Table S1: Description of 18 studies included in the review (* indicates those who showed a significant difference in patient outcomes)

First Author	Year	Country	Design/ Intervention	Analysis (unit of analysis/ power calculation)	Objective measurement/ Follow-up period	Successful Educational Aspects/Outcomes	Limitations
Tobe* [29]	2014	Canada	Prospective controlled cohort study. Comparison between immediate (5 practices) and delayed intervention (6 practices) groups with 1,201 and 1,654 patients respectively. The intervention was an evidence-informed inter-professional chronic disease management program consisting of a 2-day educational intervention with practice tools to implement the Canadian Hypertension Education Program's clinical practice guidelines. Also included a data repository to track patients.	To detect a difference of 5mmHg between the baseline period in the delayed intervention group (n=1654) and 9 months after the intervention in the immediate intervention group (n=1201), to have power of 0.9 with b¼0.10 and a¼0.05, 337 patients would be required in each group Unit of allocation – N/A. Unit of analysis – individuals.	Change in blood pressure (BP) from baseline to 9 months after the intervention between groups. BP measurements performed with validated automated office BpTRU monitors provided to each site (with training). It takes six readings, discarding the first and averaging the last five; recorded on patient's case form.	BP at baseline was 134.6/79.1mmHg (18.2/11.5) and at the end of 9 months of the program was 127.3/75.5mmHg (16.5/11.2). Greatest improvement in patients with highest BP. BP control rates increased from <50% to 75%.	Lack of randomization of the primary care sites involved to immediate intervention or delayed intervention groups.
Kruis [30]	2014	Netherlands	Pragmatic cluster randomised controlled trial. 2 day integrated disease management in practice course, including MI. The course also served as a network platform with development of an individual practice plan. Refresher course after 6 and 12 months. Also included decision support and benchmark reports. Intervention group 20 practices/554 patients;	20 practices in intervention (554 patients) and 20 practices (532 patients) in usual care group. To allow for sub-group analysis of MRC scores 1-2 V 3-5, 1080 participants needed to achieve 805 power with a significance of 0.05 including a 10% loss to follow-up.	Difference in COPD health status at 12 and 24 months measured by CCQ. Secondary outcomes: quality of life, dyspnoea, exacerbation related outcomes, self management ability, physical activity and level of integrated care. Patient questionnaire administered by nurses at baseline, 6 and 12 months. Postal questionnaire at 9, 18 and 24 months. Extraction of data from medical records at 24 months	No difference in CCQ between groups at 12 months. No difference in secondary outcomes, except follow-up/coordination) and % of self-reported physically active patients. At 24 months, no difference in outcomes, except co-ordination.	Query generalisability.

			Usual care 20 practices/532 patients.	Unit of allocation – GPs Unit of analysis – individuals.			
Vicens* [a] [31]	2014	Spain	Cluster randomised controlled trial. 77 GPs and 532 patients in 3 groups - Control (usual care),structured intervention with stepped dose reduction and follow-up visits (SIF) or structured intervention with written dose reduction (SIW). Two-hour workshop.	Sample size based on effectiveness at 12 months to detect 20% difference in proportion of patients who discontinued BZD use – 20% and 15% in SIF and SIW groups. Using ICC 0.0; cluster size of 8 and 25% patients assumed loss to follow-up. Unit of allocation – practice. Unit of analysis – individuals.	Primary outcome was benzodiazepine (BZD) use in long term users at 12 months assessed by prescription claim data.	At 12 months, 45% in SIW group, 45% in SIF group and 15% in control group had discontinued BZD use. No significant differences in anxiety, depression or sleep dissatisfaction.	Limited generalisability to more difficult to treat patients e.g. those with severe medical or psychiatric disease. Usual care differed – may impact efficacy of interventions.
Keeley* [32]	2014	USA	Cluster randomized trial in urban primary care clinics (3 intervention, 4 control); 21 PCPs (10 intervention, 11 control) and 171 English-speaking patients with newly diagnosed depression (85 intervention, 86 control). MI training included a baseline and up to 2 refresher classroom trainings, along with feedback on audiotaped patient encounters. Patients visits were as per guidelines -3 follow-up visits over the 12-week acute treatment phase, additional follow-up visits as needed during the 24-week continuation treatment phase, and prescription of	Unit of allocation – clinic. Unit of analysis – individuals.	Outcome measures include PCP MI training outcomes; depression management training outcomes and patient outcomes (change talk and physical activity). These include measures of technical (rate of MI-consistent statements per 10 minutes during encounters) and relational (global rating of “MI Spirit”) MI performance, the association between MI performance and number of MI trainings, rates of patient change talk regarding depression treatments (physical activity, antidepressant medication), PCP use of physical activity recommendations and antidepressant prescriptions and patients’ short-term	Use of MI-consistent statements and proficiency in MI Spirit was significantly higher for MI-trained versus control PCPs attending all 3 MI trainings. Although PCPs’ use of physical activity recommendations and antidepressant prescriptions was not significantly different by randomization arm, patients seen by MI-trained PCPs had more frequent change talk. Patients of MI-trained PCPs also expressed change talk about physical activity (p=	Paid training time may not be transferable to real World; small numbers hence generalisability. No primary outcome measure specified. PCPs not blind to patient participation.

			antidepressant medication over 36 weeks.		physical activity level and prescription fill rates.	.01) and reported more physical activity (3.05 vs 1.84 days in the week after the visit; P = .007) than their counterparts visiting untrained PCPs. Change talk about antidepressant medication and fill rates were similar by randomization arm (p > .05 for both).	
Kristoffersen* [b] [33]	2015	Norway	Pragmatic cluster randomised controlled trial. GPs randomised to receive brief intervention training (23 GPs/24 patients) or to control group (27 GPs/36 patients). 1-day course in headache management including a 2 hour BI role-play.	80% power with ICC of 0.5 and significance of 0.05, 18 patients or 5 GPs per arm. Taking on board pilot findings, increased to 50 GPs and 3 patients per GP. Unit of allocation – CME group. Unit of analysis – individuals.	Drug withdrawal in patients with medication overuse headache. Primary outcome measures were reduction in medication and headache days per month 3 months after intervention. Postal headache-screening questionnaire to all 18-50 year olds on practice lists. Validated diagnostic headache diary used prospectively. Other baseline data collected retrospectively at the blinded 3 month follow-up, conducted by a headache expert. Patients unable to attend were interviewed by phone. Another 2-week diary was completed prior to follow-up.	BI significantly better than usual care for primary outcomes. Chronic headache resolved in 50% of BI compared to 6% of control group.	Small clusters per GP.
van Dijk-de Vries [34]	2015	Netherlands	Pragmatic cluster randomised controlled trial. 40 nurses in 77 practices. Intervention (19 nurses) trained to integrate SMS into routine consultations. Control arm (21 nurses) provided usual care. Three 8-hour training sessions; 3-4 booster sessions during the 12 month period. Intervention	Sample of 232 (at least 5 patients per PN) would have 90% power and significance level of 0,05 to detect an improvement in perceived daily functioning (DFT <=4) at 12 months in 20% in intervention compared to 5R in	Primary outcome was a dichotomised score on a VAS measuring perceived effect of diabetes on daily functioning. Secondary measures were patients' diabetes related distress, quality of life, autonomy and participation, self-efficacy, self-management and glycaemic control. Measured at baseline, 4 and 12 month follow-up.	No significant differences in outcomes between intervention and control arms.	Low exposure of study participants to the complete intervention. Discrepancy between research driven screening and nurse led detection.

			group:19 nurses and 117 patients; Control group 21 nurses and 147 patients.	control arm. Assuming 30% loss to follow-up, invited 10 eligible patients per PN. 264 patients. Unit of allocation – practice. Unit of analysis – individuals.	Postal questionnaires for patient measurements; glycaemic control measured during diabetes consultations and extracted from patient electronic records. PNs recorded process measures and outcome on a specific registration form.		
Racic* [35]	2015	Bosnia and Herzegovina	Randomised control Trial. 10 doctors and 20 nurses from 10 family practices completed MI training consisting of 3-day's session with two half-day follow ups. Half-day course on management of diabetes type-2 also. 200 type-2 diabetic patients were randomly allocated into two groups - Intervention group (I-group) of 100 patients who were included in 3 month long program of motivational interviewing. Control group (C-group) of 100 participants who obtained patient education as a part of their regular care in family practice. All patients, irrespective of group, were scheduled to see their FPs every three weeks over the 3 months.	Unit of analysis – individuals.	Treatment outcomes included fasting blood glucose level, HbA1c, blood pressure, cholesterol, body mass index and smoking status. Data were collected at the enrolment in the study and after 3 months, during last visit in both groups.	Although both groups experienced changes in treatment outcomes from baseline to follow up, statistically significant improvements in fasting blood glucose level, HbA1c levels, blood pressure and serum cholesterol level were found in the I-group at follow-up compared to the C-group. Statistically significant differences in body mass index and smoking status were not found between the groups.	Possibly limited generalisability – 10 practices. No power calculation given; study detail lacking.
Vicens* [a] [36]	2016	Spain	Cluster randomised controlled trial. 77 GPs and 532 patients in 3 groups - Control (usual care),structured intervention with stepped dose reduction and follow-up visits (SIF) or structured intervention with written	Sample size based on effectiveness at 12 months to detect 20% difference in proportion of patients who discontinued BZD use – 20% and 15% in SIF and SIW groups. Using ICC 0.0; cluster size of 8	Primary outcome was benzodiazepine (BZD) use in long term users at 36 months assessed by prescription claim data.	At 36 months, 39.2% in SIW group, 41.3% in SIF group and 26% in control group had discontinued BZD use. No significant differences in anxiety, depression	Limited generalisability to more difficult to treat patients e.g. those with severe medical or psychiatric disease. Usual care differed – may impact

			dose reduction (SIW). Two-hour workshop.	and 25% patients assumed loss to follow-up. Unit of allocation – practice. Unit of analysis – individuals.		or sleep dissatisfaction.	efficacy of interventions.
Kristoffersen* [b] [37]	2016	Norway	Pragmatic cluster randomised controlled trial. GPs randomised to receive brief intervention training (23 GPs/24 patients) or to control group (27 GPs/36 patients). 1-day course in headache management including a 2 hour BI role-play.	80% power with ICC of 0.5 and significance of 0.05, 18 patients or 5 GPs per arm. Taking on board pilot findings, increased to 50 GPs and 3 patients per GP. Unit of allocation – CME group. Unit of analysis – individuals.	Drug withdrawal in patients with medication overuse headache. Primary outcome measures were reduction in medication and headache days per month 6 months after intervention. Secondary outcomes were proportion of patients with chronic headache and mediation overuse. Postal headache-screening questionnaire to all 18-50 year olds on practice lists. Validated diagnostic headache diary used prospectively. Other baseline data collected retrospectively at the blinded 6 month follow-up, conducted by phone by a headache expert. Uncontactable patients were sent a written questionnaire.	BI significantly better than usual care for primary and secondary outcomes. Chronic headache resolved in 63% of BI compared to 11% of control group.	Small clusters per GP.
Zwar [38]	2016	Australia	Pragmatic cluster randomised controlled trial. Nurses and GPs in intervention practices were educated to develop and implement disease management plans for COPD – PNs 8 hours of training; GPs online course; combined PN and GP 3 hour workshop. Care planning templates and prompts integrated into computer systems.	Assuming cluster size of 10, ICC of 0.01, design effect of 1.09; with 80% power to detect to detect >=4 unit difference in SGQ at 0.05 significance 200 participants per group needed. Participants were current and former smokers aged 40 to 85 years newly identified as having COPD on post-	The primary outcome was health-related QoL, assessed with the St George's Respiratory Questionnaire (SGRQ). Secondary outcome measures were other QoL measures, lung function (Post-bronchodilator FEV), disease knowledge, smoking and immunization status, inhaler technique and health service use. Data and lung function test conducted by project officers at home visits at baseline and	No between-group difference in the primary outcome measure (SGRQ) at follow-up.	Recruitment below target. Unanticipated misclassification in the practices of some spirometry results. Possible selection bias of practices - therefore, the practices may not be representative of all practices in Australia.

			Intervention: 19 practices, 144 patients. Control: 17 practices, 110 patients.	bronchodilator spirometry. Unit of allocation – practice. Unit of analysis – individuals.	12 months; 6 month data collected by telephone.		
van Lieshout [39]	2016	Netherlands	Cluster randomized trial in 34 general practices involving 34 nurses (2229 patients). Tailored improvement program, which included communication skills training, online patient information, and a clinical protocol for managing depressive symptoms. Structured feedback (refresher training) to nurses on their MI skills; online education on CVRM, written guidance on relevant e-health guidance for patients. Control practices offered usual care and were offered delayed intervention.	Powered to detect a 15% difference on the primary outcome, assuming ICC of 0.05, significance 0.05, and a power of 0.80, 450 patients per group (high risk or established CVD) would be needed (15 patients at high risk for CVD and 15 patients with established CVD per cluster, sampled in 30 practices.) Unit of allocation – practice. Unit of analysis – individuals.	Primary outcome was an aggregated score of a positive score on lifestyle counselling delivered and an appropriate action on depressive symptoms. Secondary outcomes included the various elements of the primary outcome, vascular risk factors (extracted from patient records), and patient-reported lifestyle behaviors. Data were collected from medical records and a written survey among included patients. 6 month follow-up.	No effect from the intervention on the primary outcome of this study. Regarding secondary outcomes, it was found that physical exercise showed a significant improvement in the intervention group compared to the control group.	The generalizability of findings may be limited by the low recruitment rate. Intervention intensity may have been insufficient e.g. feedback on two patient consultations. Low patient exposure to intervention.
Vaillant-Roussel [40]	2016	France	A cluster randomised controlled clinical trial. Included 241 patients with chronic heart failure CHF attending 54 general practitioners (GPs) in France and involved 19 months of follow-up. The GPs in the Intervention Group were trained during a 2-day interactive workshop (including patient simulation) to provide a patient education programme (which consisted of 6 sessions). GPs in the control group received a 3-hour	To detect a difference of 12 points for quality of life outcomes (SF-36 and MLHFQ), which corresponds to an effect size of 0.6, with a statistical power of 90% and a two-sided Type 1 error of 5%, taking into account ICC between 0.1 and 0.2 and a 20% dropout, 40 GPs recruiting 5 patients each. Unit of allocation – GPs.	The primary outcome was patients' quality of life, as measured by the MOS 36-Item Short Form Health Survey and the Minnesota Living with Heart Failure Questionnaire (MLHFQ). Secondary outcomes were all-cause and HF associated mortality, number of days spent in hospital, number of deaths, number of cardiologist visits, additional GP visits, adherence to therapy, evolution of NYHA HF stage, and changes in weight and BMI.	Elderly patients with stable heart failure in the ETIC programme did not achieve an improvement in their quality of life compared with routine care. There was no difference in MLHFQ score during followup at 7 and 13 months, or at study end 19 months.	Dropout rate of 36% after randomisation among GPs, either because they withdrew consent to participate (31%) or failed to recruit patients (5%). Recruitment targets not reached.

			information session (patients had same schedule of visits w/o educational component).	Unit of analysis – individuals.	Data collected at baseline, 7, 13 and 19 months using self-administered questionnaires completed by the patient or care giver within 7 days of GP appointment.		
Griffiths* [41]	2016	UK	Cluster randomised controlled trial. 105 general practices were randomised to usual care or the education programme. The programme had two components: the Physician Asthma Care Education (PACE) programme (2 two-hour sessions using a DVD with examples of consultations); and the Chronic Disease Self Management Programme (CDSMP) for patients (South Asians aged 3 years and older with asthma) by lay trainers with an extra 2-hour session by an asthma nurse. All patients saw an asthma control nurse; those in the intervention group additional received a self-management plan and CDSMP. 105 practices participated with 183 patients in intervention and 192 in control group.	Without taking account of clustering, to detect a clinically important difference of 68% to 48% with 80% power for 5% significance requires 105 participants in each group. Unit of allocation – practice. Unit of analysis – individuals.	Primary outcomes were unscheduled care, time to first unscheduled contact with exacerbation and proportion of patients without unscheduled care. Secondary outcomes were time to first asthma review in primary care, asthma specific and general health QoL and prescribing. Data from patient records and interviews. Data extracted one year before and one year after intervention. Interviews in person at baseline and by phone at 3 and 12 months after recruitment.	At 12 months, no reduction in unscheduled care. Improvement seen in patients' quality of life and self-efficacy (confidence to control asthma) and follow up in primary care (shortening time to review and increasing the proportion followed up).	Intervention - Low number of patients attending the CDSMP sessions.
Eikelenboom [42]	2016	Netherlands	Cluster randomised controlled trial was conducted in 15 primary care group practices. After attending a dedicated self-management support training session (one two-hour session to GPs and nurses), practice nurses in the intervention arm discussed the results of	Anticipating a 30% response rate and a 33% attrition rate, 150 patients per practice were invited to participate. Based on the number of included patients after 3 months, 100 additional patients were invited from one	Primary outcome was PAM-13 score. Secondary outcomes were exercise, nutrition and smoking. Participants completed a 13-item Patient Activation Measure (PAM-13) and validated lifestyle questionnaires at baseline and after 6 months.	PAM-13 did not differ significantly between groups at follow-up. The intervention showed a positive effect on the percentage of patients performing self-monitoring (primary analysis and per protocol analysis)	Possible low exposure - approximately one-third of patients in the intervention group reported that SeMaS was not discussed in their consultation.

			SeMaS with the patient at baseline, and tailored the self-management support. One visit to practices also shortly after the training. Intervention group 7 practices/359 patients; Control 8 practices/404 patients.	control (n = 50) and one intervention practice (n = 50).		and on the number of individual care plans (per protocol analysis). The intervention was found to have no effect on exercise, nutrition, or smoking.	
Ramli* [43]	2016	Malaysia	Pragmatic cluster randomised controlled trial. 10 public primary care clinics, matched in pairs – 471 patients in intervention and 417 in control group with type 2 diabetes. The EMPOWER-PAR intervention consisted three obligatory and two optional components. Five team members from each clinic trained in workshops with each team producing a proposed intervention plan. Support to practices continued throughout the study with review workshop at 6 months. 'Green booklet' used with patients.	To detect a 25% change in proportion achieving HbA1C target, ICC 1.5, sample size of 626, allowing for 25% drop out, 84 patients in 10 clusters with 91% power at 5% significance level. Unit of allocation – practice. Unit of analysis – individuals.	Primary outcome was change in proportion of patients achieving HbA1c <6.5%. Secondary outcomes were change in proportion of patients achieving targets for BL, lipid profile, BMI and waist circumference. One year follow-up. Minimum two visits to CDM team – no maximum. Baseline and follow-up physical examination, blood sample and interview in addition to ongoing data on clinically important events.	Significantly more patients in the intervention group achieved HbA1c target. No significant improvement in the secondary outcome variables.	Generalisability given clinic selection criteria.
Keeley* [44]	2016	Canada	Cluster randomised controlled trial. To examine effectiveness of MI on rates of improvement for depressive symptoms and remission among low income patients with newly diagnosed major depressive disorder – 7 facilities housing 10 care teams. Intervention: MI with standard management of depression -4 teams; 10 providers, 88 patients. Control: SMD alone 6	Assuming group x time interaction estimate of -0.75, 66 subjects per group required for 79% power to detect the interaction for the binary outcome. Unit of allocation – care teams. Unit of analysis – individuals.	Primary outcome was remission from depression (defined as PHQ-9 < 5 over past 2 weeks). Patients assessed at 6, 12 and 36 weeks with PHQ-9. Patient inquiry and electronic records used.	Intervention group showed improved depressive symptoms and remission rate. Patients attending providers trained in MI had a more favourable PHQ-9 trajectory over 36 weeks. Group x Time interaction not show a significant difference by group for trajectories of 2-week remission.	Results may not be generalizable beyond low income populations of mixed race/ethnicity.

			teams, 16 providers, 80 patients. Intervention group received 8 hour interactive training; 4 hour refresher session at 4 and 12 months and email and face-to-face feedback over first 14 months. Control group received a 1 hour slideshow.				
Kristoffersen* [b] [45]	2017	Norway	Pragmatic cluster randomised controlled trial. GPs randomised to receive brief intervention training (23 GPs). After the initial followup, the control group (27 GPs) received BI training. Cross-over of patients then possible: BI early (n = 24): followed up for 16 months after BI, remaining BI throughout the study and representing the original blinded BI group. BI late (n = 22): followed up on average 6 months after their BI (16 months after inclusion), representing patients crossed over from the original BAU to the intervention after the blinded follow-up. BAU (n = 14) followed up for 16 months after inclusion in the study, representing patients of the original BAU group that could not be crossed over due to logistics (not possible to arrange a BI meeting with the GP during the study period).	Unit of allocation – CME group. Unit of analysis – individuals. Follow-up open after 6 months.	Primary outcome measures were reduction in medication and headache days per month and change compared to baseline. Secondary outcomes were the proportion of patients with chronic headache, medication overuse. Psychological and disability scores prior to intervention were only available for BI late as study-related contact prior to the blinded intervention was avoided in order to minimize assessment effects – study not powered for these latter outcomes. Outcomes based on telephone interview and self-report questionnaire.	In the early BI group, average of 16 month follow-up significant change from baseline in headache and medication days. In late BI group, with average of 6 month follow-up, significant change from baseline in headache and medication days. Patients remaining in the BAU group (no BI) did not improve significantly compared to baseline. Disability and psychological scores for BI late (only group where before and after data were available) - significant reduction in mean HSCL-25 anxiety score in the compared to before medication withdrawal. There were non-significant changes in total HSCL-25, HADS and headache disability.	Small clusters per GP. Five of 60 patients lost to follow-up.

Baldeón* [46]	2018	Ecuador	Randomised clinical trial. A physician-based and patient-centred counselling program was delivered to eight PCPs (113 patients) – 4 hour session with training binder and patient materials, and taught to implement 7-10 minutes patient centred counselling protocol; live examples and role play used. Further practical reinforcement training carried out after in physicians' offices. Seven PCPs (84 patients) who did not receive the training comprised the control group. Patients at risk of developing type-2 diabetes.	A PEI score difference of 2 points±2.9 was used, with ICC of 0.3, a sample of 200 subjects provided 84% power. Unit of allocation – practice. Unit of analysis – individuals.	Anthropometric and blood pressure measurements were taken at the baseline visit and at the 6-month follow-up visit at a study patient visit. To evaluate the physicians' changes in attitudes and knowledge on physical activity, diet, and lipid therapy, and any improvements in counselling skills, pre- and post-training assessments using a diet and exercise goal sheet, the Diet and Physical Activity Risk Assessment [DARA]. Patient experience assessed by exit interview (PEI). A blood sample was also taken at both time points.	Counselling steps, measured by PEI, were significantly higher in intervention group. Significant improvements in weight, BMI, HbA1C, total cholesterol and LDL cholesterol in intervention group.	Feasibility study only was carried out with a limited number of PCPs and clinics that could have specific administrative and health care provision characteristics, which could limit the generalizability of current results. Short follow-up period.

[a] Papers arising from the same project reporting on follow up at 12 and 36 months respectively.

[b] Papers arising from the same project reporting on follow up at three, six and an average of 16 months respectively.

Table S2: Risk of bias (¥ indicates low risk of bias)

		Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Selective reporting	Selective reporting	Other
First Author	**Year**	**Selection Bias**		**Performance Bias**	**Detection Bias**	**Attrition Bias**	**Reporting Bias**	**Other Bias**
Tobe [29]	2014	High – no randomisation	High – no concealment	High – no blinding	High – no blinding	Low- clear participant flow reported	Low – all expected outcomes reported	Intervention Integrity moderate. Analysis by actual not intention to treat.
Kruis¥ [30]	2014	Low - computer generated randomisation by blinded researcher	Low - concealment	High – no blinding of practices or patients	Low – outcome assessor blinded	Low- clear participant flow reported	Low – all expected outcomes reported	Intervention Integrity moderate. Pre-existing high level of COPD care may have limited potential for improvement. Analysis by intention to treat.
Vicens¥ [31]	2014	Low - computer generated randomisation	Low - concealment	Low – Nurses blinded	Low primary; High secondary	Low	Low – all expected outcomes reported	Intervention Integrity high. Possible cross contamination. Analysis by intention to treat.
Keeley [32]	2014	Low- randomised	Unclear	Unclear	High – patients	Low	Low – all expected	Analysis by intention to

		using RAND () function by independent researcher			blinded; PCPs not blinded.		outcomes reported	treat. No primary outcomes specified.
Kristoffersen¥ [33]	2015	Low - randomised to BI training	Low - concealment	Unclear	Low - blinded	Low	Low – all expected outcomes reported	Intervention Integrity moderate. Possible recall bias. Analysis by actual not intention to treat.
van Dijk-de Vries¥ [34]	2015	Low - computer generated randomisation	High - no concealment	Low- practice nurses blinded; patients not blinded	Low - blinded	Low	Low – all expected outcomes reported	Intervention Integrity moderate. Analysis by actual not intention to treat.
Racic [35]	2015	Unclear	Unclear	High - PCPs not blinded	Unclear	Low	Low – all expected outcomes reported	Unclear
Vicens¥ [36]	2016	Low - computer generated randomisation	Low - concealment	Low – Nurses blinded	Low primary; High secondary	Low	Low – all expected outcomes reported	Intervention Integrity high. Possible cross contamination. Analysis by intention to treat.
Kristoffersen¥ [37]	2016	Low - randomised to BI training	Low - concealment	Unclear	Low - blinded	Low	Low – all expected outcomes reported	Intervention Integrity moderate. Possible recall bias.

								Analysis by actual not intention to treat.
Zwar¥ [38]	2016	Low - computer generated randomisation	High - no concealment	High – nurses, GPs, PNs and patients not blinded	Low – project officers collecting outcome data blinded	Low	Low– all expected outcomes reported	Intervention Integrity moderate. Analysis by intention to treat.
van Lieshout [39]	2016	Low - computer generated block randomisation	Unclear	High – not blinded	Unclear	Unclear – data flow presented but inconsistent numbers reported	Low– all expected outcomes reported	Intervention Integrity moderate. Analysis by intention to treat.
Vaillant-Roussel¥ [40]	2016	Low - randomisation	Unclear	Unclear but possible blinding of patients	Low- blinded	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity moderate. Analysis by intention to treat.
Griffiths¥ [41]	2016	Low - randomisation	Low - concealment	Interviewers blinded GPs and Patients – not blinded	Unclear	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity moderate. Analysis by intention to treat.
Eikelenboom [42]	2016	Low – two block randomisation by one of authors	Unclear	High – researchers blinded to allocation; no blinding of clinicians or patients	Unclear	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity low. Analysis by actual not intention to treat.

Ramli¥ [43]	2016	Low - computer generated block randomisation	Unlcear	High – no blinding	Unlcear	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity high. Fidelity monitored throughout. Analysis by intention to treat.
Keeley¥ [44]	2016	Low - computer generated randomisation by independent clinician	Low - concealment	Low- Outcome assessors/data collectors blinded; providers not blinded; patients blinded	Low - Outcome assessors/data collectors blinded	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity moderate. Analysis by intention to treat.
Kristoffersen [45]	2017	High – no randomisaton; cross over	High – concealment not possible for cross-over group	Likely high	High – no blinding	Low	Low – all expected outcomes reported; some only reported for BI late group to avoid assessment effect.	Fidelity unclear - control arm cannot be considered fully untreated.
Baldeón [46]	2018	High – no random allocation	Unclear	High – no blinding	Unclear	Low - data flow presented	Low– all expected outcomes reported	Intervention Integrity moderate but contamination possible. Analysis by actual not intention to treat.