SUPPLEMENTARY FILE 3

RESULTS OF THE FEASIBILITY PROJECT

Selection of trials and data collection for the feasibility project

The abstracts and titles were independently screened by two reviewers (SE and AS) and the full texts by SE only. However, when the suitability of a publication was in doubt, it was double-checked by AS, and disagreements between the reviewers were resolved by discussion. Consensus on inclusion or exclusion was reached for both abstracts and full text reviews.

Results of the feasibility project

Identification of studies

After removing duplicates, 512 papers were identified during the initial search. Of those, 426 were excluded in the abstract screening process. The full texts of the remaining 86 studies were screened, and a further 65 studies excluded as a result; 21 papers ultimately fulfilled the inclusion criteria (see Figure 1 at the end of the text). Full texts of the excluded studies and information on the author and reasons for exclusion are shown in Supplementary File 1.

Eighteen authors were contacted for further information on the ICCs used in their studies; three authors had already provided the necessary information in their publications. Of those 18, four then provided the relevant information (see Supplementary File 2).

Description of the studies

The reported criteria in the studies were grouped thematically (see Table 1 at the end of the text).

The 21 included studies were published between 1995 and 2013; six of them were published in 2013 alone. Eight of the studies were from the United Kingdom, followed by the Netherlands and the USA (three studies each), and most of them were published in the British Medical Journal (13 studies).

The most common groups of patients were patients with respiratory and mental diseases (six studies each). Patients with diabetes mellitus were examined in four studies and five involved elderly persons.

All studies examined complex interventions but of different levels of complexity. Most of them dealt with interventions that aimed to improve outcomes by means of a multifaceted program. They also differed in terms of the persons delivering the intervention, who were either general practitioners or specialized nurses (see Table 2 at the end of the text).

Outcome measures

Our analysis revealed that the majority of the studies (67%) could not show an intervention effect on the primary patient-relevant endpoint (see Table 3 at the end of the text). Of the 21 examined studies, 14 could not demonstrate such an effect, while three studies did reveal an intervention effect on the primary patient-relevant endpoint. The feasibility project also identified four studies that had more than one primary outcome and showed effectiveness as well as ineffectiveness, depending on the endpoint (referred to as "partly effective"). As we discovered potential differences in quality between c-RCTs that may to some extent determine whether results come out in favour of a complex intervention, we decided to exclude these studies from the review. Exactly which interventions showed an effect and the size of these on primary patient-relevant outcomes are described in Table 5 at the end of the text.

Differences in study quality between studies with and without an intervention effect

As far as general information is concerned, the criteria "patient consent" and "ethical approval" were reported in the majority of trials (86 and 71%, respectively) that were unable to show an effect on the primary outcome. Less frequently, the details on consent of clusters (36%), publication of a study protocol (43%), and trial registration (36%) were provided. In comparison, studies that found a significant intervention effect more frequently provided four of the five listed criteria (see Table 4 at the end of the text).

Some quality criteria concerning the sample size calculation were provided in 86% of studies that showed no superiority. However, consideration of the ICC and involvement of the cluster size in the sample size calculation were described less frequently (64% and 50%, respectively) and only few studies (21%) provided information on whether the cluster size was identical at baseline. In studies

showing a significant effect on the primary endpoint, this information was provided in full, with the exception of the identical cluster size at baseline (see Table 4 at the end of the text).

The method of randomization was only presented clearly in 64% of the studies that showed no intervention effect but in all studies that demonstrated superiority. However, irrespective of the significance of the primary outcomes, all other criteria in this category (recruitment and identification bias, allocation concealment, blinding (patients and outcomes)) were either reported poorly or not at all (see Table 4 at the end of the text).

In terms of analysis method, most of the studies that showed no intervention effect dealt with patient drop-outs (86%) and clusters (71%), performed ITT analyses (71%) and generally accounted for clustering in the analysis (86%). In studies showing an intervention effect on primary outcomes, 67% presented information on cluster drop-outs, and all other quality criteria that were mentioned were reported completely (see Table 4 at the end of the text).

Limitations

No conclusions can be drawn as to whether or not c-RCTs conducted in a general practice setting more often fail to show the effectiveness of a complex intervention due to methodological shortcomings. Our feasibility test did not enable us to rule out that intervention effects were simply lacking, i.e., an intervention was just not effective or not effective enough. But despite our limited sample, we were able to point out some aspects which will be investigated systematically in the planned full review. Secondly, we must consider that the included studies may reflect selection bias, as we only searched for c-RCTs in certain types of journal - the aim of the full review is to correct for this and to achieve an unbiased view. Thirdly, the limited number of included c-RCTs did not allow us to prioritize from among different CONSORT items and to ascertain the methodological quality of the trial: e.g. methods after trial commencement (the way in which an intervention is delivered and implemented and whether or not the investigators defined its fidelity) may be more important than whether the term "cluster randomized trial" appeared in the title. Fourthly, our

feasibility trial did not comprehensively examine methodological shortcomings that concern the gradual development and evaluation of a complex intervention. Thus it did not attempt to answer such questions as (1) whether a study had a sound theoretical foundation, (2) whether the piloting of the intervention components, outcomes and processes justified confidence in the feasibility of the project, (3) whether the effectiveness of the intervention had been appropriately evaluated, and (4) whether process evaluation had been well planned a priori. The full review will therefore have to take these more specific aspects into consideration by examining the framework of the c-RCT, the fidelity of the intervention, and barriers and facilitators to its implementation.

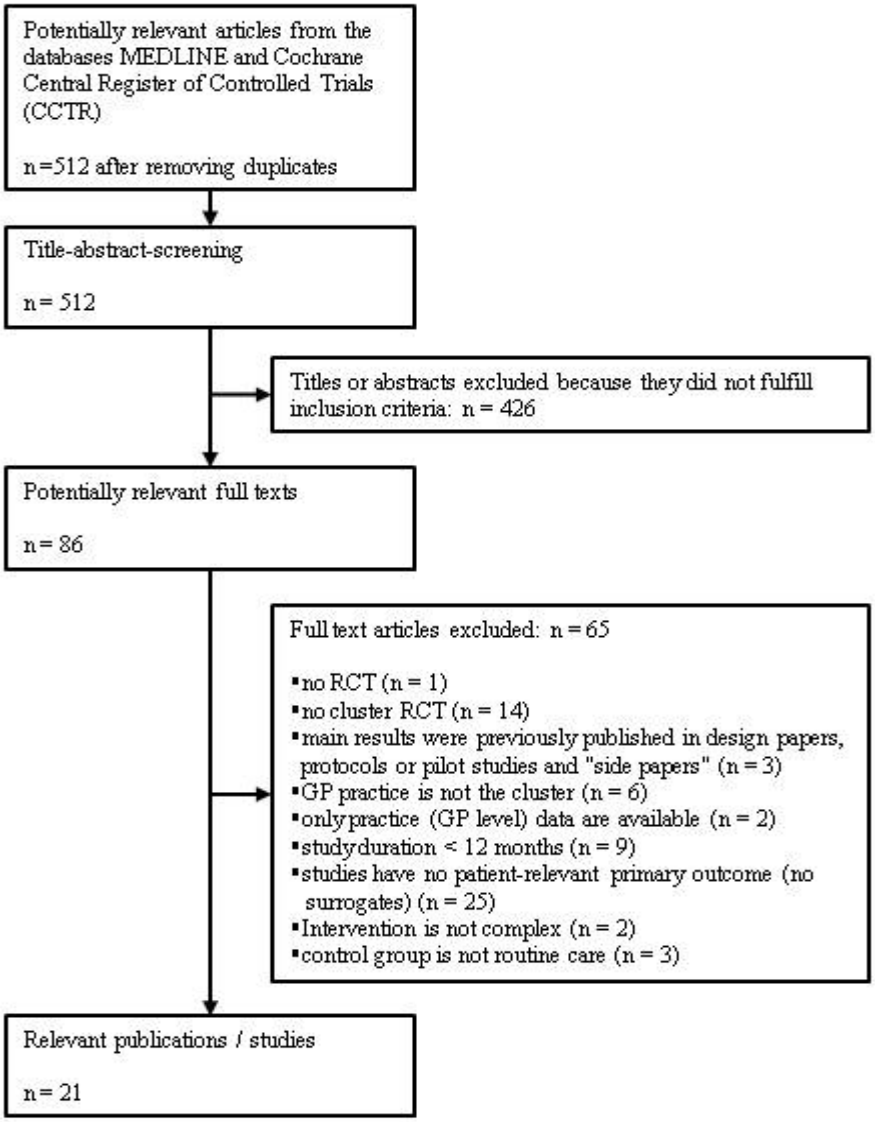

Figure 1: Result of the literature research

Table 1: Data extraction form

General information	Title, authors, journal, date of publication, country of publication, funding/conflict of interests (according to the author) and c-RCT evident from title
Study characteristics	Study design, objective (including target population/ health condition of the subject group), setting of the study, number of participating practices, cluster and cluster size (number of clusters screened, randomized and analysed), patients (number of patients screened, included, analyzed and lost to follow up), patient-relevant primary endpoint (s), not patient-relevant primary endpoint (s), patient-relevant secondary endpoint (s), not patient-relevant secondary endpoint (s)
Baseline Data	Baseline characteristics (cluster and patients), age, ethnicity, sex of the patients, disease-orientated information, inclusion criteria (cluster and patients)
Intervention Data	Run-in phase, contents of the intervention and control groups, recruiting period, follow-up period, observation period for the intervention and control groups
Outcome Data	Intervention effects on primary endpoint(s) including significance level, intervention effects on secondary endpoint(s), intra-cluster correlation coefficient (Are ICCs calculated for the primary endpoint or is information available on the effect of the design?), results of sub-group analyses, p-values (for baseline data)
Quality of the studies - general	Ethical approval, trial registration, sample size calculation method, recruitment method (cluster and patient level), consent ((clusters and patients), before or after the randomization of the practice), publication of the study protocol, involvement of the cluster in the 1st sample size calculation and 2nd. analysis, generalizability of the results for cluster and patients (according to the author), identical cluster size at baseline, recruiting/identification bias (possibility of bias adopted according to Eldridge 2008: not possible, unclear or unlikely)
Quality of the studies - risk of bias	Appropriate randomization method (acceptable: random number table, computer-generated random numbers, minimization, inappropriate: coin flip), acceptable allocation concealment (central allocation and sequentially numbered, opaque, sealed envelopes) blinding (open, blind, double-blind), dealing with drop-out (clusters and patients), intention to treat analysis (ITT), other potential bias (according to the author)
Authors' own interpretation/ explanation	Extraction of reasons why their studies did not show a positive effect e.g. loss to follow up, issues related to recruitment, adherence and data collection (outcomes).

Table 2: Description of the included studies

First author and year	Journal	Publication country	Target population / health condition	Aim / Objective
Bould 2013	Journal of General Internal Medicine	USA	Elderly people	To assess patients' functional health when guided care teams provide proactive, coordinated, comprehensive care
Byng 2004	British Journal of General Practice	United Kingdom	Mental illness	To determine patient satisfaction with care and patient perceptions with regard to unmet needs in the Mental Health Link program designed to improve communication between the teams and systems of care within general practice
Cartwright 2013	British Medical Journal	United Kingdom	COPD, diabetes, heart failure	To assess the effect of second generation, home-based tele-health on health-related quality of life, anxiety, and depressive symptoms
Elley 2003	British Medical Journal	New Zealand	Elderly people	To assess the long-term effectiveness of the green prescription programme on quality of life. The program provides advice on physical activity in a general practice setting
Gallo 2007	Annals of Internal Medicine	USA	Elderly people	To test whether an intervention to improve depression care can influence the risk of death
Gensichen 2009	Annals of Internal Medicine	Germany	Depression	To determine the effects of case management provided by health care assistants in small primary care practices on depression symptoms
Griffiths 2004	British Medical Journal	United Kingdom	Asthma	To determine the influence of specialist asthma nurses in a deprived multi-ethnic area on the percentage of participants attending a practice for unscheduled asthma care, and the time to first attendance for unscheduled asthma care the year after the intervention
Guldin 2013	Family Practice	Denmark	Relatives of patients after death by cancer	To test whether the implementation of a bereavement management program improves the general practitioner's ability to identify complicated grief and provide clinical care
Jarmann 2002	British Medical Journal	United Kingdom	Parkinson`s disease	To determine the effects of community-based specialist nurses on specific measures of health and patient well-being
Jellema 2005	British Medical Journal	Netherlands	Unspecific low back pain	To compare the differences between a minimal intervention strategy and usual care on the treatment of (sub) acute lower back pain on functional disability
Kennedy 2013	British Medical Journal	United Kingdom	Diabetes, COPD, irritable colon	To determine the effectiveness of an intervention to enhance self management support for patients with chronic conditions on generic health-related quality of life
Kerse 1999	British Medical Journal	Australia	Elderly people	To establish the effect of an educational intervention for general practitioners on the functional status of patients
Kinnersley 1999	Family Practice	United Kingdom	Dermatologic, orthopaedic, gynaecologic, rheumatic, ophthalmologic diseases	To describe whether in-house referral is practicable and acceptable for patients and whether it improves patient health outcomes and management in primary care
Metzelthin 2013	British Medical Journal	Netherlands	Elderly people	To evaluate the effect of an interdisciplinary primary care approach on disability
Murphy 2009	British Medical Journal	Ireland	Coronary heart disease	To test the effectiveness of a complex intervention designed within a theoretical framework on the rate of admissions to hospital and physical and mental health status

First author and year	Journal	Publication country	Target population / health condition	Aim / Objective
Olivarius 2001	British Medical Journal	Denmark	Diabetes	To assess the effect of a multifaceted general practice intervention on overall mortality and the patient's disease
Rubenstein 2006	Journal of General Internal Medicine	USA	Depression	To evaluate the effects of EBQI (evidence-based quality improvement) - a method for practices to self-improve depression care performance - on depression care and outcomes
Steventon 2012	British Medical Journal	United Kingdom	Diabetes, COPD, heart failure	To assess the effect of home-based tele-health interventions on the rate of admissions to hospital
Van Marwijk 2008	British Journal of General Practice	Netherlands	Depression	To test the effects of an intervention program that aims to improve the identification, diagnosis, and treatment of depression
Walters 2013	British Medical Journal open	Australia	COPD	To assess the benefits of telephone-delivered health mentoring on health-related quality of life
White 1995	British Medical Journal	United Kingdom	Asthma	To test the effects on classic patient symptoms of feeding back information on patients' asthma to primary care teams

Table 3: Effects on primary patient-relevant outcome (most recent studies first)

Studies	Effect on primary patient-relevant endpoint(s)[1]	Primary patient-relevant endpoint(s)[2]
Boult 2013	↔	Patients' functional health (-)
Cartwright 2013	↔	Treatment effectiveness (-) Treatment efficacy (-)
Guldin 2013	↔	Bereaved relatives' score (-) Relative's number of contacts with general practice (-)
Kennedy 2013	↔	Generic health-related quality of life (-)
Metzelthin 2013	↔	Disability (-)
Walters 2013	↔	Health-related quality of life (-)
Van Marwijk 2008	↔	Montgomery Åsberg Depression Rating-Scale (-) PRIME-MD Scores (-)
Rubenstein 2006	↔	Appropriate depression treatment (-) Recovery from depression (after 12 months) (-)
Jellema 2005	↔	Functional disability (-)
Byng 2004	↔	Patient satisfaction with care (-) Patient perceptions on unmet need (-)
Olivarius 2001	↔	Overall mortality (-) Incidence of diabetic retinopathy (-) Myocardial infarction (-) Stroke in patients without symptoms at baseline (-)
Kerse 1999	↔	Functional status (-)
Kinnersley 1999	↔	Patient satisfaction (-) Health status (-) Management in primary care before and after referral (-)
White 1995	↔	Classic symptoms (-)
Steventon 2012	↑	Proportion of people with an inpatient admission to hospital within the 12 month trial period (+)
Gensichen 2009	↑	Depression symptoms (+)
Griffiths 2004	↑	Percentage of participants attending for unscheduled asthma care (+) Time to first attendance for unscheduled asthma care in the year after the intervention (+)
Murphy 2009	↑/↔	Admissions to hospital (+) Changes in physical and mental health status (-)
Gallo 2007	↑/↔	Mortality: All patients with depression and major depression disorder (+) Clinically significant minor depression and patients without depression (-)
Elley 2003	↑/↔	Quality of life: General health, role physical, vitality, bodily pain (+) Physical functioning, social functioning, role emotional, mental health (-)
Jarmann 2002	↑/↔	Measures of health (-) Patient wellbeing (-) Global health question (+)

1 (↑): Upward arrow: Studies showing an intervention effect; (↔): Horizontal arrow: Studies showing no effect; (↑/↔): Studies presenting more than one primary patient-relevant endpoint with an effect on one or more endpoints but not on all of them within one and the same study

2 (+): Superiority of intervention group for a patient-relevant endpoint demonstrated; (-): No superiority of intervention group for a patient relevant-endpoint demonstrated

Table 4: Differences in study quality between studies with and without an intervention effect on the primary outcome

Study Information	Studies without intervention effect n=14 (% in brackets)	Studies with intervention effect n=3 (% in brackets)
General information		
Consent (patients)	12 (86)	3 (100)
Consent (cluster)	5 (36)	2 (67)
Ethical approval	10 (71)	2 (67)
Publication of study protocol	6 (43)	2 (67)
Trial registration number	5 (36)	2 (67)
Sample size calculation		
Sample size calculation	12 (86)	3 (100)
Assumed ICC	9 (64)	3 (100)
Involvement of the cluster in the sample size calculation	7 (50)	3 (100)
Identical cluster size at baseline	3 (21)	1 (33)
Randomization and blinding process		
Recruiting-/Identification bias	1 (7)	0 (0)
Adequate randomization method	9 (64)	3 (100)
Adequate allocation concealment	2 (14)	1 (33)
Blinding (patients)	4 (29)	1 (33)
Blinding of outcomes assessors	7 (50)	1 (33)
Analysis		
Dealing with drop-out (patients)	12 (86)	3 (100)
Dealing with drop-out (cluster)	10 (71)	2 (67)
ITT	10 (71)	3 (100)
Involvement of cluster in the analysis	12 (86)	3 (100)

Table 5: Which interventions showed an effect and the size of the effects on primary patient-relevant outcomes

Studies	Intervention effects on primary patient-relevant outcomes (with significance level)
Boult 2013	Patients' functional health: Physical Health: Difference Guided Care/Usual Care: -1.31 (CI: -3.02-0.41) Mental Health: Difference Guided Care/Usual Care: 1.05 (CI: −1.08-3.12) (adjusted for baseline age, race, sex, education level, financial status, habitation status, HCC score, SF-36 physical and mental health subscales, and satisfaction with health care)
Byng 2004	Patients' satisfaction with care: Adjusted difference between control and intervention at follow-up: -0.01 (CI: -0.21-0.18; P=0.88) (controlling for baseline scores and allowing for clustering of patients within practices) Patients' perception of unmet need: Adjusted difference between control and intervention at follow-up: -0.02 (CI: -0.56-0.51; P=0.94) (controlling for baseline scores and allowing for clustering of patients within practices)
Cartwright 2013	Treatment effectiveness with intention to treat analysis (ITT): No significant differences between the groups for the patient-relevant outcomes quality of life, depression symptoms and anxiety Complete case: $0.480 \leq P \leq 0.904$ Available case (baseline data and data of one other assessment): $0.181 \leq P \leq 0.905$ Treatment efficacy with per-protocol analysis: No significant differences between the groups for the patient-relevant outcomes quality of life, depression symptoms and anxiety Complete case: $0.273 \leq P \leq 0.761$ Available case (baseline data and data of one other assessment): $0.145 \leq P \leq 0.696$
Elley 2003	Quality of life: Difference between groups (adjusted for clustering by medical practice): general health: 4.51 (CI: 2.07-6.95; P=0.000) physical fitness: 7.24 (CI: 0.16-14.31; P=0.045) vitality: 2.30 (CI: 0.03-4.57; P=0.047) bodily pain: 4.01 (CI: 0.78-7.24; P=0.02) physical functioning: 1.23 (CI: -1.35-3.81; P=0.3) social functioning: 0.36 (CI:-3.53-4.26; P=0.9) emotional status: -0.38 (CI: -5.70-4.94; P=0.9) mental health: 0.98 (CI: -0.99-2.95; P=0.3)
Gallo 2007	Mortality: Hazard ratio for intervention effects (includes terms for baseline age, sex, education, smoking, cardiovascular disease, stroke, diabetes, cancer, cognition, and suicidal ideation): All patients with depression: 0.67 (CI:0.44-1.00) Major depression disorder: 0.55 (CI: 0.36-0.84) Clinically significant minor depression: 0.97(CI: 0.49-1.92) Patients without depression: 1.14 (CI: 0.84-1.53)
Gensichen 2009	Depression symptoms: Mean difference (P-value based on a 2-level linear mixed model for respective outcomes (T1 and T2), adjusted for intracluster correlation and baseline depression): -1.41 (CI: -2.49 to -0.33; P=0.014)
Griffiths 2004	Percentage of participants attending for unscheduled asthma care: Adjusted odds ratio (with clustering): 0.61 (CI: 0.38-0.99) Adjusted odds ratio (without clustering): 0.62 (CI: 0.38-1.01) Time to first attendance for unscheduled asthma care in the year after intervention: Hazard ratio: 0.73 (CI: 0.54-1.00)
Guldin 2013	Bereaved relatives' score - depression: Mean score, intervention group: 7.85 (CI: 6.53-9.17) Mean score, control group: 8.84 (CI: 7.41–10.28) Bereaved relatives' score - grief symptoms: Mean score, intervention group: 14.73 (CI:13.14-16.32)/ Mean score, control group: 15.57 (CI: 13.77-17.38) Relatives' number of contacts with general practice: Contact frequencies with GPs: Corresponding rate ratio: 0.92 (CI: 0.72–1.17); P=0.50 Out-of-hours contacts with GPs: Corresponding rate ratio: 0.55 (CI: 0.29-1.06); P=0.07

Studies	Intervention effects on primary patient-relevant outcomes (with significance level)
Jarmann 2002	Measures of health: Bone fracture during study: Odds Ratio: 1.20 (CI: 0.85-1.69); P=0.31 Mortality (2 years): Hazard ratio: 0.91 (CI: 0.73-1.13) P=0.38 Mortality (4 years): Hazard ratio: 0.89 (CI: 0.76-1.03); P=0.12 Patient wellbeing: Euroqol: Difference: -0.02 (CI: -0.06-0.02); P=0.30 PDQ-39 summary index: Difference: 0.47 (CI: -2.72-3.66); P=0.77 Global health question: Difference: -0.23 (CI: -0.40 to -0.06); P=0.008
Jellema 2005	Functional disability: Mean difference (adjusted for baseline values): 0.25 (CI: −0.77-1.28)
Kennedy 2013	Generic health-related quality of life: Adjusted mean difference (adjusted for model factors and covariates): -0.00 (CI: -0.02-0.01) Effect size (Adjusted mean difference (intervention minus control) divided by standard deviation in practice): -0.01 (CI: -0.05-0.04); P=0.72 P value for interaction with condition group (P value for test of whether intervention effect varies by disease condition): 0.31
Kerse 1999	Functional status: Mean effect size: 2.10 (CI: -0.94-5.1); P= 0,175 (All analyses were controlled for general practitioner billing status and effect of cluster design)
Kinnersley 1999	Patient satisfaction (mean): Intervention group (referred immediately to secondary care): 80.7 (SD: 11.1) Intervention group (not referred): 78.5 (SD: 12.2) Control group: 79.2 (SD: 10.3) Health status (mean): Intervention group (referred immediately to secondary care): 64.4 (SD: 33.5) Intervention group (not referred): 77.1 (SD: 27.9) Control group: 67.9 (SD: 29.6) Management in primary care before and after referral (mean): Intervention group (referred immediately on to secondary care): 0.25 (SD: 0.5) Intervention group (not referred): 0.56 (SD: 0.69) Control group: 0..36 (SD:0.65)
Metzelthin 2013	Disability (after 12 months): Mean difference (adjusted for age, sex, education, and significant differences at baseline (frailty and disability)): 0.47 (CI: -0.81 to 1.76); P=0.47
Murphy 2009	Admissions to hospital: Mean difference: -0.15 (CI: -0.01 to -0.29); P= 0.03 (ICC: 0.017) Changes in physical health status: Mean difference: -0.78 (CI: -2.58-1.03); P=0.39 (ICC: 0.076) Changes in mental health status: Mean difference: 0.02 (CI: -2.40 -2.35); P= 0.98 (ICC: 0.054)
Olivarius 2001	Overall mortality: P=0.82 Incidences of diabetic retinopathy: Odds ratio: 0.90 (KI: 0.53-1.52); P=0.69 Myocardial infarction: Odds ratio: 0.65 (KI: 0.31-1.35); P=0.25 Stroke in patients without these outcomes at baseline: Odds ratio: 0.89 (KI: 0.39-2.01); P=0.77
Rubenstein 2006	Appropriate depression treatment and recovery from depression (after 12 months): Effect size: 0.03; P=0.77 Intervention group (Mean): 45.6 (CI: 37.8-53.5) Control group (Mean): 47.0 (CI: 42.7-51.3) (All regressions controlled for covariates (age, sex, completion of high school, household wealth, timing of enrolment, ethnicity, count of chronic diseases, marriage, alcohol use, dysthymia) and baseline values of the dependent variable)
Steventon 2012	Proportion of people with an inpatient admission to hospital within the 12-month trial period: Unadjusted odds ratio: : 0.82 (CI: 0.70-0.97); P=0.017 Adjusted odds ratio: 0.82 (CI: 0.69-0.98); P=0.026 Combined model odds ratio: 0.82 (CI: 0.69-0.96); P=0.016
van Marwijk 2008	Montgomery Åsberg Depression Rating-Scale: Intervention group (mean): 10.80 (SE 2.85) Control group (mean): 10.09 (SE 2.50) PRIME-MD Scores: Intervention group (mean): 3.23 (SE 1.04) Control group (mean): 3.74 (SE 1.21)

Studies	Intervention effects on primary patient-relevant outcomes (with significance level)
Walters 2013	Health-related quality of life: SGRQ (mean): Intervention group: 41.9 (SD: 18.9) Control group: 40.5 (SD:17.4) SF-36 - Mental health component summary (mean): Intervention group: 50.2 (SD: 11.4) Control group: 50.5 (SD: 10.5) SF-36 - Physical component summary (mean): Intervention group: 38.5 (SD: 10.3) Control group: 38.5 (SD: 9.4)
White 1995	Classic symptoms: Breathlessness at least once a week (mean): Intervention group: 36.0 (SD:14.3) Control group: 35.0 (SD: 10.9);P=0.79 Wheeze at least once a week (mean): Intervention group: 38.0 (SD: 11.7) Control group: 31.0 (SD: 14.4); P=0.19 Cough at least once a week (mean): Intervention group: 49.0(SD:13.9) Control group: 45.0(SD: 12.1); P=0.47 Night waking at least once a week (mean): Intervention group: 27.0 (SD: 9.9) Control group: 23.0 (SD: 11.2); P= 0.39 Any time off work or studies due to asthma (mean): Intervention group: 16.8 (SD: 7.3) Control group: 19.1 (SD: 6.7); P=0.45 At least one severe attack (mean): Intervention group: 49.3(SD: 13.1) Control group: 43.3(SD: 13.2); P=0.3 Breathless on level ground (mean): Intervention group: 41.3 (SD: 17.0) Control group: 48.1 (SD: 13.0);P=0.7 Any attendance at surgery (doctor or nurse) (P=0.96), Regular use of inhaled steroids (P=0.62) and Regular use of inhaled bronchodilator (P=0.78)