods of reviewing the quality of evidence of the CTFPHC and USPSTF systematic reviews will be compared. Finally, it will demonstrate how the CTFPHC is the stronger of the two recommendations, followed by a discussion of how the media has an influence on both the correct and incorrect depression screening recommendations within Canada and the United States, respectively.

What is Depression?

According to the DSM-IV, depression is defined as a change or fall in baseline mood to a point where one experiences a loss of pleasure in daily activities for two or more weeks (American Psychiatry Association, 1952). Baseline mood refers to the normal mood, or set point, at which an individual typically behaves (Ketter & Calabrese, 2002). An individual's mood can fluctuate around this baseline, and because of the human condition, nearly everyone has the ability to experience periods of depression over his or her life course. Amongst healthy individuals, however, one's mood typically is able to return back to baseline after the event that triggered the depression has subsided (Ketter & Calabrese, 2002). People with MDD and prolonged depression differ from healthy individuals in that, when depressed, they are unable to bring their mood back to their baseline (Ketter & Calabrese, 2002). The DSM-IV symptoms state that in order to be diagnosed as clinically depressed, an individual must experience at least 5 of the 9 symptoms (table) *nearly every day (*American Psychiatric Association, 1952*).*

There are several theories for the causes of MDD; genetic predisposition, biochemical imbalances, endocrine and neurophysiological dysfunction, psychological, and/or social processes and factors all have demonstrated some degree of association with MDD (Joffres et al., 2013). That said, evidence for these theories is limited (particularly those of a biological nature) and it is impossible to narrow down a particular one-size-fits-all triggers of MDD (Hasler, 2010).

Studies have shown that various environmental stressors, such as acute or chronic stress in one's life can trigger depressions onset. Traumatic events, such as a death of a loved one, divorce, and childhood trauma (like sexual or physical abuse) all have the potential to trigger depression (O'Conner et al., 2016). Depression can also stem from problems within interpersonal relationships, such as those that occur within the family, between couples and parents and their children. Other sources may develop from one's social environment, where differences in ethnicity and social class come to play (O'Conner et al, 2016). Gender inequalities across age groups for both females and males have also been cited as a source (O'Conner et al., 2016).

Certain population groups are also at higher risk for depression than others. One group in particular who experience depression at a higher prevalence than the risk of the population is women who are pregnant or post-partum (O'Connor et al., 2016). Post-partum refers to women who have recently given birth. Several studies have shown that post-partum depression occurs in approximately 10-15% of women following childbirth, with depression most likely to develop within the first 6 months after pregnancy (Meltzer-Brody, 2011).

Health and Societal Consequences of Depression

While depression decreases one's quality of life by negatively affecting one's mood, it also has the potential to indirectly influence one's physical health. Patients with depression tend to have reduced adherence to medical treatment, reduced participation in preventive activities, and are more likely to be exposed to risk factors like obesity, smoking and sedentary lifestyles (Joffres et al., 2013). MDD has also been associated with increasing the likelihood of immune dysfunction, cardiovascular disease, and various endocrine and neurological diseases (Joffres et al., 2013). For example, those who experience depression compared to the general population have higher risk levels for stroke (2.6 times), epilepsy (4-6 times), Alzheimer's Disease (1.7-2.7 times) and cancer (1.4-1.9 times) (Mood Disorder Society of Canada).

Since depression is associated with increased risk of other chronic diseases, it has a considerable indirect cost to the health care system. In the United States, the economic burden of depressive disorders was estimated to be in excess of $83 billion (Greenberg et al., 2015). However, this number does not include the cost of health care expenses alone; rather, it also includes the productivity loss caused by depression. In Canada, it was estimated that productivity losses due to depression were approximately $4.5 billion, while total inclusive health care costs exceeded $14 billion (Joffres et al., 2013).

Depression Screening

With a high-undiagnosed prevalence of depressed patients, there exists a substantial decreased quality of life amongst a large proportion of the population as well as a large economic burden (Joffres et al., 2013, Greenberg et al., 2015). With these concerns, mass screening for depression in primary care has been suggested as a preventative measure, as it may help clinicians identify missed depression cases, catch depressed patients earlier, and initiate appropriate treatment at an earlier stage of the disease.

Screening for depression is useful only to the extent that it improves patient outcomes beyond those of standard care (Thombs et al., 2014). To be successful, a screening program must identify a substantial number of patients in whom depression has *not already been diagnosed*, *engage those patients in treatment* and *obtain sufficiently positive* results *to justify the costs and potential harms associated with the program* (Thombs et al., 2014, Joffres et al., 2013). In order to accomplish this, it is ideal for depression-screening tests to have a high sensitivity and high specificity. Sensitivity of screening tests refers to the number of cases of those who have the illness of interest that are accurately diagnosed (Aschengrau & Seage, 2013). In the case of depression screening, that would be the percentage of depressed patients who are correctly screened positive for depression. Specificity, on the other hand, refers to the number of people that do not have depression who are accurately screened as negative (Aschengrau & Seage, 2013).

Screening tests that have low sensitivity and specificity mean there is the potential for false negatives and false positives, respectively (Aschengrau & Seage, 2013). A false negative test for depression screening would occur for someone who has depression but is screened as negative, whereas a false positive would mean someone is diagnosed for depression that does not actually have it (Aschengrau & Seage, 2013). False negatives are dangerous because those who need treatment will be missed and not receive it. False positives are harmful in the sense that those without depression may receive unnecessary treatment and utilize scarce resources that could do harm to the patients and create unnecessary health care costs. Balancing sensitivity and specificity in screening instruments is important. As a general rule, valid screening instruments should have at least a sensitivity of 80-90% and specificity of 70-85% (Joffres et al., 2013).

Depression screening instruments differ from other forms of screening tools in that they utilize questionnaires to determine whether or not a patient may have depression. Most other screening tools utilize various biomarkers to determine if the illness is present. Sometimes, these screening procedures can be quite invasive to the patients (Aschengrau & Seage, 2013). Depression screening on the other hand, is not invasive, and is relatively easy to administer. That said, because it utilizes questions rather than objectively measurable biomarkers, there is a higher potential for false positives to exist, as questions are very subjective and can be interpreted differently by patients (Thombs et al., 2013). Beyond sensitivity and specificity are issues of severity of the condition, its duration, whether or not the condition can be appropriately treated, and if the condition tends to improve on its own. Conditions that regress on their own lead to over diagnosis, unnecessary treatments and potential harm to the patient and costs to the system (Aschengrau & Seage, 2013).

A Comparison between the Canadian and United States Preventative Task Force Recommendation on Depression Screening

In order to address the effectiveness of mass screening for depression, both the CTFPHC and the USPSTF conducted systematic reviews to evaluate existing literature to determine whether it was beneficial to screen for depression amongst adults within the general population. As already noted, both task forces came up with different recommendations despite evaluating the same current literature. According to the CTFPHC, when it comes to depression screening, they recommended:

"For adults at average risk of depression, we recommend not routinely screening for depression and that *for adults in subgroups of the population who may be at increased risk of depression, we recommend not routinely screening for depression [This includes pregnant and post-partum women]."*

This recommendation differs from the USPSTF recommendation, which states:

"The USPSTF recommends screening for depression in the general adult population, including pregnant and postpartum women. Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up"

The reason for this different recommendation is primarily due to differences in the methodology used by the CTFPHC and USPSTF when conducting their systematic reviews. Differences in the key questions used to guide the reviews scope, the search strategy for which studies to include and exclude from the review, and the evaluation tools to grade the quality of the studies included within the review are three major difference in methodology between reviews. The following section looks to explore these differences by comparing and contrasting the CTFPHC and USPSTF reviews in these three areas, respectively.

Evidence for depression screening amongst high-risk groups, including pregnant and post-partum women, was also assessed by the CTFPHC and USPSTF. While evidence regarding pregnant and post-partum women will be mentioned in the following sections, it will only be supplementary to the main analysis, which focuses on mass screening amongst the general adult population.

Canadian Task Force for Preventative Health Care Evidence Review

Key Questions

Before conducting the systematic review, several key questions were created. Key questions serve as guiding questions that help investigators focus their literature search on their problem(s) they wish to address. In this case, investigators wanted to identify the benefits and harms of depression screening in primary care amongst adults within the general population and certain high-risk groups who were not known to have depression. The CTFPHC's key questions for this systematic review were as follows:

KQ1: What is the evidence for benefit of screening for depression in:

1. Asymptomatic adults 18 years of age or over from the general population in (i) primary care or (ii) other outpatient settings to improve critical outcomes?
2. Adults at increased risk of depression (including pregnant and post-partum women) in (i) primary care, (ii) other outpatient settings or (iii) specialty clinics to improve critical outcomes?

KQ2: What is the evidence for the harms of screening for depression in:

1. Asymptomatic adults 18 years of age or over not at increased risk of depression in (i) primary care or (ii) other outpatient settings?
2. Adults at increased risk for depression (including pregnant and postpartum women) in (i) primary care, (ii) other outpatient settings or (iii) specialty clinics?

These two questions served as the primary questions of interest the CTFPHC sought to investigate. Several other key questions were initially included. There was no follow up on these additional key questions, as there was insufficient evidence for the primary key questions about the benefits and harms of screening to warrant investigation of these further topics of interest. Several other secondary questions were also addressed by reviewers, but will not be analyzed in this report due to the lack of overlap between the USPSTF and CTFPHC reviews.

Search Strategy and Quality Assessment via GRADE

Once the key questions were established, investigators conducted a literature search for studies that addressed the benefits and harms of depression screening. Literature was limited by date (between 1994 and January 2012), language (English and French), and unpublished literature produced by various health organizations (grey literature) that used a number of keyword terms for depression and screening (Joffres et al., 2013). Any study design was included into the search, as long as the study involved a screen versus no screen comparison and participants belonged to a generalized adult population of 18 years old or greater (Joffres et al., 2013). Studies also required that they take place in a primary care setting, and measure at least one outcome of interest associated with depression (see table). Outcomes of interest were different when collecting literature for the benefits and harms of screening.

Once evidence was gathered according to these inclusion criteria, the strength of evidence was determined using the *Grading Recommendation Assessment, Development and Evaluation (GRADE)* system for rating the quality of evidence. While there are several systems for evaluating evidence, GRADE is unique because of its systematic and explicit approach to making judgments. This method of evaluating evidence can be applied across a wide range of interventions and contexts (GRADE, 2004). Due to its systematic nature, it is becoming one of most widely used evaluation tools when making clinical recommendations, with over 40 major organizations worldwide utilizing it, including the World Health Organization (Joffres et al., 2013, GRADE, 2004).

GRADE rates the quality of evidence as high, moderate, low or very low; with each of these four levels reflecting a different assessment of the likelihood that further research will impact the estimate of effect (GRADE, 2004). GRADE's quality rating (one of the four levels described above) is based on the assessment of several factors. For instance, studies that are randomized control trials are given a high rating (GRADE, 2004). Observational studies, including case-control and cohort studies begin with a low quality rating (due to their greater likelihood of internal bias) and may be further downgraded if there are other problems with the study design (GRADE, 2004). Another factor to consider is the consistency in the direction and size/of the estimates of effects between studies. For example, if two studies demonstrate different effects when analyzing the same type of intervention, this would raise question to the true results of the intervention (GRADE, 2004). Such a scenario would result in a lower quality of evidence that the intervention would be beneficial. Lower quality ratings of the intervention under question can also occur as a result of indirectness of the body of evidence to the populations, interventions, comparators and/or outcomes of interest. Imprecision of results if there are indications of reporting or publication bias can also result in a decreased quality rating for the recommendation (GRADE, 2004). Table 2 demonstrates the criteria for assigning quality of evidence using the GRADE criteria in the appendix.

In addition to assessing quality of evidence, GRADE also considers the strength of recommendations of the intervention as a whole. The strength of the recommendation is either characterized as strong or weak. Strong recommendations imply there is high confidence adherence to recommendation outweighs the undesirable effects. Weak recommendations indicate that the desirable effects of adherence to treatment probably outweigh the undesirable effects, but there is less confidence (Guyatt et al., 2008). Four factors are considered in making the strength of a recommendation. These include the balance between desirable and undesirable consequences of alternative management strategies, quality of evidence, values and preferences of patients, and cost-effectiveness of the intervention (Guyatt et al., 2008).

Results

In total, there were 14, 226 articles scanned by investigators (Keshavarz et al., 2013). Out of these studies, only five quasi-experimental studies addressed key question 1 and met the inclusion criteria described above (Keshavarz et al., 2013). Quasi-experimental studies refer to studies that compare treatment and standard-care care interventions, but lack the randomized component allowing for comparison groups to be near identical at baseline. The lack of randomization (a component present within Randomized Clinical Trials (RCTs)), reduces the internal validity of studies. The quasi-experimental studies included in the CTFPHC review were all conducted by the same author and took place in primary care settings in rural Japan. The age of participants were over 60, and all five studies were all very similar in design (Keshavaraz et al., 2013).

In the studies, the author had developed a universal suicidal prevention program to address the high prevalence rate of suicide amongst elderly Japanese citizens who lived in rural areas (Oyama et al., 2006). Suicide is one of the potential outcome measures for those with depression (Keshavarz et al., 2013). Within this prevention program, a screening component was included, along with mental health care or psychiatric treatment (Oyama et al., 2006). Results of the five studies demonstrated that the suicide intervention program, including the screening component, led to a statistically significant reduction in suicide for women aged 75-84 (Keshavarz et al., 2013). Results for women outside of this age-range did not show any significance, and the program also had no effect on men in any range (Keshavarz et al., 2013).

When using the GRADE criteria for evaluating these studies, the CTFPHC gave a rating of *very low quality* evidence (Joffres et al., 2012, Joffres et al., 2013). As the studies included were not RCTs, according to the GRADE criteria, they automatically began with a low rating (GRADE, 2004). The studies were also downgraded for indirectness, as the studies used population groups that were not represented in Canada. The study population group in this instance was elderly women and the study setting took place in a rural community (Joffres et al., 2013). The suicide rate in Japan is substantially higher than that of Canada as well, which decreased the studies' generalizability to the Canadian population (Joffres et al., 2013). Evidence was also downgraded because the suicide prevention program included education and treatment in conjunction with screening, as one cannot attribute the reduction of suicide directly to screening itself (Joffres et al., 2013).

In regards to studies that addressed the benefits of screening amongst pregnant and post-partum women, the CTFPHC found there was no direct evidence that demonstrated benefits. When analyzing the harms of screening amongst the general adult population, there were no studies that fit the proper inclusion criteria (Joffres et al., 2013). In addition to the quality of studies, potential costs associated with depression screening were also considered; such as the high rates of false positives associated with the screening tool, potential side-effects of treatment, and the costs associated with unnecessary treatment for patients where depression would regress naturally. The time spent screening and additional time spent conducting diagnostic interviews for false positive patients were also considered as potential costs.

Summary of CTFPHCs Recommendation

There was a very-low quality of evidence regarding the benefits of depression screening amongst adults in the general population and pregnant and post-partum women in primary care, and a general lack of evidence for harms of depression screening. As a result, the CTFPHC decided to recommend against depression screening. Due to the lack of well-conducted studies, the CTFPHC also gave a weak rating for strengths of recommendation. Thus, there is potential for change with this recommendation should higher quality studies, such as randomized control trials for mass screening of depression, be conducted in the future.

United States Preventative Service Task Force Evidence Review

Key Questions

Like the CTFPHC, the USPSTF utilized several key questions to guide investigators throughout their evidence synthesis. Two questions were utilized to determine the benefits and harms of depression screening:

Key Question 1: Does primary care depression screening programs in the general adult population result in improved health outcomes through decreased depressive symptomatology, decreased suicide deaths, attempts, ideation, improved functioning, improved quality of life, and improved health status?

> *a. Does sending depression screening test results to providers (with or without additional care management or support) result in improved health outcomes?*

Key Question 2: What are the harms associated with primary care depression screening programs in the general adult?

These guiding questions were very similar to those utilized by the CTFPHC. Key questions 1 and 2 both look at addressing the benefits and harms of mass screening, respectively. However, the major difference is that the USPSTF also looked to investigate the benefits of screening in conjunction with additional support upon being screened positive in key question 1a. The CTFPHC did not include this question, and due to its inclusion in the USPSTF review, the USPSTF increased its overall scope of its systematic review beyond the effects of screening.

The benefits of screening, benefits of screening when coupled with specialized mental health treatment, and harms of mass screening amongst pregnant and post-partum women were questions that were all addressed as well in the USPSTF review (O'Conner et al., 2016).

Search Strategy and Quality Assessment

Similar to the CTFPHC search strategy, two investigators independently reviewed titles and abstracts using pre-specified inclusion/exclusion criteria (O'Conner et al., 2016). To be included into the review, studies were required to be in English, include participants aged 18 years or older, and take place in a country ranked as having a "very high" human development index according to the WHO (O'Conner et al., 2016). Studies also had to assess either the benefits or harms of screening in a primary care setting to be included (O'Conner et al., 2016). Studies that were limited to persons with other medical or mental health conditions were excluded. Where the USPSTF inclusion criteria differed from the CTFPHC is that included studies were required to be either RCTs or non-randomized controlled trials (CCTs), with observational studies being excluded (O'Conner et al., 2016). Furthermore, the USPSTF chose to include studies that included participants who already had been diagnosed or were currently being treated for depression within its review (Thombs et al., 2014; O'Conner et al., 2016).

Investigators evaluated studies according to the criteria defined by USPSTF, and did not use the GRADE rating system. Each study was evaluated individually, and assigned a rating of good, fair or poor (O'Conner et al., 2016). Criteria for good quality studies required that studies included factors like adequate randomization produces, allocation concealment, blinding of outcome assessment, having a reliable outcome measure, comparable groups at baseline, low follow-up, acceptable statistical methods, and adequate adherence to the intervention (O'Conner et al., 2016). Studies that had most of these factors (but not all) were considered fair, and still included within the review. Studies rated as poor quality were excluded from the review for reasons such as having attrition greater than 40% of study participants, and other “fatal flaws” or accumulation of multiple minor flaws throughout the study (O’Conner et al., 2016). The major difference between the USPSTF’s and GRADEs assessment of quality of evidence is that GRADE utilizes a more transparent assessment.

Results

In total, there were 3,814 articles reviewed that studied the benefits and harms of screening for adults within the general population. Out of these studies, only one fit the criteria for key question 1, which addressed the benefits of screening (O’Conner et al., 2016). Eight studies fit the criteria for key question 1a, which addressed the benefits of screening when, coupled with treatment and additional mental health services (O’Conner et al., 2016). In total, two of the studies included for key question 1a were considered ‘good’ quality studies. The remaining 6 included in the study were considered ‘fair’ quality (O’Conner et al., 2016).

In the single study that addressed the benefits of screening alone amongst adults (Williams et al., 1999), researchers had two primary objectives. The first objective was to determine if a single question-screening instrument was as effective as the current 20-question instrument (Williams et al., 1999). The second objective assessed whether screening improved outcomes amongst those who underwent screening compared to those that did not undergo screening (Williams et al., 1999). This second objective was most applicable to key question 1. As a whole, the study demonstrated a modest, but statistically insignificant increase in recognition for depression amongst patients who underwent screening (Williams et al., 1999; O’Conner et al., 2016).

Williams et al. suggested there is a potential for screening to have an impact on improving health outcomes amongst depression patients, despite having insignificant results. However, there were some methodological design limitations in the study that hurt the accuracy of this statement. Those who were in the screening group and those in the standard-care group (no screening) were both interviewed and underwent diagnostic criteria to determine if they had depression or not. Since both groups underwent diagnostic interviews, it is uncertain whether screening identified more cases of depression compared to when screening had not occurred. Thus, in this instance, the comparison group is not a truly unscreened group, and has the potential to behave differently than they would under real-world conditions (i.e. without diagnostic interview). This methodology is subject to bias in the results, as it makes it very difficult to determine if screening in fact had any influence on increasing the number of depression cases compared to when no screening occurred. Furthermore, there would potentially be a decrease in the difference in case findings between the two groups, since the standard-care group may have been more aware of the symptoms and seek more help than if they had not been formally diagnosed. However after controlling for baseline severity of depression, the mean reduction in DSM-III-R symptom counts was similar for the casefinding (1.6 symptoms) and usual-care (1.5 symptoms) groups (*P* 5 0.21) (Williams et al., 1999). Therefore it is not clear why the USPSTF reached their conclusion.

For the studies that fit the inclusion criteria for key question 1a, samples were limited to patients who already experienced depressive symptomatology (Thombs et al., 2014). Screening, in combination with specialized programs generally provided an increased likelihood of remission and positive responses with treatment in the general adult populations (Thombs et al., 2014). All studies showed some degree of remission or a response from intervention groups, however only two of the studies showed statistical significance for services provided in addition to screening (Thombs et al., 2016). Screening programs were not as successful amongst older adults at reducing depression, and were reported to even have negative effects in some instances. However whether treatment works after screening vs. no screening is a different question. These studies *failed to address the main issue, which is to address the difference in outcomes between screening and no screening.*

In regards to harm of screening amongst the general adult population for key question 2, there was no evidence.

When assessing mass screening for women who are pregnant and post-partum, the USPSTF found 6 studies (out of 11, 869) which fit their inclusion criteria, and that assessed screening alone and screening with specialized mental health treatment (O'Conner et al., 2016). Two of these six were good quality studies, and the remaining four were fair quality. Evidence demonstrated that postpartum women had a 28 to 59% reduction in risk of depression in programs that involved depression screening; with programs either involving screening alone or screening with supplementary treatment (O'Conner et al., 2016). That said, none of these studies that were included had a straightforward design that compared usual care plus screening to usual care without screening (O'Conner et al., 2016). Various treatments for pregnant and post-partum depression were also assessed. Studies demonstrated that cognitive behavioral therapy (CBT) in addition to screen-detected depression resulted in a 34% increase in remission compared to usual care (O'Conner et al., 2016). Studies also showed that second-generation anti-depressants might have harmful effects, as there was increased risk of preeclampsia, post-partum hemorrhage, miscarriage, perinatal death, and preterm birth (amongst other negative outcomes) (O'Conner et al., 2016).

Summary of USPSTF Recommendation:

Evidence that assessed the benefits of screening alone was limited, with only a single study conducted by Williams et al. existing. This study demonstrated that there were potential benefits to screening, however, it lacked statistically significant results and was only given a 'fair' quality rating. The USPSTF also included evidence that assessed the benefits of depression screening when in combination with specialized programs for the general adult population. In two studies, it was shown that there were statistically significant results for benefits of screening with these specialized services. Other studies showed benefits as well, however, without statistically significant results. Evidence demonstrated that there were no benefits of screening amongst older adults, and in some cases, even negative effects. The USPSTF did not identify any studies demonstrating harms of depression screening amongst the general adult population. Due to this lack of evidence demonstrating harms, they decided that the magnitude of harms of screening for depression in adults is small to none (O'Connor et al., 2016). As a result of their findings, they recommended that screening for depression in the general adult population should occur with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up (O'Conner et al., 2016).

A similar recommendation was made in regards to women who are pregnant or post-partum, as studies demonstrated benefits in screening both with and without specialized mental health services (O'Conner et al., 2016). That said, evidence demonstrated that CBT should be used as the preferred treatment method, as secondary-antidepressants had potentially harmful effects for both the women and child (O'Conner et al., 2016).

Where the Differences in Methodology Lay

The CTFPHC and USPSTF came up with different responses in regards to mass depression screening amongst the general adult population, and for pregnant and post-partum women as a prevention tool. The differences in these recommendations can largely be attributed to how the overall systematic reviews were designed.

Perhaps the most notable difference between the two evidence reviews is in how the key questions were asked. The CTFPHC had two major key questions, which focused solely on the benefits and harms of screening as a preventative tool. The USPSTF used similar key questions to guide their review apart from one major question; however, investigators also looked to address how screening combined with additional specialized mental services available to patient would influence depression rates compared to screening alone. As mentioned above, key questions help guide investigators to determine which problems they are trying to solve. The addition of key question 1a in the USPSTF increased the scope of the review beyond looking solely at the preventative benefits of screening. Some have argued that addressing screening in conjunction with mental health services is not appropriate, as many primary care settings do not usually have such services (Thombs et al., 2014; Thombs et al., 2013). If this is true, findings are not necessarily applicable nor can the entire adult population benefit from such services.

While the key questions that guided the CTFPHC and USPSTF systematic reviews differed, there were also differences in the inclusion/exclusion criteria for the two different task forces. As a whole, the inclusion criteria was comparable; both required participants to be 18 years or older, have studies take place in a primary care setting, and measure outcomes for both benefits and harms of screening. That said, the USPSTF allowed for the inclusion of patients who were already previously diagnosed for depression within their review, while the CFTPHC excluded studies where participants had previously been diagnosed with depression. The inclusion of patients previously diagnosed with depression raises some potential concerns. As stated above, screening programs are deemed effective largely due to their ability to identify patients whom depression has not already been identified (Thombs et al., 2014). By including patients who have already been diagnosed or have a prior history with depression within the study design, the sample is not a true reflection of the general adult population. As a result, there is the potential for bias in the results.

The USPSTF and CFTPHC also differed in that the CFTPHC found it acceptable to include observational studies, whereas the USPSTF choose to only include studies that were randomized controlled trials and controlled clinical trials. For instance, the CFTPHC included five quasi-experimental studies, while the USPSTF included one clinical screening-based study conducted by Williams et al. The CFTPHC excluded the Williams et al study because of a methodological issue discussed above, which prevented it from being a true screening versus no-screening trial. This study was excluded because despite being randomized into the screening and no-screening groups, all patients went through a diagnostic interview to determine if they had depression. The CTFPHC viewed this design factor as defeating the purpose of finding out if it is the screening itself that leads to improvement in health outcomes, resulting in its removal. The USPSTF excluded the five quasi-experimental studies as they were observational, rather than neither RCTs nor CCTs.

Finally, differences between the two systematic reviews existed in how the overall quality of the evidence was evaluated by investigators. As described, the CTFPHC utilized the GRADE methodology for rating the quality of the included studies results, while the USPSTF utilized a self-defined criteria (Harris et al., 2001). The difference between these two evaluation tools is that GRADE is more transparent in its assessment of quality, and is one of the most highly utilized quality assessment tools internationally (GRADE, 2004). The USPSTF criterion, on the other hand, is not as transparent in its assessment of study quality. This difference in the evaluation tool has the potential to lead to differences in how recommendations are made, respectively. As stated, GRADE makes its recommendations based off of the study with the least beneficial results. In the case of the five studies included, none showed benefits of screening. However, in the case that several showed benefits while others did not, the studies demonstrating no benefits would hold more weight in the screening recommendations than the studies that demonstrated the beneficial effect. In contrast, the USPSTF had only two studies that showed statistically significant benefits of screening with specialized mental health services. The remaining studies showed no statistically significant results. However, those two studies with significant results were held with greater weight in the decision to screen. Had the USPSTF been using GRADE criteria, the studies with the non-statistically significant results would hold more weight in the decision for screening recommendations. GRADE also accounts for the strength of its recommendations. The CTFPHC recommendation recognized there was a general lack of high quality of studies used to make its recommendation. They gave a weak strength of recommendation, acknowledging that if more high quality studies (such as a properly conducted RCT), this recommendation could change.

Different Methodologies, Different Interpretations, Different Recommendations

Both the USPSTF and CTFPHC systematic reviews were conducted under highly rigorous conditions. Each task force had set criteria it followed when conducting its evidence review for depression screening. Key questions 1 and 2 were similar between the CTFPHC and USPSTF reviews, respectively. However, as demonstrated above, the addition of key question 1a and methodological designs served as a major reason for differences in interpretation of the evidence. These differences in interpretation thus lead to different recommendations. This discrepancy in recommendations between the CTFPHC and USPSTF raises the issue of which recommendation is the *most* correct.

The CTFPHC made their decision to recommend not screening for depression amongst the general adult population, largely due to a lack of existing literature supporting benefits of screening (Joffres et al., 2013). As demonstrated, only five studies fit the CTFPHCs inclusion criteria (which were basically the same study), and all of them received a very low quality of evidence for demonstrating the benefits of screening. The lack of randomized screen versus no screen trials prevents one from fully knowing if any true benefits of screening for depression exist.

While there was no evidence regarding harms of depression, the CTFPHC recognized that despite the lack of evidence, there is still the possibility of potential harms existing. For instance, it is estimated that only 10-20% of positive depression screening in primary screening would be true positives (Thombs et al., 2013; Thombs et al., 2014). This suggests that if mass screening were implemented, there would be a high rate of false-positives. To conduct diagnosis interviews with many people without depression could lead to a considerable amount of time wasted in an already time-constrained health care system (Thombs et al., 2014). Another potential harm is that anti-depressive treatment is not effective amongst those who are at the lower end of the depression scale (Thombs et al., 2014). Many people experience depression at some point in their lives, but it does not last for a long period of time, and regresses naturally. The majority of individuals who experience depression fall into this category of mild depression. Studies have shown that giving treatment to those with mild depression has little benefits, and can even lead to negative side effects (Thombs et al., 2014). For women who are post-partum, the majority of depression cases also regress naturally with time. Anti-depressants can lead to increased blood pressure, and potential drug-to-drug interactions for patients taking other medication. In addition to negative side effects, studies have found that overall health care costs were about $2000 higher for patients taking anti-depressants (Thombs et al., 2014). For pregnant and post-partum women, there can also be other serious consequences for both the women and her in-utero child.

The National Institute of Health Care and Excellence (NICE) in the United Kingdom have come up with similar recommendations when it comes to depression screening. They too were concerned about the high rates of false positives and the lack of effect of treatment on those with mild depression (O'Conner et al., 2016). As a result, NICE concluded that while it is not recommended to screen for the average adult population, it is recommended that physicians be alert to possible depression, particularly when there has been a previous history for depression amongst the patient (Joffres et al., 2013).

The lack of quality evidence demonstrating the benefits of screening and the potential harms of screening for depression led the CTFPHC to make their recommendation against screening. When assessing the USPSTF recommendation, their evidence review determined to that mass screening amongst adults in the general population *should occur in the presence of specialized mental health programs*. This recommendation has come under heavy criticism from critics despite the USPSTF interpreting that evidence demonstrating screening plus specialized mental health treatment was beneficial (Thombs et al., 2014).

In order for screening to be a preventative tool, one of the main goals is to identify patients who would not have otherwise been diagnosed with depression (Thombs et al. 2014). When patients are already diagnosed, they are more likely to be aware that they are susceptible to depression, thus are likely to reach out to their health provider to receive treatment. When addressing key question 1a, the USPHTF included studies where participants had already been diagnosed with depression. The inclusion of these participants hurts the overall quality of the studies, as the target of the screening intervention would be for those who have not already been identified with depression (Thombs et al., 2014).

While the inclusion of those who were already depressed in the studies included in the systematic review hurt the quality of these studies within the USPSTF review, perhaps the greatest issue with the USPSTF recommendation is that it looked to address the benefits of screening with the addition of specialized treatment services. The CTFPHC chose not to include screening coupled with such services due to the fact that the goal was to identify the effects of screening alone. If specialized treatment were included, it would be impossible to determine whether it was the screening itself that made a difference, especially considering that the majority of primary care clinics would only provide the screening component. As demonstrated above, the five quasi-experimental studies in the CTFPHC were given a decreased quality in rating because they included mental health education and treatment interventions in combination with the screening component. The USPSTF, on the other hand, chose to address both questions separately. This was an issue in both the general adult population group and for pregnant and post-partum women in the USPSTF studies.

The USPSTF, like the CTFPHC, also identified no evidence regarding the harms of depression screening. However, rather than considering the potential harms (as the CTFPHC did), they interpreted this to mean that there are *little to no harms* associated with depression screening amongst adults (O'Conner et al., 2016). Due to this interpretation, they felt that the benefits of depression screening greatly outweigh these (almost) non-existent harms. In turn, this led them develop their (albeit generous) recommendation for depression screening with mental health services.

So… Which Recommendation is *Most Appropriate* at this time?

As detailed in the preceding section, the CTFPHC made its decision to not recommend mass screening on the philosophy that it is better error on the side of caution. From their perspective, insufficient evidence gives reason not to spend valuable (and scarce) resources on a public health intervention that may not actually be beneficial, and could even potentially cause harm. They also recognized that screening in primary care is not usually followed up with specialized mental health services, and that combing screening with specialized mental health services addresses a separate issue. Different groups have supported this belief as well and have likened this scenario to testing usual case-finding for cancer plus less-than-ideal cancer care versus screening plus state-of-the-art treatment (O'Conner et al., 2016, Thombs et al., 2013, Thombs et al., 2014). The USPSTF, on the other hand, used a methodology in their review that provided evidence that encouraged screening. This methodology, however, has been subject to question for several reasons, which have been described above. While the decision of which recommendation ultimately comes down to one of philosophical values*, this analysis' results support that the CTFPHC was correct in their decision not to recommend screening* compared to the USPHTF recommendation to screen even with the presence specialized mental health services.

Media influence on Depression Screening Recommendations

As argued above, the USPSTF recommendation to screen even with specialized mental health services is viewed by many as questionable. The U.S. media has a profound impact on the dissemination of health information; including these guidelines and recommendations. While dissemination of information through the media can (and often does) serve as a useful medium to transfer important health information, in the case of depression screening, this transmitting of information to the general public has the potential to have a negative impact.

Multiple wide-reaching newspapers have spread the depression screening recommendation from USPSTF guidelines. For example, Scientific America states all adults in the U.S. population should be screened for depression when they visit the doctor. As argued above, mass screening should not occur amongst the general adult population, as there is no evidence demonstrating its benefits to society and may potentially be harmful and unnecessarily costly. This is especially true for screening without specialized mental health services, and even the U.S. recommends against screening alone. *By stating all adults should be screened for depression is even incorrect according to the USPSTF guidelines*, as it fails to state that patients should only be screened with specialized mental health services available (although, even this recommendation is subject to criticism as argued above). It also does not demonstrate that screening can be harmful to the elderly population. Thus, we see how the media can disseminate misleading information, leading to negative effects within society. One may expect increased rates of depression screening in primary care settings without specialized mental health services, and amongst the elderly population.

USA Today can demonstrate another example of the negative impact the media has on disseminating these results in this quote:

"in its previous recommendation….. the U.S. Preventative Services Task Force recommended screening adults for depression only when mental health services were available…. In its new report, the task force says this limitation is no longer needed because mental health services are widely available today."

This statement sends information as if to say all primary care settings have mental health services attached to them. While specialized mental health services may be more available than in the past, it is unlikely that all primary care settings have these specialized services. With this statement, however, one could expect more patients may go to their primary care providers asking to be screened. With increased screening, there is a greater risk for false positive results, and over-treating patients who would not benefit from taking anti-depressant medication.

From an economic perspective, this will lead to a considerable waste of valuable resources in addition to the potential harm to patients.

From a Canadian standpoint, there has been relatively little distribution of depression recommendations in the media compared to the U.S. for depression screening. However, the United States media still heavily influences Canada's media. Due to the U.S.'s influence, several news articles have stated that Canada should have depression screening for all adults and pregnant and post-partum women, despite the CTFPHCs *correct* recommendation against screening for these groups. Although there has been little development as of yet, there is a concern that Canadians may call for mass depression screening similar to the U.S. despite the lack of evidence that promotes screening.

Conclusion and Future Recommendations

Depression can be a debilitating illness for those who experience it, and its prevalence in western culture is high. The CTFPHC and USPSTF conducted systematic reviews to evaluate existing literature to determine if depression screening served as a viable preventative measure to reducing the burden of this highly prevalent condition. The two task forces came up with separate recommendations in regards to depression screening amongst adults. The CTFPHC made their recommendation against depression screening as they found there was a lack of evidence demonstrating the benefits of the disorder. The USPSTF, in contrast, recommend for depression screening in the presence of specialized mental health services.

The differences in recommendations were largely due to differences in methodology in how the systematic reviews were constructed. A major difference in key questions was that the CTFPHC decided to address the impact of screening alone, while the USPSTF choose to also address the impact of screening with mental health services. As a result, the nature of the studies included for the systematic review differed substantially. The two task forces also had different designs in the inclusion/exclusion criteria for literature and how they evaluated the quality of the included studies, respectively. The different inclusion/exclusion criteria and quality assessment tools had different interpretations about which studies were deemed as acceptable into influencing the review. Critics have generally supported the CTFPHC's process in the conduction of its systematic review, and praised it for its decision to recommend against mass screening, while these same critics view that the methodology of how the USPSTF came to its recommendation was somewhat flawed.

When developing guidelines and recommendations like those for mass screening for depression, it can be very difficult to come to a common conclusion, especially when different organizations are responsible for conducting their own systematic reviews. Different organizations have their different beliefs and opinions, and these beliefs are present in the way they conduct their reviews and interpret their results. However, drastically different recommendations create confusion, be harmful to individuals and unnecessarily costly, both to individuals and the health system.

From this analysis, several recommendations can be made for future guidelines for depression screening, in the hopes of making sure guidelines are based on high quality evidence. The first is that there is a need to conduct randomized clinical trials that study the impact of true screening vs. no screening trials. Such studies could be conducted similarly to the Williams et al study; however, only patients screened positive would go for the diagnostic interview rather than all patients in the study. This would provide a high quality study on which task forces could use to base their recommendations.

The second recommendation is that reviews of mass-screening studies must be based on populations that are representative of the populations of interest. This is a basic principle of public health and clinical research, however. In addition, the USPSTF decided to include patients who had depression in their study design. The objective of screening is to identify patients who have not yet been identified as having depression.

A third recommendation, which is similar to the second, is that organizations should try to adopt a similar method of quality assessment when evaluating the quality of studies. The CTFPHC and many other organizations worldwide have utilized the GRADE method when making recommendations, as it provides a systematic and transparent way of evaluating studies. GRADE also uses strength of recommendations (classified as either strong or weak) to acknowledge that further information could change the current recommendation. It is suggested that the USPSTF adopt GRADE instead of its own quality assessment criteria, to provide continuity across organizations and improve its transparency in how it identifies studies that are of appropriate quality, as well as strength of its current recommendations.

It is important to recognize that because different organizations are responsible for conducting their own reviews, some with major conflict of interests, others with different questions and methodology, they may not come to the same conclusions when making recommendations for interventions like mass depression screening. These decisions may result in a wide scale impact on both the health of population and costs to the health care system. While one does not know if the decision to recommend mass screening amongst the general adult population and pregnant and post-partum women in the U.S. will be beneficial or not, it appears at this time that the CTFPHC has made the right choice to recommend against this screening.

References:

American Psychiatric Association (1952). Diagnostic Criteria for Major Depressive Disorder and Depressive Episodes. Retrieved from: http://www.psnpaloalto.com/wp/wpcontent/uploads/2010/12/Depression-Diagnostic-Criteria-and-Severity-Rating.pdf on Friday May 27th, 2016.

Aschengrau, A., Seage, G. (2013). Essentials of Epidemiology in Public Health. Chapter

Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group (2004). Grading quality of evidence and strength of recommendations. British Medical Journal. 328. 1490-1494.

Greenberg, P. Fournier, A., Sisitsky, T., Pike, C., Kessler, R. (2015). The Economic Burden of Adults with Major Depressive Disorder in the United States (2005 and 2010). Journal of Clinical Psychiatry. 76 (2). 155-162.

Guyatt, G., Oxman, A., Kunz, R., Falck-Ytter, Y., Vist, G., Liberati, A. (2008). Grade: Going from Evidence to Recommendations. British Medical Journal. 336. 1049-1051.

Hasler, G. (2010). Pathophysiology of depression: do we have any solid evidence of interest to clinicians? World Psychiatry. 9. 155-161.

Harris, R., Helfand, M. Woolf, S. Lohr, K. Mulrow, C., Teutsch, S. Atkins, D. (2001). Current Methods of the U.S . Preventive Services Task Force. American Journal of Preventive Medicine. 20 (3). 21-35.

Joffres, M., Pottie, K., Dickinson, J., Lewin, G., Shaw, E., Tonelli, M (2012). *Screening for Depression*. Canadian Task Force for Preventative Health Care. McMaster Evidence Review and Synthesis Center.

Joffres M, Jaramillo A, Dickinson J, et al. (2013) Recommendations on screening for depression in adults. CMAJ 2013 Jun 11;185(9):775-82.

Ketter, T. Calabrese, JR. (2002). Stabilization of mood from below versus above baseline in bipolar disorder: a new nomenclature. Journal of Clinical Psychiatry. 63 (2). 146-151.

National Institute for Health and Clinical Excellence. (2013) Depression in adults: recognition and management. London: NICE

O'Connor, E., Rossom, R., Henninger, M., Groom, H., Burda, B., Henderson, (2016). Screening for Depression in Adults: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force.United States Preventive Services Task Force. Kaiser Permanente Research Affiliates Evidence-based Practice Center

O'Connor, E., Whitlock, E., Bell, T., Gaynes, B. (2009). Screening for Depression in Adult Patients in Primary Care Settings: A Systematic Evidence Review. Annals of Internal Medicine. 151. 793-803.

Oyama H, Fujita M, Goto M, Shibuya H, and Sakashita T. Outcomes of community-based screening for depression and suicide prevention among Japanese elders. Gerontologist. 2006; 46(6):821-826.

Pignone, M., Gaynes, B., Rushton, J., Burchell, C., Orelans, T., Mulrow, C. (2002). Screening for Depression in Adults: A Summary of the Evidence for the U.S. Preventative Services Task Force. Annals of Internal Medicine. 136. 765-776.

Public Health Agency of Canada. (2014). What is Depression? Retrieved from:http://www.phac-aspc.gc.ca/cd-mc/mi-mm/depression-eng.php

Thombs, B., Ziegelstein, R. (2014). Does Depression Screening Improve Depression Outcomes in Primary Care? British Medical Journal.

Thombs, B., Ziegelstein, R. (2013). Depression Screening in Primary Care: Why the Canadian Task Force on Preventative Health Care did the right thing. Canadian Journal of Psychiatry. 58 (12). 692-696.

Thombs BD, Ziegelstein RC, Roseman M, et al. There are no randomized controlled trials that support the United States Preventive Services Task Force guideline on screening for depression in primary care: a systematic review.

Williams, J., Mulrow, C., Kroenke, K., Dhanda, R., Badgett, R., Omori, D., Lee, S. (1999). Case-Finding for Depression in Primary Care: A Randomized Trial. The American Journal of Medicine. 106. 36-43

Appendix:

Table 1: DSM-IV Symptoms for Major Depressive Disorder (APA, 1952)

1. Depressed mood or irritable most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful).

2. Decreased interest or pleasure in most activities, most of each day

3. Significant weight change (5%) or change in appetite

4. Change in sleep: Insomnia or hypersomnia

5. Change in activity: Psychomotor agitation or retardation

6. Fatigue or loss of energy

7. Guilt/worthlessness: Feelings of worthlessness or excessive or inappropriate guilt

8. Concentration: diminished ability to think or concentrate, or more indecisiveness

9. Suicidality: Thoughts of death or suicide, or has suicide plan

Table 2: GRADE Criteria for Evaluating Study Quality (GRADE, 2004)

Type of Evidence:

Randomized trial = high quality

Observational study = low quality

Any other evidence = very low quality

Decrease grade if:

-Serious (-1) or very serious (-2) limitation to study quality

-Important inconsistency (-1)

-Some (-1) or major (-2) uncertainty about directedness

-Imprecise or sparse data (-1)

-High probability of reporting bias (-1)

Increase grade if:

-strong evidence of association – significant relative risk of >2 (<0..5) based on consistent evidence from two or more observational studies with no plausible confounders (+1).

-very strong evidence of association – significant relative risk of > 5 (<0.2) based on direct evidence with no major threats to validity (+2)

-Evidence of dose response gradient (+1)

-All plausible confounders would have reduced the effect (+1)

Table 3: Definitions of grades of Evidence (GRADE, 2004)

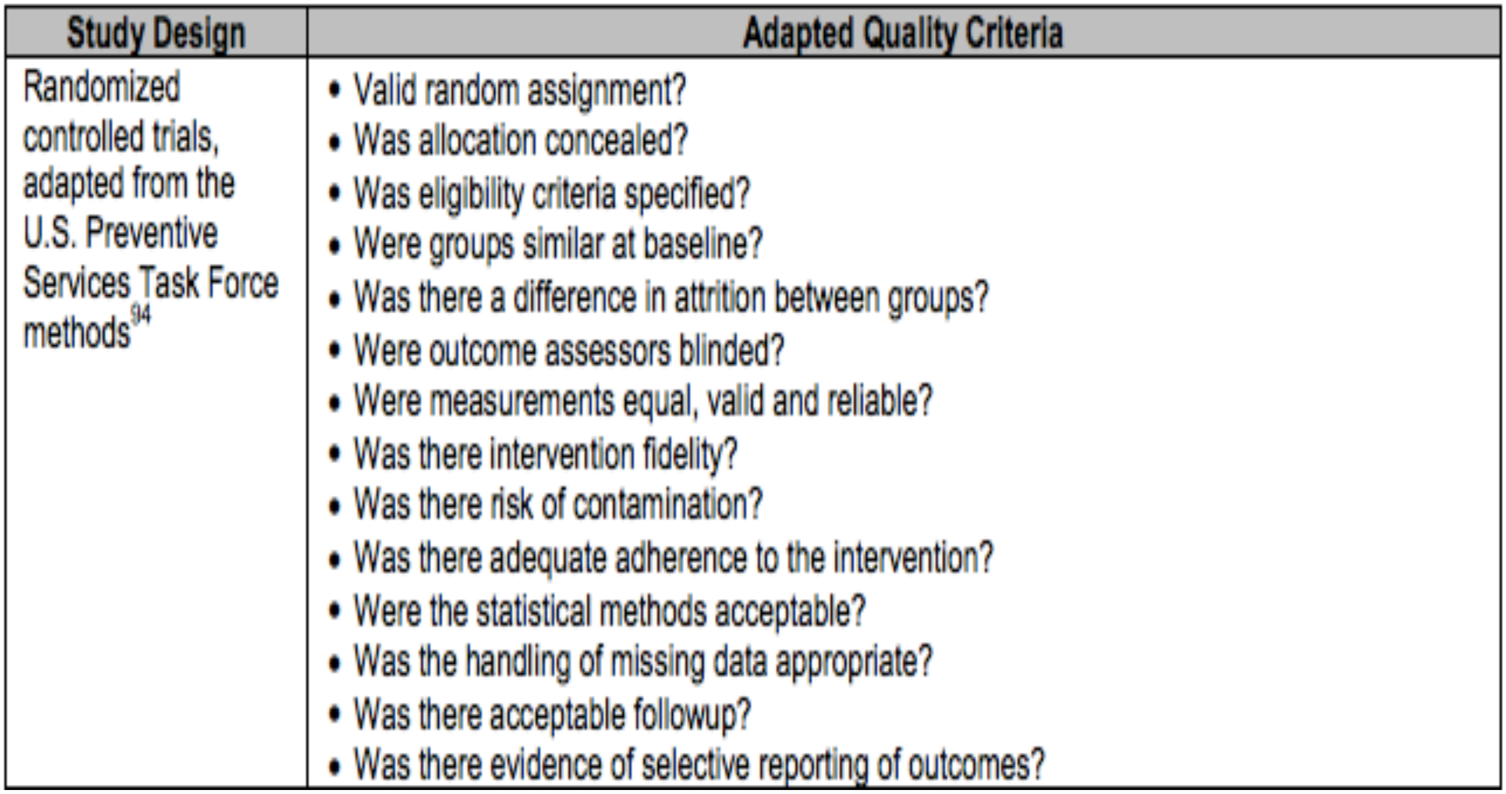
High = Further research is unlikely to change our confidence in the estimate of effect

Moderate= Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate

Low = Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate

Very Low = any estimate of effect is very uncertain

Table 4: USPSTF Self-Defined Criteria for Quality Assessment of Studies (O'Conner et al., 2016)



Study Design	Adapted Quality Criteria
Randomized controlled trials, adapted from the U.S. Preventive Services Task Force methods[94]	• Valid random assignment? • Was allocation concealed? • Was eligibility criteria specified? • Were groups similar at baseline? • Was there a difference in attrition between groups? • Were outcome assessors blinded? • Were measurements equal, valid and reliable? • Was there intervention fidelity? • Was there risk of contamination? • Was there adequate adherence to the intervention? • Were the statistical methods acceptable? • Was the handling of missing data appropriate? • Was there acceptable followup? • Was there evidence of selective reporting of outcomes?