Chapter 8

General discussion
GENERAL DISCUSSION

In this general discussion, the main findings of the presented work are described and discussed in light of the current literature. Furthermore, strengths and limitations of the studies presented will be explicated and implications for further research and practice will be discussed.

This thesis is aimed to answer the following central questions:

1) Which different trajectories can depression take within an older population with a high rate of medical illnesses? Chapter 2 presents data on the trajectories of depression in a longitudinal cohort of 396 patients with major- or subthreshold depression. Predictors of these trajectories and associations with health care costs and quality of life are described.

2) How can depression and depressive symptoms best be measured in patients with diabetes mellitus type 2 (DM2) and/or coronary heart disease (CHD)? Chapter 3 contains a systematic review to establish whether self-reported questionnaires can be used to evaluate depressive symptoms in patients with diabetes. Chapter 4 presents data on the diagnostic accuracy of the Dutch version of the PHQ-9 to establish major- and subthreshold depression in patients with DM2 and/or CHD.

3) Is the Step-Dep program (cost-)effective as compared to usual care in preventing the onset of depression in patients with DM2 and/or CHD? Chapter 5 presents the design of a pragmatic cluster randomized controlled trial to evaluate the effectiveness and cost-effectiveness of a flexible stepped care approach (Step-Dep). The results of Step-Dep are subsequently described in Chapter 6 (effectiveness) and Chapter 7 (cost-effectiveness).

The previous chapters provide information to answer these three questions. In these chapters the distinction is made between major depression, (a major depressive disorder according to the Diagnostic and Statistical Manual of mental disorders – fifth edition (DSM-5) [1]), and subthreshold depression (experiencing clinically relevant depressive symptoms without completely fulfilling the DSM-5 criteria for major depressive disorder). Throughout the different chapters the terms ‘minor depression’ and ‘subclinical depression’ are also used to indicate subthreshold depression. In this general discussion the terms ‘major depression’ and ‘subthreshold depression will be used.
Main findings and embedding in the literature

Part 1) Characteristics of depression trajectories in an older primary care population
In Chapter 2, three distinct trajectories of depression were identified in a population of older primary care patients with subthreshold depression and a high prevalence of chronic medical illnesses. Most patients showed predominantly persistent subthreshold depression. Two relatively small groups showed a more adverse pattern: one trajectory followed an intermittent course of depressive symptoms over time, the other trajectory was characterized by initially severe depression improving over time, but never reaching total remission. In all three identified trajectories, subthreshold depression remained present during the whole three-year follow-up period. Other studies mainly found the same distinct courses of depression in older adults [2, 3]. However, some differences exist in the number and nature of the identified trajectories. These differences are possibly explained by differences in model specifications within the statistical analysis. For example, Byers et al. [2] and Cui et al. [3] did not identify an intermittent trajectory of depressive symptoms. This may be explained by the fact that these studies estimated only trajectories with linear slopes. Forcing a linear slope only allows one-directional and gradual change over time, thereby overlooking any fluctuating patterns. In addition, differences in study population may also have resulted in differences in the found depression trajectories. Byers et al. performed their analyses in women from the general population and Cui et al. used a representative sample of older adults from primary care that was not limited to patients who already experience subthreshold depression.

The strongest predictor of a more adverse depression trajectory was severity of baseline depression. Living without a spouse and not feeling in control of the situation were predictive for an intermittent depression trajectory. Other studies found comparable predictors for adverse depression trajectories [2-5], and also showed that the prognosis is worst for older depressed adults who also experience problems in other domains (who for example experience more social isolation or a lower health related quality of life). Recently, a prospective cohort study indicated that having chronic illnesses increases the risk for a chronic course of depression in older depressed adults even more, underlining the poor prognosis of patients who also experience problems in other domains [6].

The study presented in this thesis assessed the association of depression trajectories with health-related costs and quality of life in a depressed older population. Patients in the intermittent trajectory reported lower health related quality of life than in the persistent subthreshold trajectory over a three-year period. After controlling for possible confounders, the total effect is comparable to a loss of over two months in perfect health over a period of two years. Health care costs are higher in...
the more adverse trajectories than in the persistent subthreshold trajectory, but this difference is not statistically significant. Considering the skewed distribution of costs, this analysis may be underpowered for costs which is also reflected in the wide confidence intervals. Although previous work describes depression trajectories and their predictors in comparable samples [2, 3], currently no other studies are performed in older primary care patients to relate depression trajectories to health outcomes over time. However, in the general primary care setting some data is available confirming the presented findings [7].

These findings provide an answer to the first central question of this thesis: **Which different trajectories can depression take within an older population with a high rate of medical illnesses?** Overall, the presented results add up to the existing literature by demonstrating the chronicity of depression in primary care patients who are 55 years of age or older [8, 9]. It is clear that once (subthreshold) depression emerges in this population with high rates of chronic medical illnesses, chances for complete and lasting recovery are slim. It seems that older adults with subthreshold depression who also experience problems in other domains have the poorest prognosis [2, 3, 5, 6, 8, 9]. Furthermore, the patients in the most adverse trajectories report lower health related quality of life. Since only few studies relate depression trajectories in (older) adult primary care patients to outcomes, the presented results are still preliminary in this regard and need replication in other studies.

**Part 2: Measurement of depression in patients with DM2 and/or CHD**

The systematic review in Chapter 3 included 21 studies evaluating the measurement properties of nine different self-reported questionnaires for evaluating depressive symptoms in diabetes patients. The overall recommendation of the review was to use the CES-D [10], with strong evidence for a positive internal consistency, structural validity, and construct validity, moderate evidence for a positive criterion validity and limited evidence for positive cross-cultural validity. The WHO-5 [11] and the PHQ-9 [12] were also considered. However, the WHO-5 was originally developed as a general questionnaire to measure health related quality of life and evidence for the structural validity of the PHQ-9 was contradictory. Thus, users should be cautious when using the WHO-5 or the PHQ-9. For all other questionnaires, evidence was too limited to give any recommendations. Moreover, since for none of the questionnaires complete information is available on all measurement properties, no definitive conclusions can be drawn [13].
Other reviews assessing the measurement properties of depression questionnaires in patients with chronic medical illnesses like cancer [14] and Parkinson’s disease [15] provided comparable recommendations, suggesting that our findings are robust.

In Chapter 4, the diagnostic accuracy of the PHQ-9 for identifying depression in primary care patients with DM2 and/or CHD was assessed. Four diagnostic strategies were considered: the sum score to identify patients at high risk for major depression, the sum score to identify patients at high risk for subthreshold depression, an algorithm based diagnosis of major depression and an algorithm based diagnosis of subthreshold depression.

The performance of the PHQ-9 is satisfactory when using the sum score. To identify patients at high risk for major depression, a cut-off score of 10 was most appropriate (sensitivity 84%, specificity 82% and an Area Under the Curve (AUC) of 0.88). Patients at high risk for subthreshold depression could best be identified using a cut-off score of 8, resulting in a sensitivity of 71%, a specificity 71% and an AUC of 0.74). It was not recommended to use the pre-specified algorithm to diagnose major or subthreshold depression in the current population. The positive predictive value of these algorithms was low, 33% and 25% respectively, indicating a high false-positive rate.

The presented findings are largely in line with previously published work on the diagnostic accuracy of the PHQ-9 [16-18] in chronically ill patients. However, recommended cut-off scores for subthreshold depression differ greatly between studies (even in comparable populations) [16-18]. This raises questions about the most appropriate cut-off score and practical possibilities of using the PHQ-9 for this purpose in daily practice. Also, we found that the cut-off score to identify patients at high risk of subthreshold depression is close to the cut-off score for major depression (the difference between these cut-off scores is only two on a scale of 0-27), indicating practical difficulties when it is aimed to distinguish between patients with possible major or subthreshold depression.

Based on these findings a partial answer can be provided to the second central question of this thesis: **How can depression and depressive symptoms best be measured in patients with diabetes mellitus type 2 (DM2) and/or coronary heart disease (CHD)?**

First, several frequently used questionnaires seem suitable to measure and evaluate the severity and course of depressive symptoms, and to compare scores within individuals. The use of the CES-D for DM2 patients is best supported by evidence, followed by the PHQ-9 and the WHO-5. In the light of other literature [19, 20], it is plausible that comparable conclusions can be drawn for patients with CHD.
To identify patients at risk for major or subthreshold depression from a large group of patients, and to compare scores between patients, it is difficult to recommend a specific questionnaire. While the presented evidence shows that one of the most used screening tools (the PHQ-9) is reasonably sensitive and specific, recommended cut-off scores vary between studies and cut-off scores for subthreshold and major depression are close to each other. This complicates the practical use of the PHQ-9 for this purpose. Previous studies have shown that the same holds true for other screening tools [20, 21]. On the other hand, the alternative, i.e. identifying chronically ill patients at risk based on help seeking behavior for depressive symptoms and the clinical evaluation of the general practitioner, might result in substantial under-diagnosis of depression in DM2 and CHD patients. This complicates addressing symptoms at an early stage [22-24].

For diagnosing major or subthreshold depressive disorder it is not recommended to use a PHQ-9 algorithm, because of high false-positive rates. Although it would be best to use a validated psychiatric interview to exclude or confirm a suspected major or subthreshold depressive disorder in both practice and research settings, we acknowledge that this may be costly, time consuming, and difficult to apply. It is therefore important for general practitioners to make good use of up-to-date guidance in a structured way and use careful monitoring, since non-structured clinical assessment by general practitioners could result in over-diagnosis of depression [25].

Part 3: effectiveness and cost-effectiveness of the Step-Dep study

In Chapter 5, the design of a cluster-randomized controlled trial (RCT) to assess the effectiveness and cost-effectiveness of Step-Dep is described. Step-Dep is the first stepped care program designed to prevent major depression in primary care patients with DM2 and/or CHD who experience subthreshold depression. It consists of four subsequent evidence-based treatment steps with increasing intensity based on the level of depressive symptoms experienced by the patient. Step-Dep was compared to usual care. Case finding of subthreshold depression was originally planned to be done using two different screening tools: first, the PHQ-2 for initial screening, and second, the PHQ-9 to further screen all patients with a positive result on the PHQ-2. However, due to practical restraints (i.e. a large number of false positives and a too complicated procedure to be able to administer a diagnostic interview to rule out depression within an acceptable time after completion of the PHQ-2), the PHQ-2 was dropped from the screening strategy shortly after the start of the study. The effectiveness and cost-effectiveness of the Step-Dep intervention as compared to usual care are reported in Chapters 6 and 7. In the effectiveness study (Chapter 6), the 12-month effect of Step-Dep on the
cumulative incidence of major depressive disorder was small (a difference of 0.2 percent in favor of the usual care group) and not statistically significant. Also, no effect of Step-Dep was found on the secondary outcomes (depression severity, anxiety and perceived recovery). Furthermore, Chapter 7 shows that Step-Dep was not cost-effective in comparison with usual care to prevent depression, improve health related quality of life, reduce depressive symptoms over time or enhance perceived recovery of patients.

Several preventive stepped care studies in other populations found similar results regarding effectiveness. For example, a study from Hong-Kong among primary care patients with subthreshold depression and/or anxiety found no difference between stepped care and usual care in cumulative incidence of major depressive disorder after 15 months [26]. Also, in older adults from primary care no effect was found [27]. In community dwelling older adults only a small short-term effect on depression severity was found in favor of stepped care after three months, but this effect was not lasting and had disappeared after six months [28, 29]. In contrast, stepped care was found to be effective in preventing the onset of major depressive disorder in older primary care patients [30-32], in older patients who lived in a nursing home [33] and in visually impaired older adults [34].

The results regarding cost-effectiveness of Step-Dep are in contradiction with other cost-effectiveness studies. Stepped care to prevent major depressive disorder proved to be cost-effective in comparison with usual care in older general practice patients after 12 months [35]. A similar minimal psychological intervention for chronically ill primary care patients was found to be cost-effective in comparison with usual care as well [36]. Anyway, in this field, it seems that few interventions show effect sizes over 0.3, which is clinically only marginally relevant (i.e., the mean difference between groups at the end of a trial is 1.6 points on a range of 0-27, assuming a mean of 10 and a standard deviation of 5) [30, 34].

A possible explanation for these mixed results is that the incidence of depression among patients participating in the Step-Dep study was much lower than in earlier studies [30, 33-36]. This may be explained by the fact that our screening strategy identified patients who were less at risk to develop major depressive disorder than in studies using other screening strategies. For example, van ‘t Veer et al. used ‘double’ screening where patients were only eligible for participation in the study when they had two consecutive elevated depression scores on the screening tool instead of relying on one assessment [30]. It is possible that by using such a strategy risk stratification could have been better. Also, to screen for elevated depressive...
symptoms, a cut-off of 6 on the PHQ-9 was used, while in the same population a cut-off score of 8 appeared to be most appropriate to identify subthreshold (or minor) depression (Chapter 4). Since symptoms of DM2 and/or CHD may overlap with symptoms of depression, this higher cut-off score may have been more appropriate for identifying patients who were at high risk for major depressive disorder. Consequently, a large proportion of the sample may not have been experiencing clinically relevant depression symptoms and may not have been at higher risk for major depressive disorder. This explanation is in accordance with the previously described risk of over diagnosis of depressive symptoms when using a screening tool (see chapter 4) [21].

A possible alternative explanation for the low incidence of depression is that most studies in which incidences of psychopathology were higher, used both depression and anxiety as outcomes, whereas we focused on depression only. This may have resulted in study populations with more mental health complaints and therefore a higher risk to develop major depressive disorder.

Another possible explanation for the heterogeneity of study results lies in the difference in uptake rates of the intervention, which was relatively low in our study as compared to other studies. Only 40% (n=36) of all patients experienced prolonged depressive symptoms to be offered a sequential treatment step after watchful waiting. Of those, 70% (n=25) accepted one or more offered treatment steps. That means that almost one third of patients who were eligible for active interventions declined to participate. These patients did not drop out of the study and continued to participate in the measurements. This was also true in studies that did not show any effects [26-29]. In contrast, in all other studies that found a stepped care intervention (cost)-effective in preventing major depressive disorder as compared to usual care, the uptake of the intervention was much higher [30, 33-36]. Thus, the uptake of the intervention may have influenced the outcomes and should be improved in future studies.

Additionally, Step-Dep was compared to usual care. For primary care patients with DM2 and/or CDH in the Netherlands usual care is already relatively enhanced. Patients are seen by their general practitioner or primary care nurse on a regular basis. These caregivers are already instructed to question these patients about their mental health. As a result, also patients in the usual care condition have easy access to (mental) health care options. This may have diminished contrast between the Step-Dep intervention and usual care. Also, patients in the Step-Dep intervention possibly did not perceive the need for an additional program next to the care they already
received. This can explain the previously described low uptake of the intervention. Combined, these explanations are most plausible as to why no effect was found in the Step-Dep study, while other comparable studies demonstrated (cost-)effectiveness. It is unsure whether there is a need for stepped care to prevent depression in the population at hand.

The findings from the Step-Dep study can answer the third central question of this thesis: **Is the Step-Dep program (cost-)effective as compared to usual care in preventing the onset of depression in patients with DM2 and/or CHD?**

Step-Dep was neither effective nor cost-effective in comparison with usual care to prevent major depressive disorder in primary care patients with DM2 and/or CHD who screen positive for subthreshold depression. No significant effect on secondary outcomes (severity of depressive symptoms, severity of anxiety symptoms, perceived recovery and health related quality of life) was found either. Therefore, it is not recommended to implement the Step-Dep protocol as studied into primary care. However, definite conclusions about the effectiveness and the cost-effectiveness of different forms of stepped care to prevent major depressive disorder in this population cannot be drawn. Perhaps the case finding strategy identified patients with clinically less relevant types of subthreshold depression, and who as a result did not perceive the need for depression care. This might have resulted in the relatively low uptake of the intervention and low contrast between treatment groups. Evidence from Part 2 of thesis strengthens this hypothesis, demonstrating high false positive rates when screening for clinically relevant depressive symptoms using the PHQ-9 with a cut-off of 6. The general problem in this field seems to be that few interventions show effect sizes over 0.3, which is clinically only marginally relevant (i.e., the mean difference between groups at the end of a trial is 1.6 points on a range of 0-27, assuming a mean of 10 and a standard deviation of 5). When this is the main reason for Step-Dep to be ineffective, another case finding strategy targeting patients who really are at risk for major depression may be helpful. On the other hand, it is possible that Step-Dep was not effective because the experienced depression symptoms do not usually develop into major depression in this population. Finally, because usual care is already relatively enhanced in the Netherlands, patients do not need an extra care program to prevent depression. In those cases, it is not feasible to further research or use the Step-Dep intervention in primary care patients with DM2 and/or CHD.

Based on our current findings and existing evidence, it seems that the key to improving the results of stepped care to prevent depression in patients with subthreshold depression and CHD and/or DM2 is to identify the patients who are at
highest risk of developing major depression and experience a need for preventive depression care. The main remaining question is whether these patients exist and can be found using a different case finding strategy.

**Methodological considerations**

This thesis aimed to contribute to the scientific knowledge regarding the course, measurement and prevention of depression in (older) primary care patients who also suffer from chronic illnesses such as DM2 and/or CHD. Different study designs were employed to unravel this: a prospective cohort study following the course of depression in late life, a systematic review of the literature, a validation study of a screening tool for depression and a cluster RCT. Within these designs, multiple different statistical techniques were used for data analyses: latent class growth mixture analysis, a multinomial regression prediction model, an association model based on linear regression, analyses of sensitivity, specificity and AUC, multilevel analyses to assess effectiveness and cost-effectiveness analyses of the cluster RCT data. Altogether, the data presented in this thesis provides multi-faceted answers to the posed central questions. To correctly interpret the presented results, it is important to acknowledge the following methodological considerations:

**The study design:** In Chapter 2, utilization of mental health care services and antidepressants was not taken into account in the estimation of the depression trajectories. This means that it is possible that patients in the more benign depression trajectory had more access to mental health care which resulted in a more positive course of depressive symptoms. Although it is impossible to entirely rule this out, this seems unlikely since post-hoc analyses revealed that in the more adverse trajectories patients are using significantly more mental health care resources. Also in Chapters 6 and 7 differences in care seeking behavior between treatment groups could have influenced the outcomes. Patients in both groups were not restrained from seeking mental health care outside of the Step-Dep study as a consequence of the pragmatic nature of the study. This may have reduced the contrast between the Step-Dep and usual care group. However, since information on cost-effectiveness is mainly used by health care decision makers for allocation decisions, it is important that results resemble a real primary care setting. This reduces internal validity as compared to a strictly controlled trial, but the generalizability of the results increases.

In Chapter 3, We deliberately excluded studies which only assessed diagnostic accuracy, since our review aimed to review the evaluative (within-patient) properties of the self-reported questionnaires. Consequently, the results only provide information
about instruments for measurement of severity of depressive symptoms and not for screening or diagnostic purposes. To be able to compare and recommend instruments on these discriminative qualities, a systematic review focusing on these aspects is necessary. Results presented regarding diagnostic accuracy (Chapter 4) only apply to the PHQ-9.

Due to the nature of the performed cluster RCT (Chapter 6 and 7), blinding of caregivers, patients and researchers to the intervention status was not possible. Also in Chapter 4, blinding of the researchers to the screening result of the PHQ-9 before administering the psychiatric interview was not possible due to practical reasons. This might have resulted in overestimation of the effects of the Step-Dep intervention and the diagnostic accuracy of the PHQ-9. In the case of the effectiveness and cost-effectiveness of Step-Dep, this does not seem to have caused bias. No effect was found and it seems highly implausible that Step-Dep would have worsened the situation in comparison to care as usual due to the nature of the intervention. For the assessment of the diagnostic accuracy of the PHQ-9 not blinding the interviewer who administered the reference standard (Mini International Neuropsychiatric Interview (MINI) [37]) to the results of the PHQ-9 possibly caused bias. This would have been true when the interviewer was more reluctant to score positive answers on the MINI when a patient had a low PHQ-9 score or vice versa. Since we were not able to check whether or not this actually occurred, the results should be interpreted with caution.

In the cluster RCT (Chapter 6 and 7), the interval between measurements and the duration of the follow-up period (12 months) may have introduced bias. First, the recall period for the questionnaires addressing health care utilization and productivity losses was three months to cover the full period between measurements. This may have resulted in some recall bias, although a recall period of three months in combination with closed questions is considered reliable [38]. However, we expect that any bias that possibly occurred is similar in both groups, and therefore does not affect the difference between the groups. Second, due to limited staff availability and time restraints diagnostic interviews were only administered every six months. As the median of spontaneous remission (i.e., without any treatment) of a major depressive episode seems be three to six months [39], it is possible that incident major depression has remained unnoticed in several patients. This might provide an additional explanation for the low incidence of major depressive disorder found in the Step-Dep study. When this is the case, and the missed incident depression episodes have occurred more often in one of the treatment groups, results could have been influenced. However, since the PHQ scores (that were
measured every three months) did not show fluctuations hinting at this scenario, it is
not very plausible this issue has influenced our outcomes.

**The study population:** Chapter 4 shows that using a cut-off score of 6 on the PHQ-9 to
identify patients with subthreshold depression may have resulted in overdiagnosis
of subthreshold depression and, thus, in an overestimation of the risk of developing
major depressive disorder within the study population (Chapter 6 and 7). A possible
consequence is the lower than expected incidence of major depressive disorder in
the study population. This may have reduced the need for depression care, thereby
diminishing the contrast between treatment groups. As shown in Chapter 4, a cut-off
score of 8 on the PHQ-9 might have been more appropriate. Yet, this alternative cut-off
is very close to the most appropriate cut-off score for major depressive disorder,
which is 10. Using a cut-off of 8 instead of 6 might have resulted in a very narrow
window to detect subthreshold depression reducing the number of patients eligible
to participate. Another solution might have been the use of a second screening
moment, as previously employed in a study where higher rates of incident major
depressive disorder were found [30]. This way it could have been ensured that only
patients with lasting symptoms would be selected to participate. However, in the
Step-Dep study this was not feasible, due to very small yield and time restraints.
These problems were also previously described by van Weele et al. [40].

Even considering the above, it can be difficult to overcome the problem of correctly
identifying patients with subthreshold depression, especially in the specific
population of patients with DM2 and/or CHD in our cluster RCT. Symptoms of
depression and chronic somatic illness, such as fatigue, problems concentrating,
problems sleeping and altered eating and appetite, overlap. Consequently, scores
on depression screening questionnaires may be easily elevated without patients
experiencing depressive symptoms. Unfortunately, this not only applies to screening
questionnaires, but also to frequently used reference standards (such as the MINI)
since these are based on the same reflective model of depression (ea. list of
symptoms that together comprise the construct of depression [41]). Following this
argumentation, results of diagnostic accuracy of instruments to detect depression
that seem to be acceptable when a screening questionnaire such as the PHQ-9 is
compared to a commonly accepted reference standard such as the MINI (Chapter
2) might be inflated. After all, also the reference standard constitutes a list of
symptoms that might overlap with symptoms of chronic somatic illness and may not
be accurately diagnosing depression in a chronically ill patient group [20]. This poses
a problem, since no gold standard is available to diagnose major depressive disorder.

Considering these challenges in measuring subthreshold and major depression in patients with DM2 and/or CHD, the results of the presented work must be interpreted with care.

**Power of the performed analyses:** Chapters 2 and 7 comprise analyses in which health care costs and lost productivity costs are included as an outcome. Although in both chapters considerable differences in costs are found between the compared groups, these differences are not statistically significant. This is likely due to the fact that both studies were not originally powered for costs. This increases chances of a type 2 error, thus possibly concluding there is no difference between the compared groups, while in reality there is a difference.

Also in Chapter 2, it is difficult to derive practical implications from the results since two of the identified trajectories are based on small groups of patients. On the other hand, despite two of the trajectories being very small, statistically significant predictors have been established and a significant association with health-related quality of life was found. This indicates that – at least for these predictors and this outcome – a type 2 error can be ruled out and the results are a promising venture point for further exploration.

Finally, since the cumulative incidence of major depressive disorder in the cluster RCT (Chapter 6 and 7) was much lower than expected, insufficient power possibly also compromised the validity of these analyses. However, since the absolute observed difference between treatment groups was close to zero, it is unlikely that a clinically relevant difference has been missed.

**Generalizability:** The pragmatic design of the cluster RCT (Chapter 6 and 7) improved the external validity of the results of the trial considerably. By closely mimicking everyday practice, it is likely that the results of the Step-Dep study are generalizable to primary care. Therefore, it is possible to make a firm recommendation not to implement the current Step-Dep program as it stands because it is not effective, nor cost-effective in comparison to usual care.

To be able to correctly interpret the results of the economic evaluation of Step-Dep, it is important to consider the following generalizability concerns:

In cost-effectiveness analyses prizes and health care services are assessed that are
available within the health care system of the country (or even region) in which the intervention is implemented. This can be substantially different between countries. The results of the presented economic evaluation are therefore only valid in the Dutch setting [42, 43].

Costs and outcomes of the Step-Dep program are measured from a societal perspective, meaning all cost and outcomes are taken into account, regardless of who eventually pays for them or benefits from them [42]. Although this is the recommended perspective in the Netherlands, in the UK the health care perspective is preferred. Therefore, comparability of study results with UK studies may be difficult. However, employment of the societal perspective allows for sensitivity analyses using a health care perspective [44]. In the current study, adoption of a health care perspective resulted in comparable findings.

An important generalizability issue on the latent class growth mixture model to identify different trajectories of late life depression (Chapter 2) needs to be discussed. The used technique is data driven, which means that the model is estimated to match the data at hand, without a pre-imposed theory about the number and course of the different trajectories [45]. This can be viewed as a strength, since estimating trajectories that have unique associations with other variables in the model is facilitated. These associations are not on forehand specified and can thus be estimated to closely match the observations, making as much use of available information as possible. On the other hand, the identified trajectories possibly apply to the current sample only, enlarging incidental patterns that only apply to the current sample, but do not exist in the target population. A systematic literature review in which studies estimating data-driven subtypes of depression in the general population were validated, shows that this problem may occur in the estimation of depression trajectories [46]. The small number of patients in two of the found trajectories (Chapter 2) increases the risk of incidental findings that match only the current sample. External validation is, thus, still needed and this should be kept in mind when interpreting the findings presented in Chapter 2.

Also, the estimation of most optimal cut-off scores on the PHQ-9 to identify patients at risk for subthreshold depression or major depressive disorder is data driven. Therefore, the same risks apply as described above. It is clear that the cut-off scores work best in the study sample, but in the target population this might be different [47]. Given the fact that in comparable studies different cut-off scores are found, it is import to determine the most appropriate cut-off scores across study populations in a meta-analysis or an external validation study.
In the systematic review on measurement properties (Chapter 3) of self-reported depression questionnaires in diabetes patients, two generalizability issues arise that are important to consider when interpreting the results. Firstly, the included questionnaires are studied in a large variety of languages, cultures and settings, which makes it hard to combine results and draw one conclusion regarding a measurement property of an instrument. The fact that very few studies performed a proper cross-cultural validation of translated questionnaires makes this even harder. We tried to combine the existing evidence in the best possible way, but it is important to consider this issue while interpreting the results. Secondly, there was no questionnaire of which all measurement properties were researched. Especially data are missing on measurement error, responsiveness and interpretability. To draw definitive conclusions it is important to have complete information on all measurement properties for all questionnaires [13].

**Implications for research**

External validation of the presented findings on depression trajectories in older adults is necessary. Only by external validation it can be assured that the trajectories identified, their predictors and associated outcomes exist in the real world. While performing the external validation, it is key to use a large sample, overcoming power issues that arise when one or more of the trajectories are relatively small. Also, it would be interesting to assess whether current results from the general older primary care population are transferrable to a population of patients with DM2 and/or CHD.

To definitively recommend a patient reported depression questionnaire for evaluation of depressive symptoms in patients with diabetes, it is necessary to conduct validation studies on the measurement properties that were not yet researched: reliability, responsiveness and interpretability [13]. It is also important to assess whether the recommended questionnaires (CES-D, PHQ-9 and WHO-5) were performed in the same way across languages, cultures and health care settings. Complementary, research is needed to assess the performance of depression questionnaires in patients with CHD and other comorbid somatic disorders.

Furthermore, it is important to evaluate strategies that can best be used in primary care patients with DM2 and or CHD to identify patients at high risk for clinically relevant depression. Issues that need to be clarified are: a) which instruments can be validly used and b) identification of reliable cut-off scores for the best performing questionnaires. This can best be done with a systematic review of the existing
literature which includes meta-analyses to compare frequently used instruments and define cut-off scores that are valid across study samples.

Finally, the following recommendations for future studies can be made:

- To perform a systematic review of the existing literature with a meta-analyses to establish which depression questionnaire(s) to screen for subthreshold depression can be used best and to determine the most appropriate cut-off score.
- To use two consecutive measurements to include patients with elevated depression scores in a study to minimize the risk of selecting patients with a false positive score and/or transient depressive symptoms;
- To study which casefinding strategy works best to identify patients with a need for preventive depression treatment and to check whether patients indeed perceive a need for preventive depression care before they are included in a study.

Implications for practice
At present, the CES-D, or alternatively the PHQ-9 or the WHO-5, is recommended for monitoring or evaluating the course of depressive symptoms of a patient in primary care. This recommendation could be altered when more information regarding the measurement properties of these questionnaires becomes available in the future.

To identify patients at risk for subthreshold- or major depressive disorder, a screening tool, such as the PHQ-9, can be used. It is not yet advised to use the PHQ-9 for the whole primary care population with DM2 and/or CHD as part of routine practice, since too little is known about the most effective screening procedure. When patients at risk are identified using a screening questionnaire, it is important to always confirm a possible diagnosis with preferably a validated psychiatric interview performed by trained personnel, or by making use of other up-to-date guidance in a structured way and use careful monitoring.

In practice, the presented findings in Chapter 2 implicate that it can be beneficial to monitor depressive symptoms closely in older adults and offer depression care timely. This way it can possibly be prevented that depressive symptoms are already severe when treatment starts. Timely treatment is important, because the presented work shows that the strongest predictor for an adverse depression trajectory is high severity of initial depression symptoms. Also, using this strategy, adverse health outcomes such as loss of quality of life, declining medication adherence and a rise in
health care costs can possibly be averted. However, it is yet unclear which protocol should be used to find patients who are in need of an early intervention for depression. Current literature suggests that other intervention methods than stepped care can be used for preventive depression treatment in the general population [48], and in older patients [49]. For now, in practice, it is not recommended to use Step-Dep as a preventive depression treatment among patients with DM2 and or CHD.

**General conclusion**

Late life depression is likely to be chronic. Patients who have more severe symptoms at the time of detection of depression seem to be most at risk for an adverse course and poor depression outcome. Intervening early, when patients do not yet have severe depression, may be an opportunity to reduce the personal and societal burden associated with depression. However, to be able to do so, it is important to identify older patients who have subthreshold depression at an early stage. Measurement of depressive symptoms poses a challenge in (older) patients with chronic illnesses, such as DM2. Based on current evidence, the CES-D seems most appropriate to evaluate depressive symptoms among diabetes patients using patient reported outcome measures. Alternatively, the PHQ-9 or the WHO-5 can be used. However, evidence for the validity of self-reported questionnaires for diagnosing clinically relevant depression is limited. Therefore, using only questionnaires for diagnostic purposes is not recommended.

In its current form the Step-Dep program, which was especially tailored to answer to the need of patients with subthreshold depression in combination with DM2 and/or CHD, was neither more effective nor cost-effective compared with usual care. In view of all the studies that have been performed, we probably need to go back to the drawing board to develop better case finding strategies to identify patients in need of preventive depression treatment in primary care and a care program that is more tailored to the health perception of patients. Whether the concept of indicated prevention, with its limited uptake and limited effects will eventually be considered worth the investment and a public health improvement is still an open question.
REFERENCES


